"Think" pragmatically: Children’s interpretation of belief reports
Shevaun Lewis, Valentine Hacquard, & Jeffrey Lidz

Abstract: It has often been reported that children under 4-5 years of age evaluate belief reports based on reality instead of beliefs. They tend to reject sentences like, “John thinks that giraffes have stripes” on the grounds that giraffes do not have stripes, even if that’s what John really believes. Previous accounts have proposed that these non-adult-like judgments reflect either a lack of conceptual understanding of belief, or incorrect syntactic/semantic representations. We argue that the problem is actually pragmatic. Adults frequently use belief reports in situations where the main point is to provide information about reality, rather than someone’s mental state (e.g., “I think the stove is still hot”). Young children have difficulty determining when adults are talking about reality (the stove situation), and when they’re talking about mental states (John’s ideas about giraffes). We show that if the context emphasizes beliefs, children are more able to evaluate belief reports appropriately (Experiment 1). The pattern of children’s truth value judgments demonstrates that they do understand the literal meaning of think sentences, despite their pragmatic difficulty grasping the speaker meaning (Experiment 2).

Keywords: belief reports; false belief; Theory of Mind; pragmatic inference; truth value judgment task

1 Introduction

“Mental verbs” like think, hope, and want pose a considerable challenge for children learning their first language. They refer to aspects of the world that are not directly observable and rarely play a salient role in conversation. It is unsurprising, then, that children seem to acquire mental verbs relatively late compared to other verbs (de Villiers J. G., 1995; Johnson & Maratsos, 1977; Shatz, Wellman, & Silber, 1983).

Here we focus on the belief verb think. Children seem to have non-adult-like interpretations of belief reports when the reported belief conflicts with reality, as in (1) (de Villiers J. G., 1995; de Villiers & de Villiers, 2000; de Villiers & Pyers, 2002; Johnson & Maratsos, 1977; Sowalsky, Hacquard, & Roeper, 2009). For example, 3-4 year olds would tend to say that (1) is false, even if John does think that dogs quack.

1. John thinks that dogs quack.

Children’s non-adult-like interpretations of think in false belief scenarios align with their notoriously poor performance in traditional “false belief tasks”. In the change-of-location false belief task, children are asked how a character with a false belief will behave. In a representative story from Wimmer & Perner (1983), a boy named Maxi is helping his mom put away groceries. He puts some chocolate in the blue cupboard before going out to the playground. While he is gone, Maxi’s mom uses the chocolate to make a cake, and puts the leftovers into the green cupboard. Then Maxi returns to eat some chocolate. Children are asked the test question in (2).

2. Where will Maxi look for the chocolate?
In Wimmer and Perner’s study and the dozens of similar studies that followed, 3-year-olds’ performance is quite poor. They appear to be biased by their own knowledge: in response to a question like (2), they would often say that Maxi would look for the chocolate in its actual location (the green cupboard) rather than where he put it (the blue cupboard). In a meta-analysis of 178 studies, Wellman, Cross and Watson (2001) found that children respond based on reality more often than not until about 3;5 (3 years, 5 months). They start responding based on beliefs at about 4;0 on average. Thus, the “curse of knowledge” seems to affect children’s performance in both linguistic and non-linguistic tasks involving false beliefs during similar timeframes of development.

However, previous studies did not provide a fine-grained assessment of children’s truth conditions for think sentences, which would allow us to infer exactly how their interpretation of belief reports differs from adults’. We compare three potential sources of difficulty with belief reports: (1) belief concepts, (2) the semantics of belief reports, and (3) the pragmatics of conversational use of belief reports.

1.1 The conceptual hypothesis
It is reasonable to assume that understanding a concept at some non-linguistic level is a prerequisite for mapping a word to that concept. Under this assumption, children’s non-adult-like understanding of belief reports follows inevitably from their poor performance on the non-linguistic tasks. A likely explanation of both is that children lack an understanding of false belief at the conceptual level. The change that occurs around 4 years of age, causing improvements on both tasks, is that children become able to attribute false beliefs to others.

This hypothesis encounters two main empirical challenges. The first is that although performance on linguistic and non-linguistic tests of belief understanding are closely correlated in development, it seems that children succeed earlier on the linguistic tasks. For example, in a longitudinal study with 3-4 year-old children, de Villiers and Pyers (2002) found that there were strong correlations at each time point between false belief task performance and language measures, particularly one involving memory for false complements. Children were told a short story with pictures, and then asked a belief question, as in (3).

(3) This girl saw something funny at a tag sale and paid a dollar for it. She thought it was a toy bird but it was really a funny hat. What did she think she bought?

Critically, children were much more likely to pass the language task earlier than the false belief tasks, suggesting that language is the causal factor driving their improved performance. Evidence for an apparent causal role of language development in performance on false belief tasks has been found in several different populations, including children with autism (Tager-Flusberg & Joseph, 2005), children with specific language impairment (de Villiers, Burns, & Pearson, 2003), and deaf children and adults with either delayed or normal language development (de Villiers P. A., 2005). In a meta-analysis of 107 studies, Milligan and colleagues found large effects for several different kinds of language measure on false belief performance, and early language was more likely to predict later false belief performance than vice versa (Milligan, Astington, & Dack, 2007). Finally, training studies have provided more direct evidence that understanding of sentential complements, especially false complements, is causally related to success on false belief tasks (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003).
The second main empirical challenge for the conceptual hypothesis is that children seem to be able to reason about false beliefs long before their fourth birthday. Many researchers have argued that traditional false belief tasks are fundamentally flawed as measures of young children’s underlying competence since they require an explicit decision and response. Young children succeed earlier in false belief tasks if the critical question or task is less direct (Buttelmann, Carpenter, & Tomasello, 2009; Knudsen & Liszkowski, 2012; Rubio-Fernández & Geurts, 2012; Scott, He, Baillargeon, & Cummins, 2012; Southgate, Chevallier, & Csibra, 2010). Infants as young as 13 months show understanding of the behavioral consequences of false beliefs when they are tested using more implicit measures in preferential looking or eye-tracking studies (Baillargeon, Scott, & He, 2010; Clements & Perner, 1994; He, Bolz, & Baillargeon, 2012; Onishi & Baillargeon, 2005; Song & Baillargeon, 2008; Song, Onishi, Baillargeon, & Fisher, 2008; Southgate, Senju, & Csibra, 2007; Surian, Caldi, & Sperber, 2007). 7-month-olds are able to at least track the beliefs of others (Kovács, Téglás, & Endress, 2010). These findings align with evidence that even 2-year-olds engage in numerous behaviors that would seem to require them to attribute belief states to others. For example, they attempt to deceive, which requires an understanding that it is possible for others to have false beliefs (Chandler, Fritz, & Hala, 1989). They also attempt to help people that they know to have a false belief about the location of an object (O'Neill, 1996).

