Cognitive Development 46 (2018) 40–50

Contents lists available at ScienceDirect

Cognitive Development

journal homepage: www.elsevier.com/locate/cogdev

Replications of implicit theory of mind tasks with varying representational demands

Lindsey J. Powell[a,*], Kathryn Hobbs[b,c], Alexandros Bardis[b,d], Susan Carey[b], Rebecca Saxe[a]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[b] Department of Psychology, Harvard University, Cambridge, MA, USA
[c] Collaborative for Educational Services, Northampton, MA, USA
[d] Calcot Services for Children, Reading, UK

ARTICLE INFO

Keywords:
Cognitive development
Theory of mind
Infancy
Replication
Social cognition

ABSTRACT

We attempted to reproduce three tests of theory of mind in infants using implicit tasks that have been previously reported in the literature. These efforts were intended as initial steps in larger projects aimed at building on past research to better understand infants' theory of mind capacities and their relationship to preschoolers' explicit theory of mind. One task fully replicated evidence of 2-year-old children's sensitivity to others' ignorance. The results of another task testing for similar capacities in 18-month-old infants also elicited behavior similar to the original findings, although in that case we only conducted one of two conditions critical for demonstrating that performance depended upon theory of mind capacities. In contrast, our violation of expectation tasks failed to reproduce evidence that, by 18 months of age, infants form specific expectations about the actions an agent will engage in on the basis of false beliefs. Instead, looking times were consistent with infants having no clear prediction about the agent's actions under conditions of false belief. We discuss factors that might account for our failure to reproduce the previously reported violation of expectation results on which we were attempting to build. However, we also discuss the consistency of our data with other findings and hypotheses regarding early-developing theory of mind, and consider the possibility that they reflect the veridical abilities of 18-month-old infants, who may track others' knowledge and ignorance but may not consistently represent the contents of others' beliefs.

1. Introduction

For twenty-five years, children's understanding of the inner workings of other minds was studied almost exclusively through what children could say, or predict, explicitly (Flavell, 1999). Children first systematically predict future actions based on false beliefs, explain past actions caused by false beliefs, and talk spontaneously about divergent beliefs, all in the fourth year of life and not before (Bartsch & Wellman, 1995; Perner, Lang & Kloo, 2002; Wimmer & Perner, 1983; Wellman, Cross & Watson, 2001). Although there were always critics, and alternative views (Leslie, 1994; Bloom & German, 2000), the data suggested a fairly coherent and cumulative scientific picture of the child's developing "theory of mind": an earlier stage when preschoolers understood others' perceptions and desires, followed by a later stage when children could also conceptualize meta-representational states like beliefs (Wellman & Liu, 2004; Saxe, Carey & Kanwisher, 2004).

* Corresponding author at: Department of Brain and Cognitive Sciences, 46-4017, 77 Massachusetts Ave, Cambridge, MA 02139, USA.
E-mail address: ljpowell@mit.edu (L.J. Powell).

http://dx.doi.org/10.1016/j.cogdev.2017.10.004
Received 16 February 2017; Received in revised form 21 September 2017; Accepted 4 October 2017
Available online 10 October 2017
0885-2014/

Early evidence that even one-year-old infants understand others' false beliefs was, consequently, a scientific sensation. The first demonstration used infants' looking as a measure of their expectations (Onishi & Baillargeon, 2005). When the infant saw a person reach for an object where it *really was* instead of where the person *falsely believed it to be*, infants looked longer, showing that the reach violated their expectations. Since 2005, many groups, using many different methods, have found evidence for some understanding of others' beliefs in children under three years of age (e.g. Buttelmann, Carpenter & Tomasello, 2009; Kovács, Téglás & Endress, 2010; Knudsen & Liszkowski, 2012; Hamlin, Ullman, Tenenbaum, Goodman & Baker, 2013; Southgate, Senju & Csibra, 2007; Surian, Caldi & Sperber, 2007; Träuble, Marinović & Pauen, 2010; see Scott & Baillargeon, 2017 for a review). In contrast to the many failed attempts to elicit verbal reasoning or predictions about false beliefs from young 3-year-old children, these tasks indirectly assess infants' and toddlers' understanding of others' beliefs or their awareness of the state of the world by measuring the participants' spontaneously generated attention or behavior. Why the apparently conflicting results? One possibility is that infants' and toddlers' ability to think about other minds is limited by available cognitive resources. Answering an experimenter's direct question about hypothetical future actions (explicit tasks) might demand most of a toddler's available cognitive resources, leaving nothing left over for also thinking about false beliefs (Baillargeon, Scott, & He, 2010; Keen, 2003; Cheng & Holyoak, 1985). By contrast, passive viewing and spontaneous responding (implicit tasks) have minimal cognitive demands, so infants and toddlers can reveal richer abilities.

The key challenge now is to build a picture of the child's developing theory of mind that incorporates these new results. If infants and toddlers *can* track others' knowledge, where do the boundaries and limits of their abilities lie? Do infants and 4-year-old children have similar knowledge about others' false beliefs, and differ mainly in their ability to express that knowledge? Or is there developmental change in the scope, flexibility, or content of young children's concepts of beliefs? Relatedly, many implicit tasks are subject to richer and leaner interpretations. Some tasks could be passed by participants tracking only others' knowledge or ignorance (i.e. their "awareness") of an object or event, whereas other findings suggest a much richer capacity for inferring and representing the contents of others' beliefs (Baillargeon et al., 2010; Martin & Santos, 2016; Perner & Ruffman, 2005; Wellman, 2011). And, as in all active areas of research, more specific tests of infants' knowledge have yielded conflicting findings across research groups (Grosse Wiesmann, Friederici, Singer & Steinbeis, 2016; Kulke, Reiß, Krist, & Rakoczy, 2017; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012; Yott & Poulin-Dubois, 2016).

To address these important questions, many researchers, including ourselves, sought to build on the early reports of belief-understanding in 1- and 2-year-old children. Such attempts often involve initial phases in which a laboratory seeks to replicate the findings on which it wishes to build. Here we report the results of four such attempts to reproduce published findings, conducted in anticipation of performing follow up experiments to better understand how infants' early theory of mind, measured by implicit tasks, is related to their performance on explicit tasks as preschoolers. Our original plan was not to publish standalone replications – successful or failed – but to publish the results of additional, original experiments. The data are presented here in the same spirit: to help future researchers who may wish to build on existing studies.

We conducted close replications of studies by Knudsen and Liszkowski (2012) and Buttelmann et al. (2009), as well as two experiments based on published violation of expectation paradigms (Onishi & Baillargeon, 2005; Song, Onishi, Baillargeon & Fisher, 2008; Träuble, Marinović & Pauen, 2010), with procedural changes based on both our novel experimental goals and the original authors' advice. We found evidence consistent with the claim that infants and toddlers routinely track others' awareness of the state of the world. Across two large, well-powered samples, however, we failed to replicate evidence that infants form action expectations based on specific false beliefs. We also tested the possibility that variable performance reflected a real individual difference between infants who do, versus those who do not, understand false beliefs, and found no evidence for this hypothesis either. Both practical and theoretical lessons can be learned from these experiments.

2. Experiment 1

Several of our attempts at reproducing prior findings were components of a study testing if theory of mind performance on implicit tasks in infancy shares two features of explicit theory of mind performance in older children: coherent individual differences across tasks (Wellman & Liu, 2004) and a relationship to individual differences in executive function (e.g. Carlson & Moses, 2001). The study included measures of working memory and inhibitory control, as well as two implicit theory of mind measures: one interactive helping task and one violation of expectation (VOE) looking time task. Because our goal was to measure individual differences in the ability to track others' mental states, for both task types we ran only trials that tested for the attribution of false belief or ignorance, and not control trials in true belief conditions.

We tested an initial version of the VOE task by collecting a large, well-powered sample (over 2.5 times larger than the samples in the studies on which it was based; Simonsohn, 2015), and then used a revised version of this task in the main experiment on the relationship between implicit theory of mind task performance and executive functions. These two experiments are reported separately below as Tasks 2 and 3. Many of the participants in the initial VOE task also completed the interactive helping task, and the procedure for the helping task was not changed when we transitioned to the main experiment. In the Task 1 report below, we have thus collapsed across all participants tested with this common procedure for the sake of power, but we also discuss how performance on this task related to looking times elicited by the two VOE procedures when presenting Tasks 2 and 3.

2.1. Task 1: Interactive helping

Interactive tasks test infants' or toddlers' capacity to represent others' awareness or beliefs by observing how participants choose

to interact with an experimenter who was either present or absent when something about the world changed. For instance, if an infant is asked to help an experimenter, he first has to infer what the experimenter's goal is. Goal attribution relies on information about what others perceive and know about the world (Baker, Saxe & Tenenbaum, 2009; Hamlin et al., 2013; Luo & Baillargeon, 2007), so if infants can represent others' awareness or beliefs, then they may infer different goals, on the basis of the same actions, for individuals who do or do not hold mistaken beliefs about current circumstances, leading to distinct helping behaviors.

Our interactive task was based closely on a study by Buttelmann et al. (2009). They asked if 16-, 18-, and 24-month-old infants' attempts to help an experimenter would be influenced by the experimenter's beliefs. The experimenter was attempting to open a locked box, struggling with a locking mechanism that participants had been trained to master. The experimenter either did or did not know that a toy she had placed in that box had been moved to an adjacent box, which was also locked. Buttelmann and colleagues found that the way in which participants chose to help was influenced by this factor. If the experimenter *had* seen a confederate move the toy, participants were more likely to help her open the box she was acting on and presumably knew was empty. In contrast, if the experimenter had not seen the toy moved, participants were more likely to help by opening the box the she did not act on, which now contained the toy.

The difference in performance between the two conditions demonstrates that the participants were sensitive to the experimenter's knowledge state, but it is open to two interpretations. One (rich) possibility is that when the experimenter was away while the confederate moved the toy, infants attributed to her the specific, false belief that the toy remained in the original box. Another (leaner) possibility is that infants in the "false belief" condition only recognized that the experimenter was (after the transfer) unaware of toy's location. Without the ability to reference a specific belief to guide goal attribution, infants observing the experimenter pull on the empty box could plausibly have inferred that her goal was to find the toy inside either box. (Other commentators have also argued that the differences in E1's actions in the two conditions, e.g. seeing the toy moved without protesting or approaching, make toy retrieval a more plausible goal in the false belief than the true belief condition [Allen, 2015; Priewasser, Rafetseder, Gargitter & Perner, 2017].

When selecting this task for our battery, we were aware of this ambiguity, but did not intend to resolve it. Instead we simply hypothesized that helping in this context could be a measure of the robustness of infants' use of whatever capacity they had to represent others' epistemic relationship to the state of the world. We ran only the false belief condition (FB) from this between-subjects design in our battery, as the true belief (TB) condition does not test for the ascription of either false beliefs or ignorance and the task is not easily adaptable to a within-subjects design.

2.1.1. Participants

Ninety four 17–20-month-old infants participated and were included in the sample (35 female; age range: 17 months, 6 days – 20 months, 13 days; mean age: 18 months, 18 days). An additional 34 infants were tested but excluded from analyses, 30 for declining to approach and help the experimenter within the allotted time limit (see below), and four for experimenter error.

2.1.2. Procedure

When replicating this task, we adhered as closely as possible to the original methods as describe by Buttelmann et al. (2009). Two experimenters sat together on the floor of a testing room with the participant and his or her parent. They engaged the participant in free play for several minutes. Experimenter 1 (E1) then approached two boxes (approximately 1′ × 1′ × 1′) placed side by side near one wall of the room. The participant watched E1 lift the lid and look inside each box twice. She then said "Bye bye" and left the room. Experimenter 2 (E2) then invited the participant to approach the boxes and showed the participant how to lock them by inserting large, wooden pins into holes that went through both the front lid and a portion of the interior of the box, preventing the lid from being lifted. After encouraging the participant to practice locking and unlocking the boxes, E2 unlocked both boxes, placed the pins on the floor, and guided the participant to resume playing on the other side of the room.

E1 then returned with a small, stuffed tiger, which she showed to the participant. After several minutes of playing with the tiger, E1 took the tiger and sat behind and between the boxes, facing the participant. She looked at each box and then chose one, opening its lid, placing the tiger inside, and closing it. E1 then said "Bye bye" and left the room again. E2 looked at the participant with a "sneaky" expression, then tiptoed to the boxes and sat behind them, facing the participant. She took the tiger from the box it was in and moved it to the other box, then used the two pins to lock both boxes. Then E2 went back to the other side of the room and resumed playing with the participant. E1 came back into the room and sat behind the two boxes facing the participant. She looked at both boxes and then tugged on the lid of the box in which she had left the tiger, which did not open due to the lock. She tugged on the box a second time and then sat back with a confused expression, looking either at the participant or the floor between the boxes. If the participant did not approach and unlock one of the two boxes in the 10 s after this action sequence, E1 then asked, "Can you help me?" E2 also encouraged the participant saying, "Go on, help her!" If the participant did not help after an additional waiting period, then E2 signaled the parent to encourage the participant using the same language ("Go on, help her!"; the parent was instructed before the study *not* to tell the participant to find the toy). If the participant did not help within 2 min, they were excluded from data analysis, similar to original exclusion criteria reported by Buttelmann et al. (2009).