This evidence suggests that the development occurring in the preschool years is not the emergence of a “new” concept of belief. Rather, it must involve some more subtle aspect of children’s representational or processing abilities, such that they can access their knowledge in some tasks but not in others. One possibility that has been suggested is that attributing false beliefs is a cognitively demanding task that easily breaks down under stress. Standard false belief tasks prompt children to engage in an explicit reasoning process that contrasts someone else’s (false) belief with their own (true) belief. Considering both belief states simultaneously may overwhelm children’s limited processing capacity, causing them to respond based on the most salient and available representation—their own beliefs. This hypothesis gains support from the fact that children’s performance on false belief tasks is improved when the conflict with their own beliefs is reduced (Wellman, Cross, & Watson, 2001)—for example, when the object is removed from the scene in Wimmer and Perner’s (1983) “Disappear” condition.

To summarize, the conceptual hypothesis is that children misunderstand belief reports for the same reason that they perform poorly on false belief tasks: they have difficulty representing, tracking, or otherwise reasoning about false beliefs.

1.2 The syntax/semantics hypothesis

The semantic hypothesis takes the opposite approach from the conceptual hypothesis, by proposing that the problems with linguistic belief reports are primary, rather than the problems with non-linguistic false belief understanding. This account is motivated by the findings (described in the previous section) that belief report understanding seems to be a prerequisite for success on traditional false belief tasks.

Jill de Villiers has argued that the crucial development that occurs between the ages of 3 and 4 is the acquisition of the semantic structures necessary to represent attitude reports with false
complements (de Villiers J. G., 2005; 2007; de Villiers & de Villiers, 2000; 2009; de Villiers & Pyers, 2002). The critical features of false complements seem to be that they are truth-evaluable and that they provide a way to represent a particular perspective on the world. De Villiers suggests that children’s mastery of false complements for communication verbs—which correspond to observable events in the world—allows children to bootstrap their way into an understanding of mental verbs like think. These semantic structures for representing mental attitudes are then crucially involved in successful performance on a false belief task.

The semantic hypothesis, like the conceptual hypothesis, is challenged by the recent evidence that very young children and infants have some concept of false belief. If it were true, as de Villiers argues, that a semantic representation of think sentences is a necessary precursor to belief attribution, pre-linguistic infants should not show any understanding of false beliefs. However, linguistic representations could affect the non-linguistic attribution of mental states in some way other than by providing access to the concept. One possibility is that pre-linguistic “implicit” understanding of false belief reflects the output of a qualitatively different mental system than the “explicit” understanding demonstrated in traditional false belief tasks (Apperly, 2011; Apperly & Butterfill, 2009; Carruthers, 2002; 2009). Humans may come equipped with an innate system for attributing and processing mental states. This system would be fast and informationally encapsulated, and thus limited in scope. It might be able to generate simple expectations about behavior—which would be reflected in eye movements—but not the explicit, reasoned decisions required in the traditional false belief task. A later developing, more general system reliant on language and executive function would be able to use the outputs of the low-level system more flexibly. Language might play a crucial role in this high-level system by representing information in a way that is interpretable by multiple mental systems (Carruthers, 2002).

A more fundamental problem with the semantic hypothesis is that, as more basic structural properties (for example, sentential complementation) are ruled out as the key feature for children’s acquisition of false complements, the account becomes less explanatory. It is difficult to distinguish a “point of view” feature on a complement clause (de Villiers & de Villiers, 2009) and the conceptual representations that it is supposed to enable. Even so, the abundant evidence for the facilitative effect of language on false belief reasoning must be accounted for.

1.3 The pragmatic hypothesis
We propose that children’s difficulty with belief reports is pragmatic. Specifically, 3-year-olds’ non-adult-like interpretations reflect the computation of inappropriate speaker meanings, not

---

1 J. de Villiers initially hypothesized that the main difficulty was the syntax of the tensed sentential complements that think selects (compared to the tenseless complements that want selects). However, cross-linguistic evidence has demonstrated that belief reports are more difficult than desire reports even if they have the same kind of complements (German: Perner, Sprung, Zauner, & Haider, 2003; Mandarin: Tardif & Wellman, 2000). While the conceptual hypothesis has no problem accounting for these cross-linguistic consistencies, since the relevant conceptual development would presumably unfold in the same way cross-culturally (Callaghan, et al., 2005; Liu, Wellman, Tardif, & Sabbagh, 2008), they are problematic for any version of the syntax/semantics hypothesis that emphasizes tensed complements or any other single syntactic property (but see Hacquard (to appear) for a proposal on how the syntactic properties of complements to think and want converge cross-linguistically at a more abstract level, beyond tensedness).
incorrect literal meanings. This hypothesis has two main motivations. First, adults tend to use belief reports in contexts where the beliefs being discussed are not salient and the speaker implicitly endorses the truth of the complement clause. Second, children’s early uses of think and other mental verbs are consistent with this endorsement meaning.

Belief reports can be used to express different speaker meanings in different contexts (Hooper, 1975; Simons, 2007; Urmson, 1952). Consider the different implications of the same belief report in the different contexts in (4)-(5).

(4) A: Why didn’t Mary invite John to the meeting?  
   B: She thinks he’s working from home.

(5) A: Where is John? It’s time to start the meeting.  
   B: Mary thinks he’s working from home.