2.1.3. Data analysis

The experimenters recorded which box the participant touched and/or opened first: the one representing the tiger's former location, which E1 attempted but failed to open, or the one representing the tiger's current location. In Buttelmann and colleagues' original experiment, they compared the number of children who approached each of these two boxes in an FB condition like the one conducted here with the results of a TB condition, in which E1 knew that the former location box was empty but still tried to open it.

As we did not conduct a TB condition, we could not conduct such a comparison. Instead, we compared the proportion of participants who chose the tiger's current location to chance using a binomial test. Although it is unlikely that the null hypothesis (i.e. what infants would do if they did not represent E1's false belief) would be best represented by a random choice between the two locations, the original study also conducted binomial tests as a secondary analysis, and found that both 16- and 18-month-olds in the FB condition were more likely to open the current location box than would be expected by chance. We could thus assess the replicability of that result with our sample.

2.1.4. Results

Fifty-nine of the 94 participants (62.8%) first touched or opened the box containing the tiger; the remaining 35 first touched or opened the empty box E1 acted on. This distribution differed significantly from chance ($z(93) = 2.37$, $P < 0.05$). Thus, as in Buttelmann and colleagues' original study (2009), 1.5-year-old infants were more likely to help not by opening the box the experimenter was pulling on, but by retrieving the toy the experimenter falsely believed to be in that box. As in the original experiment, this may reflect infants' understanding that the experimenter is unaware that the toy is no longer in the box she left it in. A critical caveat in the current case is that we did not run the TB condition and show that infants would behave differently if the experimenter knew the box she pulled on was empty. In the original study, and in some replications (e.g. Fizke, Butterfill, van de Loo, Reindl & Rakoczy, 2017), the percentage of infants in the TB condition who opened the current location box was below 50%, so comparison to chance may not represent a relaxed standard. However, at least one group observed approach to the current location box slightly more than 50% of the time in the TB condition. (Approach to the current location was still more frequent in the FB than the TB condition; Priewasser et al., 2017.)

2.2. Task 2: Violation of expectation, version 1

Violation of expectation tasks rely on the tendency of infants to look longer at events they find surprising or inconsistent with expectations guided by preceding events (Aslin, 2007). VOE experiments designed to measure false belief attribution generally have in common a situation in which an experimenter, who has established a pattern of reaching for or approaching a particular object, develops a false belief about the object's location. In separate trials measuring infants' attention, either between or within participants, the experimenter then reaches for the object either where she ought to think it is or where it is actually located. Looking times to these two trial types are then compared (e.g. Onishi & Baillargeon, 2005; Surian et al., 2007; Song et al., 2008; Scott & Baillargeon, 2009; Träuble et al., 2010; Scott, Baillargeon, Song & Leslie, 2010; for a review see Baillargeon et al., 2010).

One advantage of these paradigms is that they have the potential to differentiate between accounts of implicit theory of mind capabilities in which infants represent the contents of others' false beliefs and accounts which hold that infants track others' knowledge versus ignorance of current circumstances, but fail to maintain representations of false beliefs that could support specific behavioral predictions. The former accounts predict that infants will be surprised, and thus look longer, when someone reaches for an object where it ended up relative to events in which the person reaches where she falsely believes the object to be. In contrast, the latter account is consistent with infants simply failing to make a behavioral prediction, and thus looking equally at different reaching outcomes when observing someone they believe to be unaware of the current state of affairs. Both predictions can be differentiated from the expectation that a person will reach to the current location of a desired object, either always or when knowledgable about its location.

A number of VOE experiments have found data consistent with the hypothesis that infants generate specific expectations about actions based on false beliefs (i.e. that the actor will reach where she thinks an object is; e.g. Onishi & Baillargeon, 2005; Song et al., 2008; Träuble et al., 2010). However, it is worth noting that some studies have yielded results that are more consistent with the conclusion that 1-year-old infants represent an actor's awareness, or ignorance, regarding the status of an object, but do not routinely make specific predictions on the basis of false beliefs (e.g. Poulin-Dubois & Yott, 2017; Surian et al., 2007; Yott & Poulin-Dubois, 2016) and may even fail to use true beliefs to predict behavior when the actor's access to information about the situation is intermittently disrupted (Sodian & Thoermer, 2008). Similar findings have led some comparative researchers to conclude that non-human primates possess an understanding of others' awareness, but not a representational theory of mind capable of representing specific false beliefs (Marticorena, Ruiz, Mukerji, Goddu & Santos, 2011; Martin & Santos, 2016).

Our initial VOE task was based on the 'FB-green' condition from Onishi and Baillargeon's study (2005).[1] In the critical trial of this condition, an experimenter was absent, hidden behind closed, opaque doors, when a watermelon she had placed into one box moved along a track in the floor of the stage to a second box. When the experimenter returned, she reached either into the box in which she had left the watermelon (an 'expected' trial) or the box to which it had moved (an 'unexpected' trial), in a between subjects design. Fifteen-month-old infants who observed an unexpected trial looked longer than those who observed an expected trial.

Our version involved several small changes to the perceptual aspects of the displays and several procedural differences introduced in order to conduct a within-subjects comparison between expected and unexpected looking times, as we reasoned that this contrast would provide a better individual measure of false belief reasoning than an unexpected trial alone. (See the SI for a list of all differences.) Our tasks were thus not direct replications of Onishi and Baillargeon's study. However, all procedures were consistent with other published studies that are considered to provide evidence for infant false belief ascription (e.g. Surian et al., 2007; Träuble

[1] This was preceded by a pilot study with a smaller sample attempting to directly replicate Träuble and colleagues (2010). See the Supplementary Information (SI) for details.

et al., 2010). Thus there are empirical grounds for expecting infants' preference for unexpected trials, in which an actor behaves in a manner inconsistent with her false belief, to be robust to these changes.

2.2.1. Participants

Forty-three 17–20-month-old infants participated in the first VOE task (18 female, age range: 17 months, 6 days – 20 months, 0 days; mean age: 18 months, 11 days). A sample this size provides power of 0.96 for detecting a within-subjects effect of the size reported by Träuble and colleagues (Cohen's $d = 0.58$), and of 0.99 for detecting a between-subjects effect of approximately the size reported by Onishi and Baillargeon (Cohen's $d = 1.4$, estimated from figure). Thirty-two of these infants also provided data in the interactive helping task. Nine of the eleven excluded did not help within the time limit, and two were excluded for experimenter error. An additional two infants were recruited for the VOE task but excluded from analysis due to fussiness or inattentiveness.

2.2.2. Procedure

Participants were seated on their parent's lap, approximately 2 feet from the front of stage that could be covered by an opaque, black screen. An orange and a purple box were positioned on the stage so that open, fringe-covered sides on each box faced one another. Experimenter 1 (E1), wearing a hat that allowed the participant to see the position of her head and gaze but not to make eye contact, sat at the back of the stage, in a large opening that could be covered by a curtain, obscuring E1 from view. A toy tomato was positioned in a track that extended between the two boxes, and could be moved either by E1 in view of the participant, or by a second experimenter (E2) pulling the tomato along the track from underneath the stage, making the tomato appear to move on its own.

Two familiarization trials each began with the toy tomato sitting between the two boxes. E1 looked at the tomato, moved it into one of the two boxes, and then closed the curtain, disappearing from view. After a brief pause, she opened the curtain, reached into the box with the tomato and slid it back to the center. She briefly held the tomato in that position and then slid it into the other box. She paused with her arm outstretched, holding the toy in the box, until the participant looked away for 2 consecutive seconds or reached a maximum cumulative looking time of 30 s. These thresholds were based on those used by Onishi and Baillargeon (2005) and chosen in advance of data collection. Looking time was measured by an online coder in another room, who communicated when a threshold was reached via walkie talkie. E2 then lowered the screen in front of the stage and E1 repositioned the tomato in the center of the stage for the start of the next trial.

Test trials also began with the tomato positioned between the two boxes. E1 looked at it, slid it into one of the two boxes (Fig. 1a), then closed the back curtain. At that point, E2 slid the tomato along the track in the stage floor, so that the participant saw the tomato move from the box in which E1 left it into the other box, seemingly of its own volition (Fig. 1b). After a brief pause, E1 returned and reached into one of the two boxes (Fig. 1c). On expected trials, E1 reached into the box in which she had left the tomato; on unexpected trials, she reached into the other box, where the tomato was located. In both trial types, E1 paused, with her arm reaching into one of the boxes, until participants looked away for 2 consecutive seconds or reached 30 s of cumulative looking. At the end of each test trial, E2 lowered the screen over the front of the stage.

2.2.3. Data analysis

All participants' looking times were recoded by a trial-blind offline coder. Correlation between the coders' looking times was 0.98, and the offline coder's data were used in analyses. We compared looking times to expected and unexpected trials in three ways. First, we conducted a repeated measures ANOVA comparing both trial times for each subject, and including test order as a between-subjects factor. Second, following Onishi and Baillargeon, we conducted an independent samples *t*-test, comparing looking times to the first test trial only for infants who saw expected vs. unexpected trials first. Finally, as all except two participants also completed the helping test described in Task 1, we performed an ANOVA testing for any differences in looking to expected and unexpected reaches between participants who passed, failed, or did not help during the helping task.

2.2.4. Results

The repeated-measures ANOVA comparing looking times to expected trials ($M = 21.0$ s) and unexpected trials ($M = 18.3$ s) across all infants found no main effect of trial type nor interaction with test order (P's > 0.1; Fig. 2). There was also no difference in looking to the first test trial for participants who saw an expected trial ($M = 20.7$ s) versus an unexpected trial ($M = 18.8$ s; $t(41) = 0.76$, $P > 0.4$). Thus we failed to find evidence that 1.5-year-old infants expected the experimenter to reach for an object either where she falsely believed it to be, or in its true location. (See SI for Bayesian statistics on the likelihood of these data under the null

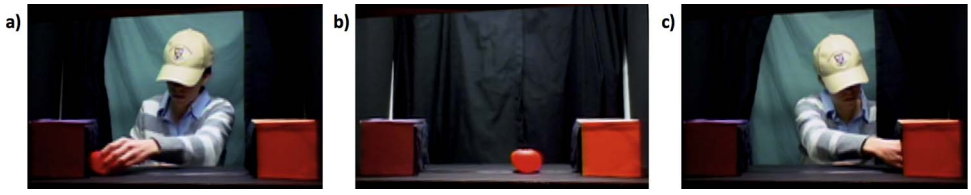

Fig. 1. Scenes from a test trial in Experiment 1, Task 2. The experimenter (a) put the tomato into one of two boxes, then (b) closed the back curtains and remained out of sight while the tomato moved independently from that box into the other box. The experimenter returned and (c) reached into one of the two boxes, either the one she left it in ('expected' trial; not pictured) or the one it had moved to ('unexpected' trial; pictured above).

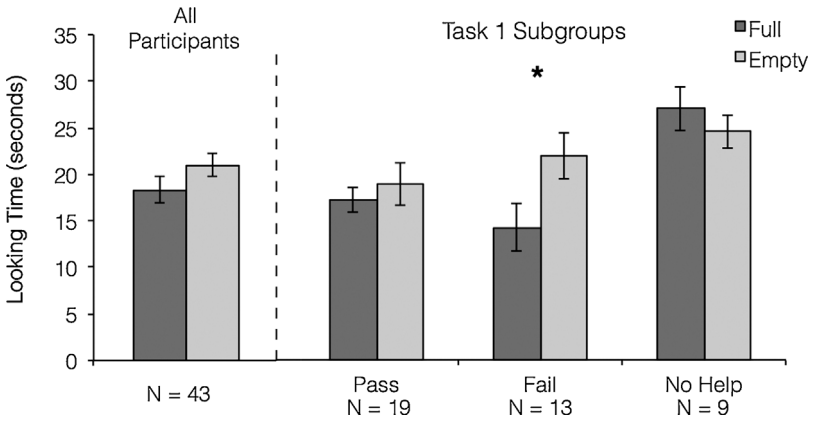

Fig. 2. Results from Task 2. There was no reliable difference in looking to trials in which the experimenter reached into the full box (the unexpected outcome) and those in which she reached into the empty box (the expected outcome). When subdivided into groups based on performance on Task 1, the only difference observed was in the group that "failed" Task 1, helping by opening the empty box the experimenter acted on, rather than the full one with her toy. Infants in this group looked significantly longer to the empty box reach (i.e. the expected outcome). Error bars represent standard error of the mean (SEM).

and experimental hypotheses.)