The exchange in (4) is about explaining Mary’s behavior, so her mental states are highly relevant. B’s utterance is therefore intended as a straightforward report of Mary’s belief, which may or may not be true in the actual world. The exchange in (5) is about John’s whereabouts, so Mary’s mental states are less relevant. The main point of B’s utterance is to provide information about John, but it also notes that the source of the information is Mary. Thus, B’s main contribution to the conversation is the proposition expressed by the complement clause (John is working from home), rather than that expressed by the whole sentence. Given that conversational participants are generally assumed to contribute only what they know to be true (Grice’s maxim of Quality), B implicitly endorses the truth of the complement clause in the actual world. However, if B were entirely certain about John’s whereabouts, he probably would have simply said so directly. So the inclusion of a source of evidence implies that B only endorses the proposition that John is working from home to the degree that Mary can be trusted. This hint of uncertainty is amplified in first person belief reports like (6).

(6) I think he’s working from home.

Studies of child directed speech have found that adults’ uses of belief reports overwhelmingly occur in the first person and have this endorsement meaning (Diessel & Tomasello, 2001; Howard, Mayeux, & Naigles, 2008; Naigles, 2000), as in the examples in (7)-(8).

(7) I think you should try and finish one game.  
(8) I don’t think you’ll be working on the roof.  
   (Howard et al., 2008)

Given the preponderance of endorsement uses in adult speech, it is unsurprising that children’s early uses of think have the same flavor. Children begin using think to refer to mental states before age three (Shatz, Wellman, & Silber, 1983). Although their sentences with think often include clausal complements, they occur overwhelmingly in the present tense with the first person singular subject ‘I’ (Bloom, Rispoli, Gartner, & Hafitz, 1989; Diessel & Tomasello, 2001). These early uses function more often to direct the conversation or express uncertainty, as in (9)-(13), than to comment on someone’s belief state.

(9) I thoughted we’d eat some cake.  
(10) I think this is a lamb. 
(11) I think I’m go in here. [3;1]  
   (Bloom et al., 1989)  
   (Diessel & Tomasello, 2001)
(12) Think some toys over here too. [3;2]
(13) It’s a crazy bone I think. [3;5]

Later, children begin to use *think* with a wider variety of subjects, tenses, and aspects, and in interrogative sentences. The wider variety of syntactic contexts corresponds to a greater proportion of uses to describe mental states (Diessel & Tomasello, 2001).

Our pragmatic hypothesis is that children tend to assume endorsement uses of belief reports, even in situations where beliefs are actually relevant. We suggest that although children are capable of computing the literal meaning of belief reports, they often misjudge the discourse context and fail to recognize when beliefs are relevant to the conversation.

To make the hypothesis more concrete, let’s walk through an example. Suppose we walk into a room and observe my dog frantically sniffing and pawing at an empty treat box. I might utter the belief report in (14).

(14) Jasper thinks that there are still treats in there.

As an adult, you understand that the implicit Question Under Discussion I just introduced is, “Why is Jasper so interested in that empty treat box?” We both already know that there are actually no treats in the box, so it would be irrelevant to comment on that fact.

Now suppose that I’m having this conversation with a 3-year-old child instead of an adult. When we walk into the room and I utter (14), the child has to figure out why I would have said such a thing. One possibility is that I’m commenting on Jasper’s erroneous belief state, but this possibility may not be particularly salient. An alternative possibility is that I don’t know whether there are any treats in the box, and I’m citing Jasper’s behavior as a source of evidence that there are. That is, the child assumes that I’ve just introduced a different Question Under Discussion: “Are there any treats in the box?” Under this assumption, the child might deny my statement with (15). This is the behavior that has been observed in previous studies: denying a belief report based on reality.

(15) No there aren’t! (I already gave her the last one!)

The pragmatic hypothesis we are proposing is that children often fail to recognize the relevance of belief in context, and therefore assume that the speaker meaning has to do with reality. Why do children so often fail to recognize that beliefs are relevant? It may be because they do not track people’s beliefs as automatically or as quickly as adults do. Alternatively, they may not understand when people are likely to be talking about beliefs, even if they accurately track beliefs. Young children do not seem to be exposed to much discussion about belief states, so they may not be able to guess when a belief state is noteworthy enough to be the topic of a conversation. This latter possibility opens up a possible explanation of children’s difficulty in traditional false belief tasks, which we will discuss in the last section.

The pragmatic hypothesis makes two clear predictions. First, we should be able to induce children to judge belief reports based on beliefs rather than reality by drawing their attention to the relevance of belief in a particular context. We test this prediction in Experiment 1. Second, if children have access to the literal meaning of belief reports, they should be able to reject belief
reports that are literally false, regardless of whether they compute an inappropriate speaker meaning. We test this prediction in Experiment 2.

2 Experiment 1: Context sensitivity \(^2\)

The goal of Experiment 1 was to test the prediction that if the relevance of belief is made more salient in the context, children will infer the correct speaker meaning for belief reports more often and show a more adult-like pattern of judgments. We presented children with stories about hide-and-seek and asked them to judge belief reports about the seekers’ beliefs. In one condition, we enhanced the relevance of belief by introducing a conflict of belief between two seekers. If children are influenced by this contextual manipulation, they should show more adult-like responses in the critical false belief conditions. Neither the conceptual nor the syntax/semantics hypothesis predicts that children should be influenced by context, so they would not predict any improvement.

To determine how much of children’s difficulty could be attributed to difficulty inhibiting their own knowledge, we included a condition in which the child was ignorant of the location of the hider. If children’s difficulty in the false belief condition is due to difficulty processing the conflict with their own knowledge, they should perform better in this “ignorance” condition. By contrast, the pragmatic hypothesis does not predict any improvement in the ignorance condition, since the children could still mistake the intended speaker meaning.

2.1 Methods

2.1.1 Participants
36 children aged 3 years, 10 months (3;10) to 4 years, 5 months (4;5) participated in the study (3.8-4.4 years, mean = 3.9, 17 girls). Participants were recruited from the Center for Young Children preschool or the Infant Studies Database at the University of Maryland. All participants were typically-developing monolingual English-speakers.

2.1.2 Design
Children were presented with stories about hide-and-seek. After each story, a puppet uttered a target sentence containing _think_, and the child was asked to judge whether the puppet was “right” about what happened.

2.1.2.1 Sample story
All stories followed the same template, illustrated in the following sample story.

In the first scene, the characters (Swiper and Dora) are named and the experimenter confirms that the child can identify them. Swiper is identified as the Hider, Dora as the Seeker: _Swiper is gonna hide, and Dora will look for him. So she’ll wait in the other room where she can’t see._

Dora leaves, and the child watches as Swiper hides behind the curtain. His yellow tail remains visible, protruding from behind the curtain. Then a squirrel (the Distractor) hides behind the toy

\(^2\) These results were reported previously in Lewis, Hacquard, & Lidz (2012).
box, leaving an identical yellow tail visible (Figure 1a). The experimenter points out the two clues to ensure that the child knows what evidence Dora will be using to guess Swiper’s location.

Dora reappears to state a guess about Swiper’s location based on one of the clues: *Hmm, where should I look? Oh! I see a yellow tail behind the toy box! I know--Swiper is there! I'll look for Swiper behind the toy box.* The Seeker’s script is intended to establish that she is just guessing based on the first clue she noticed, but she is nevertheless confident—she believes what she’s saying. Dora moves toward the toy box as she speaks and remains there for the rest of the story as a cue to her stated belief (Figure 1b).

At this point, the experimenter asks the puppet to say something about what’s going on in the story. The puppet delivers a target sentence like (16). After the child responds, the puppet delivers a filler sentence. Once the child has responded to both the target sentence and the filler, the Hider and Distractor emerge from their hiding places.

(16) Dora thinks that Swiper is behind the toy box.

2.1.2.2 Manipulations
Within the stories, we manipulated whether the child had KNOWLEDGE of the Hider’s true location. In the knowledge condition, the child watched as the Hider and Distractor hid in the scene (as in the sample story). In the ignorance condition, the screen was obscured during the hiding, so the child did not know which clue corresponded to the Hider until after responding to the sentences.

We also manipulated the BELIEF TYPE: whether the target sentence referred to a Seeker with a true belief or a false belief. In the sample story, the target sentences are about a seeker with a false belief. Note that in the ignorance condition, it is unknown at the point of the target sentence whether the Seeker has a true or false belief. The truth of the target sentences (i.e., the target

---

Figure 1 Experiment 1: Sample scenes.
(a) Identical clues for the Hider (Swiper, behind the curtain) and Distractor (squirrel, behind the toy box). (b) The Seeker (Dora) guesses the location of the Hider.
response) was counterbalanced. Table 1 shows the set of possible think target sentences for the sample story.

The most important manipulation for our pragmatic hypothesis was the NUMBER OF SEEKERS who looked for the hider. In the 2-seeker stories, a second seeker guessed the other location, so that one seeker had a true belief and the other a false belief. The 2-seeker stories were intended to heighten the relevance of belief in context by introducing an important conflict of belief. The NUMBER OF SEEKERS was a between-subjects factor, so each child saw only 1-seeker stories or only 2-seeker stories.

2.1.3 Materials
We created 14 stories including a variety of scenes and characters. The locations of the Hider and Distractor were spread across the different hiding spots across trials, and the characters playing the Hider and Seeker rotated from story to story. The stories were illustrated and animated in Adobe Flash. We recorded narration for each story and added it to the animated videos.

There were two lists of target sentences. In each list, the order of sentences with respect to Belief Type and sentence truth was pseudo-randomized. Two filler sentences, one true and one false, were created for each story. The fillers did not involve belief. They were created using templates exemplified by (17)-(21). (18)-(19) were only appropriate in knowledge stories, and (21) in ignorance.

(17) Dora is looking for Swiper {behind the toy box/behind the curtain}.
(18) Swiper is really hiding {behind the curtain/behind the toy box}.
(19) There's really a squirrel {behind the toy box/behind the curtain}.
(20) We can see a yellow tail {behind the toy box/under the bed}.
(21) Swiper is {behind the curtain or behind the toy box/behind the door or under the bed}.

Each participant saw 2 practice trials, followed by 3 trials in each of 4 conditions (KNOWLEDGE × BELIEF TYPE). The distribution of true and false sentences was counterbalanced across conditions. Since there were an odd number of trials per condition, the distribution is only fully balanced when both lists are taken together.

<table>
<thead>
<tr>
<th>Sample sentence</th>
<th>Belief Type</th>
<th>Sent. Truth</th>
<th>Comp. Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boots thinks that Swiper is behind the curtain.</td>
<td>TB</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>Boots thinks that Swiper is behind the toybox.</td>
<td>TB</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>Dora thinks that Swiper is behind the toybox.</td>
<td>FB</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>Dora thinks that Swiper is behind the curtain.</td>
<td>FB</td>
<td>F</td>
<td>T</td>
</tr>
</tbody>
</table>

Table 1 Experiment 1: Target sentence types.
Belief Type (TB = true belief; FB = false belief), truth of the sentence, and truth of the complement clause.
2.1.4 Procedure
Sessions took place in a quiet room with the child seated in front of a laptop. The experimenter sat alongside the child, operating the puppet with one hand and coding responses with the other. Sessions were videotaped so that children’s responses could be coded later by an independent viewer.

The experimenter began by explaining the task, introducing the puppet (“Drog”, a baby dragon who wants to learn how to play hide-and-seek), and obtaining the child’s assent to participate. To ensure that the child was comfortable telling the puppet whether he was right or wrong, the experimenter asked the puppet to label a few objects, and prompted the child to say whether the puppet was correct. Once the child had produced at least two yes and two no responses, the experimenter continued with the experiment.

In each trial, the child watched the animated video alongside the puppet. After the story, the puppet uttered the target sentence. The experimenter prompted the child to judge the sentence by asking, Is Drog right? For the two practice trials (which included only filler sentences), the experimenter provided feedback if the child responded incorrectly. The form of the feedback was flexible, but often involved pointing out relevant parts of the scene, repeating parts of the story, or modeling the correct response. After the practice trials, the experimenter did not provide feedback. In general, the experimenter reacted to the child’s response by giving feedback to Drog: Good job, Drog—you got it right! or Silly Drog, you got that one mixed up!

The filler sentence for each trial was chosen based on the child’s response to the experimental sentence. If the child accepted the sentence, a false filler was chosen; if the child rejected it, a true filler was chosen.

2.1.5 Data analysis
Children’s responses were coded online by the experimenter and again from the video recording by a different person. Responses were coded as yes, no, I don’t know, or unclear. Only clear yes or no responses that were never revised were counted in accuracy rates. Video coders rejected trials in cases of experimenter error (3 out of 1420 trials), or when the child was clearly not attending or distracted (28 trials). Since most of the 4-year-old participants had fragile attention spans, coders only rejected trials in extreme cases where the child was out of her chair or talking over the story.