The ANOVA testing for a relationship between looking times in the current task and performance on the helping task did find a trend toward a trial type x helping performance interaction ($F(2,38) = 2.51$, $P < 0.1$; Fig. 2). Examining looking times within categories of helping performance revealed that infants who failed the helping task (i.e. opened the box the experimenter acted on, rather than inferring based on her ignorance or false belief that she was looking for the toy located in the alternate box) looked longer to expected trials (i.e. reaching to the currently empty box, $M = 21.8$ s) than unexpected trials ($M = 14.2$ s; $t(12) = 2.67$, $P < 0.05$), while participants who passed and those who did not help did not differentiate the expected and unexpected VOE trials (both $P > 0.3$). There was also a significant main effect of helping category on overall looking time ($F(2,38) = 4.99$; $P < 0.05$): participants who did not help in Task 1 spent more time looking at the VOE displays across trial types ($M = 51.29$ s) compared to participants who failed the task ($M = 36.0$ s; $t(20) = 2.82$, $P < 0.05$) and to those who passed ($M = 35.9$ s; $t(26) = 3.22$, $P < 0.01$).

These latter results suggest a possible relationship between participants' performance on the two implicit theory of mind tasks. However, the marginal size of the interaction effect, and the small number of participants in each category of helping performance (all N < 20) emphasize the need for additional evidence before drawing conclusions about stable ignorance or false belief attributions across tasks. One additional concern about the current data is that the maximum looking threshold, based on the method reported by Onishi and Baillargeon, resulted in nearly one quarter (24.4%) of trials being ended at 30 s of looking, before participants' own behavior indicated that they were no longer interested in the display. Task 3 addressed these concerns by recruiting a larger sample and raising the threshold for maximum looking time to 60 s.

2.3. Task 3: Violation of expectation, version 2

On the basis of the trending interaction in Task 2, we conducted a new study with the aim of testing the relationship between individual differences in implicit theory of mind performance and executive function skills. Over the course of two visits, no more than 2 weeks apart, participants completed both the looking time task described below (always the first task run in the first visit) and the helping task described in Task 1, as well as multiple tasks designed to measure working memory and inhibitory control.

Aside from increasing the maximum looking time threshold, the differences between this paradigm and that in Task 2 were based on advice received from Onishi and Baillargeon in personal communication. Specifically, they advised that using doors, rather than a curtain, to block the back of the stage would provide better evidence to the participant that the experimenter was unaware of the location transfer and that the experimenter still wanted the toy despite not reaching for it during the transfer. Second, they advised having a second, visible experimenter move the object during the transfer, rather than the object appearing to move itself, to eliminate concerns that the participant might infer that the experimenter knew the object could move on its own, disrupting inferences about her belief as to the object's location (see Song et al., 2008).

2.3.1. Participants

Seventy-nine 17–20-month-old infants participated in the second VOE task (32 female; age range: 17 months, 8 days – 20 months, 13 days; mean age: 18 months, 23 days). Fifty-seven of these infants also provided data in the helping task, 18 were excluded for not helping within the time limit, and 4 were excluded for experimenter error. An additional 10 infants were recruited to participate in the VOE task but excluded due to fussiness, inattentiveness, or experimenter error.

2.3.2. Procedure

The procedure was as in Task 2 except as follows. The boxes were purple and green, rather than purple and orange. The opening

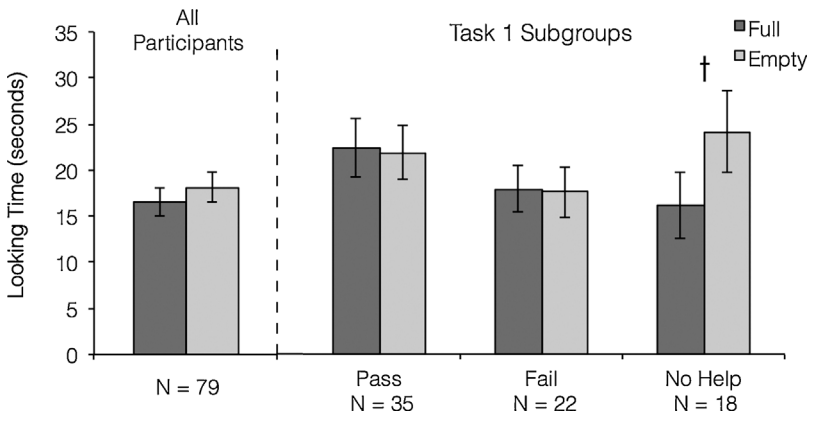

Fig. 3. Results from Task 3. There was no reliable difference in looking to trials in which the experimenter reached into the full box (the unexpected outcome) and those in which she reached into the empty box (the expected outcome). When subdivided into groups based on performance on Task 1, there were also no reliable differences within or between performance groups. Error bars represent SEM.

in the back of the stage could be covered by two hinged doors that met in the middle, rather than a curtain. Experimenter 2 (E2) sat in an opening on the side of the stage (to the participant's right), and was visible in profile from the waist up. The maximum cumulative looking time for all familiarization and test trials was increased to 60 s. And finally, the method of object transfer during test trials was changed. After E1 closed the back doors and disappeared from view, E2 retrieved the tomato from the box E1 put it in, looked at it, then placed it in the other box. E2 then resumed her typical posture before E1 opened the back-of-stage doors. E1 then returned and reached into either the empty box (expected trial) or the box with the tomato (unexpected trial). The order of the expected and unexpected trials, as well as the box in which the tomato was located, were counterbalanced across participants.

2.3.3. Data analysis

Eighty-three percent of participants' looking times were recoded by a trial-blind offline coder. Correlation between the online and offline coders' times was 0.88. Comparisons of looking to expected and unexpected trials, and of looking times to performance on Task 1, followed the same procedure as Task 2.

2.3.4. Results

The repeated measures ANOVA comparing looking times to expected trials (M = 18.1 s) and unexpected trials (M = 16.5 s) within participants again failed to find any main effects or interactions of trial type and order (all $P > 0.3$; Fig. 3). There was also no between-subjects difference in looking to the first trial when that trial presented an expected reach (M = 17.6 s) versus an unexpected reach (M = 17.4 s; $t(77) = 0.09$, $P > 0.9$; see SI for Bayes Factor calculations). In contrast to Task 2, there was also no evidence of a relationship between performance on the helping task and looking to expected and unexpected reaches ($F(2,72) = 1.46$, $P > 0.2$; Fig. 3), nor of a difference in overall looking time across helping categories ($P > 0.5$). Participants who failed the helping task spent similar amounts of time looking at expected reaches (M = 15.1 s) and unexpected reaches (M = 15.4 s; $t(21) = 0.12$, $P > 0.9$), as did participants who passed (expected M = 18.8 s; unexpected M = 19.3 s; $P > 0.8$). Participants who did not help open either box in Task 1 showed a non-significant trend toward looking longer to expected reaches (M = 20.7 s) than unexpected reaches (M = 13.8 s; $t(17) = 1.98$, $P = 0.06$), but given that the corresponding group in Task 2 showed no such difference, this likely represented random variation. Overall, we again failed to find evidence that 18-month-old infants form expectations about how an agent with a false belief will act, either in accordance with the world or with her false belief. Moreover, we failed to corroborate the tentative evidence from participants in Task 2 that infants' performance on interactive and VOE tasks cohere with one another (see also Poulin-Dubois & Yott, 2017).