Accuracy rates for truth-value judgments were first analyzed separately for children in the 1-seeker and 2-seeker conditions, using logistic mixed effects models with fixed effects for Belief Type and Knowledge and the maximal by-subject random effects structure (random by-subject intercepts and random by-subject slopes for both main effects and the interaction). Binomial tests were used to compare accuracy to chance levels.

Accuracy in the 1-seeker and 2-seeker conditions was compared using a model with a fixed effect for Number of Seekers as well as Belief Type and Knowledge. This model had the same by-subject random effects structure as the previous models, since the Number of Seekers did not vary within individual subjects.
2.2 Results

2.2.1 Filler accuracy
The fillers were designed to be easy to judge so they could be used as a criterion to exclude participants who could not understand or attend to the task. 4 participants who had accuracy rates below the predetermined cutoff of 65% were excluded from analysis. For the remaining 32 participants, filler accuracy ranged from 67% to 100% (mean = 87%, median = 90%). After exclusions, there were 16 participants each in the 1-seeker and 2-seeker conditions.

2.2.2 Accuracy on think sentences
See Table 2 and Figure 2 for a summary of results.

In the 1-seeker condition, there was a significant main effect of BELIEF TYPE ($\beta = 0.499$, Wald’s $z = 2.45$, $p < .05$): children were more accurate with true belief than false belief. There was also a significant interaction between BELIEF TYPE and KNOWLEDGE ($\beta = 0.698$, Wald’s $z = 2.73$, $p < .01$): the asymmetry based on BELIEF TYPE only held in the knowledge condition (as expected, since in the ignorance condition the belief type is in fact unknown). In knowledge stories, children were highly accurate in the true belief condition (83%: above chance, $p << 0.001$), and inaccurate in the false belief condition (36%: marginally below chance, $p = 0.08$). In ignorance

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Belief Type</th>
<th>1-seeker</th>
<th>2-seeker</th>
</tr>
</thead>
<tbody>
<tr>
<td>knowledge</td>
<td>true belief</td>
<td>83%**</td>
<td>89%**</td>
</tr>
<tr>
<td></td>
<td>false belief</td>
<td>36%°</td>
<td>52%</td>
</tr>
<tr>
<td>ignorance</td>
<td>true belief</td>
<td>69%*</td>
<td>81%**</td>
</tr>
<tr>
<td></td>
<td>false belief</td>
<td>56%</td>
<td>82%**</td>
</tr>
</tbody>
</table>

Table 2 Experiment 1: Accuracy rates by condition.
Stars indicate that the accuracy rate was different from chance: ° $p < 0.1$, * $p < 0.05$, ** $p < 0.001$.

Figure 2 Experiment 1: Accuracy rates by condition.
Error bars represent 95% confidence intervals based on the binomial distribution.
(collapsing across Belief Type), children were just above chance (62%, $p = 0.02$). There was no significant difference in accuracy between true belief (69%) and false belief (56%) sentences.

In the 2-seeker condition, there was again a significant interaction between Belief Type and Knowledge ($\beta = 0.822$, Wald’s $z = 2.20$, $p < .05$). In knowledge stories, children were highly accurate in the true belief condition (89%; above chance, $p << 0.001$), but no different from chance in the false belief condition (56%, $p = 0.9$). In ignorance stories, children were significantly above chance (82% overall, $p << 0.001$), with no significant difference between true belief (81%) and false belief (82%) sentences.

In the model for all the conditions together with Number of Seekers added as an additional fixed effect, there was a significant main effect of Belief Type ($\beta = 0.459$, Wald’s $z = 2.61$, $p < .01$), and an interaction between Belief Type and Knowledge ($\beta = 0.621$, Wald’s $z = 3.55$, $p < .001$). There was also a significant main effect of Number of Seekers ($\beta = 0.614$, Wald’s $z = 2.66$, $p < .01$), but no interactions between Number of Seekers and any other factor. Thus, the overall pattern of responses was similar across the 1-seeker and 2-seeker stories, but children were more accurate overall with the 2-seeker stories.

### 2.3 Discussion

Children were highly accurate in the true belief condition, inaccurate in the false belief condition, and somewhere between in the ignorance condition. Their accuracy improved across conditions when the story involved two seekers instead of one, introducing a conflict of belief that highlighted the relevance of belief in the discourse context.

The most important finding for the pragmatic hypothesis is the improved performance in the 2-seeker condition. We conclude that the heightened relevance of belief in the context helped children access a belief-based rather than a reality-based speaker meaning for the belief report. Although all of the hypotheses predicted the difference in performance between the true belief and false belief conditions—that is what they were designed to do, after all—the middling performance in the ignorance condition is potentially informative. The intention of the ignorance condition was to eliminate the conflict between the characters’ belief and the child’s belief, which should make belief attribution easier. However, children showed lower accuracy in this condition than in the true belief condition. What made it more difficult? If children were attempting to evaluate the complement clause against reality—as part of a deviant literal meaning for the sentence or an inappropriate pragmatic reading—the ignorance condition would be confusing, because it makes that evaluation impossible. Thus, the mild difficulty in the ignorance condition might provide some support for either of the linguistic hypotheses over the conceptual hypothesis.

Although the results of Experiment 1 demonstrate that children are sensitive to context when interpreting belief reports, and can provide more adult-like responses in some situations, it does not demonstrate that children compute the correct literal meaning for belief reports. We investigate this question in Experiment 2.
3 Experiment 2: Truth conditions for think

The goal of Experiment 2 was to determine whether children are capable of rejecting a belief report based on their knowledge of its literal meaning. We manipulated the literal truth of the belief report as a factor, rather than merely counterbalancing it as in our and others’ previous studies. If children can evaluate belief reports based on their literal meaning, they should always reject sentences that are literally false. They may evaluate literally true sentences based on the truth of the complement clause. Thus, children’s accuracy should be higher for literally false than literally true sentences.

In Experiment 2 we also tested a wider age range of children in order to investigate which conditions children show improvement in over the course of development.