2.4. Discussion

Experiment 1 achieved mixed success in reproducing previous reports on infant theory of mind. The VOE procedures in Tasks 2 and 3 failed to reproduce evidence for a rich capacity to attribute false beliefs. Nonetheless, results from the three tasks are all consistent with the hypothesis that 1-year-old infants do possess the leaner capacity to track others' knowledge or ignorance of current circumstances, and that this tracking informs their inferences from and expectations of others' behavior. Recognizing that the experimenter in Task 1 does not know where her toy truly is may permit the inference that she is looking for it, even when she pulls on the lid of the empty box. If so, an attribution of ignorance regarding the toy's location can account for participants' tendency to help by opening the box with the toy in Task 1. With respect to the VOE tasks, participants may plausibly fail to generate any expectations about where an ignorant actor will reach (Marticorena et al., 2011), resulting in equivalent looking times regardless of action, as observed in Tasks 2 and 3. One remaining limitation is that none of the tasks incorporated true belief conditions to demonstrate contrasting results when participants attributed knowledge rather than ignorance. Experiment 2 presents an additional replication attempt testing for participants' understanding of others' knowledge or ignorance, including both true and false belief

conditions.

3. Experiment 2

Knudsen and Liszkowski (2012) assessed 24-month-old children's sensitivity to others' knowledge or ignorance in an anticipatory helping task. All trials began with the experimenter putting the object into a container. On some trials, the experimenter then watched while a confederate moved the object to a different location, after which the experimenter briefly left the room; on other trials the experimenter was absent during the transfer of the object's location. When the experimenter returned and indicated she intended to retrieve the object, participants could inform the experimenter of the object's location. Participants were more likely to do so when she had been absent, rather than present, for the transfer event. This difference suggests they understood that in the absent condition the experimenter was no longer aware of the object's true location, without necessarily providing evidence of false belief ascription.

Amongst published implicit theory of mind tasks, this one seemed especially well suited for use with older children. We thus chose this task to address a question neglected by much of the literature: would young 3-year-old children, who typically fail explicit false belief tasks, succeed on the implicit measures passed by younger children, or might the factors leading to their failure on explicit tasks also affect implicit performance at this age? To ensure that we could replicate Knudsen and Liszkowski's findings, we repeated the task with one of the original age ranges (24-month-old children) as well as with a separate group of 3-year-old children. Here we report results from both age groups.

3.1. Participants

Sixteen 24–27-month-old children (8 female; age range: 24 months, 9 days – 26 months, 27 days; mean age: 25 months, 15 days) and 16 3-year-old children (11 female; age range: 35 months, 27 days – 45 months, 6 days; mean age: 40 months, 6 days) participated in the experiment. All participants' primary language was English, and no additional participants were excluded.

3.2. Procedure

While Experimenter 2 (E2) explained the study to the parent and obtained consent, Experimenter 1 (E1) spent 10 min playing with the participant in a waiting room area. E2 then hid behind a curtain in the testing room, with the participant, parent, and E1 entering after E2 was concealed. The participant and his or her parent sat on one side of a small table, opposite E1. (The arrangement of the table, curtain, and personnel matched that depicted in Knudsen & Liszkowski, 2012.) E1 and the participant briefly played on the table with a small ball. E1 then began a series of four experimental trials. At the start of each trial, E1 got out four identical containers, placed them in a row on the table, and told the participant that one container held a toy he wanted to show the participant. The toy was always in a container at either the far left or right end of the row, and E1 began opening containers at the opposite end, such that he opened each container to look for the toy before finding it in the last one. He happily showed it to the participant and they looked at it together for one min. Then E1 then placed the toy back in the box where he had found it, announced he needed to leave briefly but would come back, and left the room.

Two trials were FB trials. During these trials, after E1 left the room, E2 came out from behind the curtain, smiled and made a "quiet" gesture by placing her finger to her lips, and then moved the toy from the container E1 put it in to the one at the other end of the row. E2 then hid behind the curtain again. E1 returned and stated that he wanted to continue playing with the toy. He then slowly and deliberately walked along an L shaped path to the table and sat down. This sequence was timed to take 30 s, so that the participant had that window of time in which to inform E1, verbally or by gesturing, that the object had been moved. If the participant did inform E1 of the new location, E1 opened the container indicated by the participant and said "Hmm, that's funny! I wonder how it got there." If the participant did not inform E1 of the new location, E1 opened the container in which he had left the toy and said "Oh, it's not in here! I wonder where it is."

The two additional trials were TB trials. The only difference between these trials and the FB trials is that, while E2 was in the process of moving the toy from one container to another, E1 opened the door to the room and leaned in to watch. E2 turned to E1 to acknowledge his presence, and after E2 finished enclosing the toy in the new location, E1 said "Ok, I see" to verbally register his awareness of the new location. E1 then shut the door and remained outside the room until E2 had hidden behind the curtain again. Then E1 came back in the room and, as in the FB trials, stated his desire to play with the toy, giving the participant 30 s to gesture or verbally indicate the toy's position while he made his way back to the table. The FB and TB trials were administered in blocks, with block order counterbalanced across participants (i.e. either FB,FB,TB,TB or TB,TB,FB,FB).

3.3. Data analysis

A trial-blind coder watched videos of the sessions offline and recorded the number of times per trial that the participant gestured or spoke to E1 to indicate the position of the toy during the 30 s time period between E1′s return to the room and his first attempt to relocate the toy. A repeated-measures ANOVA compared the number of such attempts for FB and TB trials, with age group and trial order (FB or TB first) as between-subjects factors. To ensure that both age groups independently provided evidence of helping more in the context of false beliefs, we also conducted planned paired samples *t*-tests on the frequency of helping during FB and TB trials for each age group. We also conducted two non-parametric analyses. Following Knudsen and Liszkowski (2012), we used a McNemar test

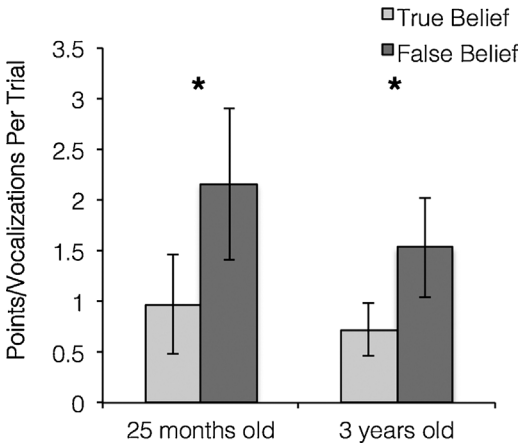

Fig. 4. Results from Experiment 2. Participants in both groups produced more points and/or vocalizations toward the box containing the experimenter's object when she had a false belief about its location compared to when she had a true belief (* $P < 0.05$). Error bars represent SEM.

to compare the proportions of participants who attempted to help at least once during FB vs. TB trials (McNemar test) for each age group, testing whether participants were more likely to inform at all on FB than TB trials. We also conducted a Wilcoxon sign rank test on the total number of attempts to inform the experimenter during false and true belief trials, to provide a non-parametric test of the hypothesis that participants would inform more frequently during FB trials.

3.4. Results

The ANOVA revealed a main effect of trial type ($F(1,28) = 11.84$, $P < 0.01$). Both age groups attempted to inform E1 of the toy's location more frequently on FB trials (2-year-olds $M = 2.16$; 3-year-olds $M = 1.53$) than on TB trials (2-year-olds $M = 0.97$; 3-year-olds $M = 0.72$). Paired samples *t*-tests were significant for both age groups (2-year-olds: $t(15) = 2.70$, $P < 0.05$; 3-year-olds: $t(15) = 2.47$, $P < 0.05$; Fig. 4). There were no other significant main effects or interactions revealed by the ANOVA. The non-parametric analyses on proportion of participants informing at least once during each trial type found a non-significant trend amongst 2-year-old participants to more frequently engage in any informing behavior during FB trials (11/16) than TB trials (5/16; McNemar, $P = 0.077$). Amongst 3-year-old children, however, there was no evidence of a difference in likelihood to provide at least one informing act between FB (9/16) and TB trials (7/16; McNemar, $P > 0.4$). The results of the Wilcoxon sign rank test were more robust. Participants informed more frequently during FB trials than TB trials ($z = 3.25$, $P < 0.005$), and this was true for both the younger age group ($z = 2.44$, $P < 0.05$) and the older age group ($z = 2.12$, $P < 0.05$).

These results replicate Knudsen and Liszkowski's (2012) findings with 24-month-old children and partially extend them to 3-year-old children, suggesting that an age group population of children well known to fail explicit false belief tasks is nonetheless motivated by an experimenter's lack of awareness to engage in more helpful, informing behaviors.

4. General discussion

We conducted exact or conceptual replications of several experimental paradigms that have been claimed to provide evidence for understanding of others' beliefs in infants and toddlers. In two paradigms, we found evidence consistent with the original reports. Both paradigms measured infants' and toddlers' interactive helping behaviors. Two- and 3-year-old children tried to communicate with an experimenter about a relocation of her object more frequently when she had not observed the transfer, replicating Knudsen and Liszkowski (2012). We also replicated Buttelmann and colleagues' (2009) finding that 18-month-old infants helped an experimenter with a false belief by unlocking the actual location of her desired object, not the empty box she misguidedly tried to open. The interpretation of the latter finding as evidence of infants' sensitivity to another's perceptual history is tempered, however, by the absence of a concurrent TB condition assessing infants behavior toward a knowledgeable experimenter under our laboratory conditions.

On the other hand, we failed to reproduce evidence that 18-month-old infants' expectations are violated when an experimenter who ought to have a false belief reaches for an object in its true location (e.g. Onishi & Baillargeon, 2005; Song et al., 2008; Träuble et al., 2010). Instead, in two independent samples, infants looked equally long when the experimenter reached to the true location of her object and when she reached to the location where she believed the object to be. Compared to the original studies, our experiments had large samples (N = 43, and N = 79 respectively), so we are confident that our effect size estimates do not reflect simple measurement error.

As in any failed replication, our failures to replicate VOE measures of infant false belief understanding raise many questions. First, we must consider whether differences in the experimental paradigm – for example, apparatus, event timing, lighting in the room, prior social interaction with the experimenter, etc. – could have obscured infants' capacities. In the current research, all experimental

details were designed, as much as possible given concurrent goals of measuring individual differences in theory of mind, to match the details provided in reports of VOE studies finding positive evidence for infant false belief understanding (Onishi & Baillargeon, 2005; Song et al., 2008; Träuble et al., 2010). Still, many details of experimental design are hard to communicate across labs and settings. In such cases, it is useful to include a manipulation check – another, theoretically neutral measure of the experiment, to ascertain that infants were, for example, sufficiently attentive to the paradigm overall. In the context of VOE studies of infants' false belief understanding, it is useful to have looking time data from the true-belief conditions, which demonstrate that a given procedure can successfully elicit the expectation that an actor will reach for an object when she has full knowledge of its location. Many labs, using varied procedures, have demonstrated such expectations, including both those that do and do not find differences in looking time under conditions of false belief (Onishi & Baillargeon, 2005; Poulin-Dubois, Polonia, & Yott, 2013; Surian et al., 2007; Song et al., 2008; Träuble et al., 2010; Yott & Poulin-Dubois, 2016). Unfortunately, we did not include a true-belief condition in the current experiment because of our concurrent goals and time constraints.

Second, we can ask whether there were differences in the populations sampled. Although we followed previous reports in recruiting infants between 15 months and 24 months of age (e.g. Onishi & Baillargeon, 2005; Song et al., 2008; Scott & Baillargeon, 2009; Träuble et al., 2010), it is possible that the population our sample drew from is different in some relevant respect from those sampled in other studies. Perhaps infants in our sample were more likely to be in full-time daycare, or more likely to be multi-lingual (although participants in interactive tasks all spoke English as their primary language) and therefore relatively delayed in language acquisition. Any difference in the children's cognitive development could be a factor moderating their sensitivity to others' false beliefs in a looking time paradigm.