3.1 Methods

3.1.1 Participants
65 children aged 3 years (3;1) to 4 years, 2 months (4;2) participated in the study (3.1-4.2 years, mean = 3.6, 26 girls). Participants were recruited from the Infant Studies Database at the University of Maryland. All participants were typically-developing monolingual English-speakers. Data from an additional 8 children (3.1-4.0 years, mean = 3.5, 3 girls) were excluded because they could not complete the task.

3.1.2 Design and materials

3.1.2.1 Truth-value judgment task
As in Experiment 1, children were presented with stories about hide-and-seek, and asked to judge target sentences containing think. All of the stories contained 2 seekers, to give younger children the best possible chance for adult-like responses.

To determine whether children are able to evaluate belief reports against the character’s beliefs instead of reality, we manipulated the literal truth of the sentences as a factor (rather than simply counterbalancing it, as we did in Experiment 1). We collapsed Belief Type and Knowledge into a single Belief Type factor with 3 levels: true belief, false belief, and unknown.

We used the same 14 stories from Experiment 1. Rather than presenting the stories in animated videos, we illustrated each stories with a series of 8-9 still images.

We created two lists of target sentences. In each list, the order of sentences with respect to Belief Type and Literal Truth was pseudo-randomized. Two filler sentences, one true and one false, were created for each story, using the same templates as in Experiment 1. Each participant saw 2 practice trials, followed by 2 trials in each of 6 conditions (Literal Truth × Belief Type).

3.1.2.2 False belief task
In addition to the truth-value judgment task, most children also completed two trials of a standard change-of-location false belief task. The story was acted out by the experimenter using toys. The story for one of the trials was as follows:
This story is about Toby and his dad. Toby is playing with his cowboy hat and pretending to be a cowboy. He’s having a great time. Then he decides to go outside and play, but he doesn’t want his cowboy hat to get dirty. So he puts it under the bucket where he can find it later. [Toby leaves the scene.] While Toby is playing outside, his dad comes in to clean up his room. He finds the cowboy hat under the bucket. He says, “Hey, this doesn’t belong here! I’m going to put it in the toy box where it’s supposed to be. ...There. Much better.” [Toby’s dad leaves the scene.]

After the story, children were asked the following series of questions.

(22)  Pre-test memory questions:
    a. Where did Toby put the cowboy hat (before he went outside)?
    b. Where is the cowboy hat now?
(23)  Test question: Toby is coming back inside, and he wants to play with his cowboy hat again. He remembers where he put it. Where is Toby going to look for the cowboy hat first?
(24)  Justification: Why will he look there?
(25)  Post-test memory question [depending on child’s answer to Test question]:
    a. Correct test: Where is the hat really?
    b. Incorrect test: Where did Toby put the hat before he went outside?

3.1.3 Procedure

Sessions took place in a quiet room. For the truth-value judgment task, the child was seated in front of an iPad. The experimenter sat alongside the child, operating the iPad. Sessions were videotaped so that children’s responses could be coded later by an independent viewer.

The experimenter began by explaining the task, introducing a little boy who appeared on the screen next to the story: This little boy is only 2 years old, so he doesn’t know how to play hide and seek. We’re going to try to help him learn to play. After each story, he’s going to try to say what happened in the story, and it will be your job to tell him if he’s right or wrong. After obtaining the child’s assent to participate, the experimenter continued with the experiment.

In each trial, the experimenter narrated the story, swiping the screen to display each scene, then delivered the target sentence (in the voice of the little boy). The experimenter prompted the child to judge the sentence by asking, Did the little boy get it right? As in Experiment 1, the experimenter provided feedback for the two practice trials, but not the experimental trials. The filler sentence for each trial was chosen based on the child’s response to the experimental sentence. If the child accepted the sentence, a false filler was chosen; if the child rejected it, a true filler was chosen. The experimenter recorded the child’s responses using buttons on the iPad.

The false belief task always came after the truth-value judgment task. The iPad was removed, and replaced with the toys for the stories. If the child initially provided incorrect answers for the pre-test memory questions, the experimenter retold the story until the child responded correctly. No feedback was provided for the test question or post-test memory questions.
3.1.4 Data analysis
Children’s responses were coded online by the experimenter and again from the video recording by a different person.

Responses for the truth-value judgment task were coded as yes, no, I don’t know, or unclear. Only clear yes or no responses were counted in accuracy rates. Video coders excluded responses in cases of experimenter error (2 out of 1396 responses) or when the child was clearly not attending or distracted (1 response). As in Experiment 1, coders were conservative, only rejecting trials in cases of extreme and obvious inattention.

For the false belief task, all children provided correct answers on the pre-test memory questions on the first or second try. Children received a score of 1 for the trial if they provided correct answers for both the test question and the follow-up memory question. Each child was given a total FB score of 0, 1, or 2 for the number of correct trials.

Accuracy rates for truth-value judgments were analyzed using logistic mixed effects models with fixed effects for Belief Type, Literal Truth, and the subject’s age, as well as a random by-subject intercept and random by-subject slopes for Belief Type, Literal Truth, and their interaction. All factors were coded orthogonally. The 3-level factor Belief Type was coded as two contrast variables: the first compared true belief to false belief; the second compared false belief to unknown.

To determine whether children’s performance on the standard false belief task was predictive of their understanding of think sentences, we used two different models. The first added children’s FB Score as a fixed effect to the original model in place of the age effect. The second was a model of accuracy in the truth-value judgment task of the false belief trials alone. The model had fixed effects for Literal Truth and FB Score, as well as a random by-subject intercept and a random by-subject slope for Literal Truth.

3.2 Results

3.2.1 Filler accuracy
17 participants who had accuracy rates below the predetermined cutoff of 65% were excluded from analysis. These participants were distributed over the full age range (3.2-4.0 years, mean = 3.6, 6 girls). For the remaining 48 participants (3.1-4.2 years, mean = 3.6, 25 girls), filler accuracy ranged from 71% to 100% (mean = 86%).

3.2.2 Truth-value judgment task
See Table 3 and Figure 1 for summaries of the results.

There were main effects for both Belief Type contrasts: accuracy was significantly higher in the true belief compared to the false belief condition ($\beta = 0.751$, Wald’s $z = 3.01$, $p < .01$), and marginally higher in the unknown compared to the false belief condition ($\beta = -0.344$, Wald’s $z = 0.669$, $p = .08$). There was also a significant interaction between the first Belief Type contrast and Literal Truth ($\beta = 1.11$, Wald’s $z = 4.68$, $p < .0001$): in the true belief condition accuracy was higher for true (88%) than false (68%) sentences, while in the false belief condition accuracy was higher for false (82%) than true (39%) sentences.
There was no significant effect of the participant’s age, whether it was entered as a continuous variable or as a 2-level categorical factor based on a median split (about 3;8).

3.2.3 False belief task
44 of the 48 participants in the analysis completed the false belief task. Children were grouped by their FB score (see Table 4). There was no significant difference in age between the three groups (one-way ANOVA, $p = 0.29$).

There were no significant effects of FB score on overall accuracy in the truth-value judgment task, or on the false belief trials alone.

<table>
<thead>
<tr>
<th>Belief Type</th>
<th>Literal Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T$</td>
</tr>
<tr>
<td>true belief</td>
<td>78%***</td>
</tr>
<tr>
<td>false belief</td>
<td>60%**</td>
</tr>
<tr>
<td>unknown</td>
<td>66%***</td>
</tr>
</tbody>
</table>

Table 3  Experiment 2: Accuracy rates by condition.
Stars indicate difference from chance: * $p < .05$, ** $p < .01$, *** $p < .001$

![Figure 1](https://example.com/image.png)

Figure 1  Experiment 2: Accuracy rates by condition.
Error bars represent 95% confidence intervals based on the binomial distribution.

<table>
<thead>
<tr>
<th>FB Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of children</td>
<td>16</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Mean age in months (sd)</td>
<td>43.3 (3.5)</td>
<td>43.0 (2.7)</td>
<td>44.6 (4.3)</td>
</tr>
</tbody>
</table>

Table 4  Experiment 2: False Belief scores.
3.3 Discussion
The pattern of children’s responses in this experiment demonstrates that they understand the literal meaning of belief reports. They correctly reject literally false sentences and accept true ones, with only one exception: they tend to reject true sentences in false belief scenarios. The contrast in children’s performance with true and false sentences in the false belief condition suggests that their well-known difficulty evaluating belief reports in false belief contexts is not due to an incorrect semantic or conceptual representation of belief. 3-year-olds are sensitive to the literal meaning of belief reports, and are able to evaluate whether the subject holds the stated belief. Rather, their difficulty with true sentences in false belief scenarios is consistent with the pragmatic hypothesis. Children may often default to a speaker meaning in which the complement clause is the main point of the utterance, and the speaker endorses its truth.

Children’s accuracy was not predicted by their age, suggesting that the literal meaning of think is in place by around 3 years of age, if not earlier. Furthermore, children’s accuracy was not predicted by their success on a traditional false belief task. Thus, the acquisition of belief reports is not dependent on the skills needed to pass a false belief task.

It is possible that children’s performance on the false belief task and their interpretation of true sentences in the false belief condition are both related to their tendency to assume that the Question Under Discussion involves reality instead of beliefs. We will discuss this possibility in the next section. However, we should be cautious about interpreting the results from the false belief task, since participants completed it after having been exposed to twelve belief reports during the truth-value judgment task. Although children did not receive any feedback on their responses, it is possible that this more concentrated exposure to mental state language would have affected their behavior on the false belief task. An effect of this sort might explain the surprising fact that there was no difference in age between children who failed the false belief task and those who passed it.

In this experiment as in Experiment 1, children’s accuracy in the unknown condition (the ignorance condition of Experiment 1) was above chance but middling compared to accuracy in the true belief condition. Accuracy in this condition, like the others, did not change with age. Middling accuracy in this condition might be expected if many of the children interpret the complement clause as the main point, since the truth of the complement clause in reality cannot be evaluated. However, the pragmatic hypothesis predicts that children should be able to reject literally false sentences in this condition just as well as in the false belief condition. One possible explanation for the difference in accuracy (82% vs. 71%) is that the relevance of beliefs in the story is greater in the knowledge conditions, so children track them more carefully. Anecdotally, in the knowledge conditions children often made comments during the story about which seeker got it right or wrong. In the ignorance condition, on the other hand, children were focused on figuring out where the hider was; their spontaneous comments were mostly about their own guess about the hider’s location (although we did try to prevent participants from committing to a guess, and excluded any trials where they explicitly did so). They didn’t care much about what the seekers thought, because they had the impression that the seeker’s beliefs were just as arbitrary as their own in this case. Thus, counterintuitively, removing reality from the equation may have reduced children’s attention to the beliefs in the story, rather than enhancing it. They are less prepared to evaluate a belief report because they have not tracked the beliefs as carefully. Although it should be possible to reconstruct the beliefs by looking at the scene, since the
seekers stand next to the location they guessed, some children may not go to the trouble, and simply guess.

Another result that deserves more scrutiny is that children’s accuracy on false sentences in the true belief condition was lower than for true sentences (68% compared to 88%). Even some of the oldest children in the study answered incorrectly in this condition. Although we might not be surprised in general when children have more difficult rejecting sentences than accepting them, this pattern stands in contrast to the other conditions in this study, where children’s performance was as good or better for false sentences. Since this result was not predicted by any of our competing hypotheses, we can only speculate as to its source. Let’s consider a sample scenario and sentence. Suppose Swiper is really hiding behind the curtain, and Boots also thinks he is (see Figure 2). A false sentence in the true belief condition would be (26).

(26) Boots thinks that Swiper is behind the toy box.

The sentence incorrectly attributes a false belief to Boots: Boots doesn’t think that Swiper is behind the toy box, and in fact Swiper is not behind toy box. One might have expected the double falsity of the sentence to make it easier for children to reject: there’s nothing temptingly right about it. Since the other seeker holds the opposite belief, a contrasting true proposition is readily available regardless of whether the subject or the complement clause is taken to be in focus, as demonstrated in (27)-(28). However, this double falsity might have confused some children: they might have found it odd that the speaker would be so wrong about what happened in the story, and second-guessed their own knowledge.

(27) No, DORA thinks that Swiper is behind the toybox.
(28) No, Boots thinks that Swiper is behind the CURTAIN.

Although the more unexpected findings from Experiment 2 warrant additional research, the main conclusions are strong. Children are capable of evaluating the literal meaning of belief reports, as

Figure 2  Experiment 2: Sample scene.
evidenced by their rejection of literally false sentences even in the false belief condition. However, they tend to assume that the speaker meaning has to do with reality, as evidenced by their reality-based evaluation of literally true sentences.