Third, critics of replication attempts sometimes comment that replicators are inexperienced experimenters, and/or that the goal to "prove a finding false" leads to a negative confirmation bias. In our case, the research team was highly experienced with measures of infant social cognition. More pertinently, the goal of the research was not to undermine, but to build on, influential previous studies. The tasks in Experiment 1 were part of an ambitious study designed to measure individual differences in the implicit tasks in 1.5-year-old infants and to relate those differences to executive function abilities. Initial results (reported in Task 2) suggested an intriguing link between performance on helping and looking time measures; as a result, a large, effortful, and high-powered confirmatory test was conducted (Task 3). The hint of an association between the two tasks disappeared in this higher-powered test. This sequence reflects the scientific process: a process of hypothesis generation and confirmatory tests, designed to build on prior results.

In sum, it is possible that our results differ from previous findings because of subtle differences in the procedure or population. However, we consider it more likely that 18-month-old infants, as a population, do not reliably form clear action predictions on the basis of false beliefs. Although this interpretation contradicts a few studies with small samples of infants between 12 and 18 months of age, it is consistent with other published studies of infants throughout the second year of life (e.g. Poulin-Dubois et al., 2013; Surian et al., 2007; Sodian & Thoermer, 2008; Yott & Poulin-Dubois, 2016). These studies all find that infants form expectations about the experimenter's actions (to correctly reach for her desired object) when she has maintained perceptual contact with the event, but suspend this prediction when she has a false belief or disrupted access. Similar patterns of results have been observed in macaques (Marticorena et al., 2011; Martin & Santos, 2016). Both human infants and macaques may consistently track an agent's awareness of an object, and use that awareness to form action predictions, but not consistently track, or form predictions based on, the specific contents of false beliefs.

Is this proposal also consistent with our interactive helping task data and published results from similar studies? We think so. Helpful informing is an appropriate response to the lack of awareness, or ignorance, of someone about to look for an object. Moreover, an infant who considers a social partner to be unaware of the specifics of the situation and location of objects around her may make inferences not based on particulars of the partner's actions, but on likely goals in a given situation. This could explain why participants in Buttelmann et al. (2009) false belief paradigm were more likely to help by opening the box containing the toy, a probable goal, rather than the empty box the experimenter is acting on. It is also consistent with Priewasser and colleagues' finding (2017) that if a third, always-empty box is added to the scenario, infants in an FB condition assume an experimenter who pulls on this box is looking for the toy. If infants were tracking the experimenter's specific beliefs, they ought to have inferred she was trying to open the always-empty box, as it did not match her false belief about the toy's location. However, if infants simply considered the experimenter to be ignorant of the toy's location, they may plausibly interpret attempts to open any box as a search for the toy.

Of course, there are a number of other implicit test of theory of mind on which infants have been shown to succeed over the course of the second year of life, not all of which are compatible with positing that 1-year-olds' theory of mind is limited to tracking awareness relations. We await additional tasks and high powered replication attempts to give a clearer picture of how the 18-month-old population as a whole gathers information about others' knowledge, and how they use that information to predict others' behavior.

Acknowledgments

We thank West Resendes, Rebecca DiStefano, Brooke McDowell, Joseph Rodini, Mai Hinton, and Meg Barrow for assistance with data collection and coding. This research was funded by the National Science Foundation (DRL-0940140) and a grant from the Spencer Foundation to Susan Carey.

Appendix A. Supplementary information

Supplementary information associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.

cogdev.2017.10.004.

References

Allen, J. W. P. (2015). How to help: can more active measures help transcend the infant false-belief debate? *New Ideas in Psychology, 39*, 63–72.
Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*(1), 48–53.
Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*, 110–118.
Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.
Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* Oxford University Press.
Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), B25–B31.
Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*(2), 337–342.
Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032–1053.
Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*(4), 391–416.
Fizke, E., Butterfill, S., van de Loo, L., Reindl, E., & Rakoczy, H. (2017). Are there signature limits in early theory of mind? *Journal of Experimental Child Psychology, 162*, 209–224.
Flavell, J. H. (1999). Cognitive development: children's knowledge about the mind. *Annual Review of Psychology, 50*(1), 21–45.
Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2016). Implicit and explicit false belief development in preschool children. *Developmental Science*. http://dx.doi.org/10.1111/desc.12445.
Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science, 16*(2), 209–226.
Keen, R. (2003). Representation of objects and events. Why do infants look so smart and toddlers look so dumb? *Current Directions in Psychological Science, 12*(3), 79–83.
Knudsen, B., & Liszkowski, U. (2012). Eighteen-and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science, 15*(1), 113–122.
Kovács, M., Téglás, E., & Endress, A. D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.
Kulke, L., Reiß, M., Krist, H., Rakoczy, H. (2017). Implicit Theory of Mind: Replicability across the life span. *Cognitive Development*.
Leslie, A. M. (1994). ToMM, ToBy, and Agency Core architecture and domain specificity. *Mapping the Mind: Domain Specificity in Cognition and Culture,* 119–148.
Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition, 105*(3), 489–512.
Marticorena, D. C., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science, 14*(6), 1406–1416.
Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences, 20*(5), 375–382.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.
Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: how deep? *Science, 308*(5719), 214–216.
Perner, J., Lang, B., & Kloo, D. (2002). Theory of mind and self-control: more than a common problem of inhibition. *Child Development, 73*(3), 752–767.
Poulin-Dubois, D., & Yott, J. (2017). Probing the depth of infants' theory of mind: disunity in performance across paradigms. *Developmental Science,* e12600.
Poulin-Dubois, D., Polonia, A., & Yott, J. (2013). Is false belief skin deep? The agent's eye status influences infants' reasoning in belief-inducing situations. *Journal of Cognition and Development, 14*(1), 87–99.
Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2017) Helping as an early indicator of a theory of mind: Mentalism or teleology? *Cognitive Development*.
Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual Review of Psychology, 55*, 87–124.
Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80*(4), 1172–1196.
Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249.
Scott, R. M., Baillargeon, R., Song, H. J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology, 61*(4), 366–395.
Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.
Sodian, B., & Thoermer, C. (2008). Precursors to a theory of mind in infancy: perspectives for research on autism. *The Quarterly Journal of Experimental Psychology, 61*(1), 27–39.
Song, H. J., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition, 109*(3), 295–315.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586.
Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology, 30*(1), 172–187.
Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: do infants understand others' beliefs? *Infancy, 15*(4), 434–444.
Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523–541.
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.
Wellman, H. M. (2011). Developing a theory of mind. *The Wiley-Blackwell Handbook of Childhood Cognitive Development, 2*, 258–284.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.
Yott, J., & Poulin-Dubois, D. (2016). Are infants' theory of mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development, 17*(5), 683–698.