4 General Discussion

We have argued that children’s non-adult-like interpretations of belief reports should be attributed to a pragmatic problem, rather than a syntactic/semantic or conceptual deficit. Children are less able than adults to determine what a speaker means by uttering a belief report because they do not always know that beliefs are under discussion in the discourse. Their default assumption is that beliefs are not relevant. This “setting” for the default could be attributed to children’s general difficulty tracking beliefs or to the low frequency of conversations about belief states in their experience, or both.

An alternative explanation for the results is that the pattern of children’s judgments is generated by a semantically-encoded endorsement reading, rather than a pragmatic-level speaker meaning. That is, children’s semantic representation for think is something like think correctly. This is a tempting hypothesis: if children constantly hear uses of belief reports in which the speaker endorses the truth of the complement clause, why shouldn’t they conclude that think means think correctly?

There are differences between the truth and felicity conditions of think correctly and those generated by a pragmatically-derived endorsement interpretation, but they are subtle. To explain them, we first have to decide what we mean by think correctly. There are no verbs in English which simultaneously assert both an attitude attribution and the speaker’s perspective on the truth of the content of the attitude. One of these meaning components is always presupposed, that is, taken by the speaker to be shared background information. For a meaning close to think correctly, we have two options: know and be right. The know sentence in (29) asserts that Dora believes that Swiper is behind the curtain, and presupposes that Swiper actually is behind the curtain. The be right sentence in (30) asserts that Swiper is behind the curtain, and presupposes that Dora believes it. Of course it is possible to assert both parts simultaneously, but for that you need two words: think correctly. The truth conditions for each of these three ways of encoding the meaning are represented in the truth table in Table 5, along with those of a pragmatic endorsement interpretation.

(29) Dora knows that Swiper is behind the curtain.
(30) Dora is right that Swiper is behind the curtain.

We can test whether children treat part of the meaning as presupposed by seeing how they interpret negated sentences. Dudley, Orita, Hacquard, and Lidz (to appear) have demonstrated that three-year-olds do not treat think like know—they know that think does not presuppose its
complement. In a setup where a toy is hidden in either a red or a blue box, children understand that in (31), the toy must be in the blue box, while in (32) it is likely to be in the red box.³

(31) Lambchop doesn’t know that the toy is in the blue box.
(32) Lambchop doesn’t think that the toy is in the blue box.

Thus, we can eliminate the possibility that children treat think as know, but thus far we have no conclusive evidence that they do not treat think as either be right or think correctly. However, there are at least two pieces of evidence that speak against these possibilities. First, children’s sensitivity to context—their improved performance in the 2-seeker condition in Experiment 1—is only expected under a pragmatic account. Second, children use think with an uncertainty implication quite often in their own production (Shatz, Wellman, & Silber, 1983; Bloom, Rispoli, Gartner, & Hafitz, 1989; Diessel & Tomasello, 2001). This reading would be impossible if children assumed that the literal meaning of think was think correctly or be right. Indeed, children would have to be quite ignorant of pragmatic principles to assume that speakers frequently literally say, “I’m right that…” or “I think correctly that…” Since such statements would not yield any uncertainty implicatures, they only state redundantly what is already assumed: speakers say things that they believe to be true.

We conclude that the pattern of children’s responses is best explained by pragmatic difficulty. The pragmatic hypothesis may be extended to explain children’s difficulty in traditional false belief tasks as well. Children show the most difficulty in tasks where they must evaluate a statement or question posed to them directly. This setup requires the child to track the experimenter’s perspective and conversational intentions in addition to the perspectives of the characters in the story (Helming, Strickland, & Jacob, 2014). For example, after hearing the story about Maxi and the chocolate described in the introduction, children under 4 years are very poor at answering the direct question in (33) (Wimmer & Perner, 1983). They are better at acting out a

³ Not all 3-year-olds seemed to know that know presupposes the truth of its complement, and would thus pick the red box after a clue like (30). However, all 3-year-olds seem to understand that think doesn’t presuppose the truth of its complement.
response to a prompt like, “What happens next?” (Rubio-Fernández & Geurts, 2012), or looking in response to the experimenter’s “wondering” aloud, as in (34) (He, Bolz, & Baillargeon, 2012). Findings like these suggest that what makes the false belief task difficult is the need to puzzle out the experimenter’s intentions in asking the question. When that need is removed, children are able to demonstrate that they can track false beliefs. Under this broader view of the pragmatic hypothesis, children’s difficulty is not specific to the verb think.

(33) Where will Maxi look for the chocolate?
(34) I wonder where Maxi will look for the chocolate…

Given that young children show variable success in tracking beliefs and difficulty identifying conversations that are about beliefs, it is remarkable that they manage to learn the literal meaning of think by age three, if not earlier. We suspect that children rely heavily on syntactic information to learn the meanings of mental state verbs like think (Fisher, Gleitman, & Gleitman, 1991; Gleitman, Cassidy, Nappa, Papafragou, & Trueswell, 2005; Hacquard, to appear; Papafragou, Cassidy, & Gleitman, 2007; White, Dudley, Hacquard, & Lidz, 2012).

5 Conclusion

We have investigated the “curse of knowledge” in children’s interpretation of belief reports— their tendency to be influenced by reality when judging sentences about false beliefs. We argued that children’s non-adult-like interpretations are not due to non-adult-like conceptual or syntactic/semantic representations, as has been previously proposed. In fact, they seem to know the literal meaning of think by the youngest age we tested (37 months): they are able to reject think sentences that report a belief incorrectly. Children’s difficulty relates to determining the speaker’s meaning in context. They assume that speakers are generally talking about reality, not mental states, and therefore sometimes reject true sentences in false belief scenarios.

6 Acknowledgements

This work was funded in part by NSF grant #BCS-1124338. Many thanks to our undergraduate research assistants (Amber Frazier, Faina Kostyukovsky, Jessica Lee, Sara McVeigh, Laura Sherry, and Leah Whitehill), and the Attitudes and Theory of Mind research group (Rachel Dudley, Kate Harrigan, Morgan Moyer, Naho Orita, and Aaron Steven White).

7 References


