Cognitive Development 46 (2018) 112–124

Contents lists available at ScienceDirect

Cognitive Development

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

Invited Commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations☆

Renée Baillargeon[a,*], David Buttelmann[b], Victoria Southgate[c]

[a] University of Illinois at Urbana-Champaign, United States
[b] University of Bern, Switzerland
[c] University of Copenhagen, Denmark

ARTICLE INFO

Keywords:
Theory of mind
False-belief understanding
Replication
Implicit false-belief task

ABSTRACT

There are now over 30 published reports, spanning 11 different methods, providing convergent evidence for false-belief understanding in children ages 6–36 months (for a review, see Scott & Baillargeon, 2017). The negative findings reported in this special issue of *Cognitive Development* are inconsistent with this body of data, and the aim of this commentary is to try to shed some light on the discrepancies between studies. We examine the negative findings reported with violation-of-expectation tasks (written by R. Baillargeon), interactive tasks (written by D. Buttelmann), and anticipatory-looking tasks (written by V. Southgate). In many cases, procedural differences between studies may explain failures to replicate. In other cases, apparent participant motivation and attention differences may be important in explaining failures, raising doubts about the utility of some paradigms to elicit the behaviors on which they rely. Our hope is that this commentary will provide a useful analysis that will inform the design of future studies in order that a higher level of replication can be achieved.

1. Introduction

There are now over 30 published reports with evidence of some capacity for false-belief understanding in children ages 6–36 months (for a review, see Scott & Baillargeon, 2017). These reports have used 10 different methods, including (a) behavioral spontaneous-response tasks (violation-of-expectation, anticipatory-looking, preferential-looking, anticipatory-pointing, and affective-response tasks); (b) interactive or elicited-intervention tasks (helping and referential-communication tasks); (c) traditional or elicited-prediction tasks with reduced processing demands; and (d) neural spontaneous-response tasks (neural action-prediction and neural sustained-representation tasks). To this last group, we must now add one more method, neural belief-processing tasks: Using functional near-infrared spectroscopy, Hyde and colleagues recently found that like adults, 7-month-olds tested with transfer-of-location scenarios showed more activation in a region corresponding to the temporal-parietal junction when the agent did not witness a toy's transfer than when she either did witness it or could infer it (Hyde, Aparicio Betancourt, & Simon, 2015; Hyde, Simon, Ting, & Nikolaeva, 2018). Together, these reports contradict claims that the evidence for early false-belief understanding is minimal or lacking in convergent validity.

However, in this special issue of *Cognitive Development* and beyond, negative findings have recently been reported with some of these same false-belief tasks. In the following sections, we examine the negative findings with violation-of-expectation tasks (written

☆ The three authors have contributed equally to this article and are listed in alphabetical order.

* Corresponding author.
E-mail addresses: rbaillar@illinois.edu (R. Baillargeon), david.buttelmann@psy.unibe.ch (D. Buttelmann), victoria.southgate@psy.ku.dk (V. Southgate).

https://doi.org/10.1016/j.cogdev.2018.06.001
Received 25 May 2018; Accepted 1 June 2018
Available online 28 June 2018

by R. Baillargeon), interactive tasks (written by D. Buttelmann), and anticipatory-looking tasks (written by V. Southgate). In some cases, we identify procedural differences that might have contributed to the observed negative findings. In other cases, we concede that some tasks may provide less sensitive assessments of false-belief understanding, and we discuss possible reasons for why that might be the case. We hope that our comments will support constructive investigations of the nature of early false-belief understanding and the conditions under which it can best be observed.

2. Violation-of-expectation tasks

2.1. Negative findings with children

The violation-of-expectation (VOE) method takes advantage of young children's natural tendency to look longer at events that contradict, as opposed to confirm, their expectations. Thus, a positive result in a VOE task typically involves longer looking at an unexpected than at an expected event, whereas a negative result involves equal looking at the events. To date, over 15 VOE reports of early false-belief understanding have been published (Scott & Baillargeon, 2017). Several of these reports included internal replications, with positive results in two or more conditions. In addition, several reports contrasted false-belief and ignorance conditions, making it unlikely that children were merely tracking whether an agent was ignorant, as opposed to mistaken, about some aspect of a scene.

In light of this substantial positive evidence, how can we make sense of the negative VOE false-belief findings reported in this issue (Dörrenberg, Rakoczy, & Liskowski, 2018; Powell, Hobbs, Bardis, Carey, & Saxe, 2018) and elsewhere (Poulin-Dubois, Polonia, & Yott, 2013; Yott & Poulin-Dubois, 2016)? As is the case with most methods, seemingly small procedural differences can lead to negative results. Below are examples of four such differences.

2.1.1. No clear basis for expectation

Because VOE false-belief tasks depend on children's tendency to look longer at events that contradict their expectations about an agent's actions, it follows that negative findings will be obtained when the scene is ambiguous or confusing and does not support the formation of a clear expectation about these actions (i.e., where there is no expectation, there is nothing that can be contradicted).

In the task of Onishi and Baillargeon (2005), which was the first VOE task to demonstrate early false-belief understanding, children received three familiarization trials that supported a simple expectation. In the first trial, an agent played with a toy, hid it in box-A, and then paused with her hand inside the box; in the next two trials, she reached and paused inside box-A, as though wanting the toy she had placed there. Yott and Poulin-Dubois (2012) conducted a close replication of this task, and they obtained the same positive results. However, Poulin-Dubois et al. (2013) introduced several changes, with negative results. These changes included making the boxes transparent and adding a fourth familiarization trial in which the agent ignored the toy and put on a blindfold (she also wore the blindfold in test). One factor that might have contributed to this task's negative results is that this fourth trial created an ambiguous scene (e.g., was the agent no longer interested in the toy, and was she now playing a new game with the blindfold?), making it difficult for children to form a clear expectation about what the blindfolded agent would do next.

A similar argument could be made for the negative results of Powell et al. (2018), who attempted conceptual replications of Onishi and Baillargeon (2005) in two VOE tasks. Subjects received two familiarization trials, each with a complex sequence of events. In the first trial, the agent picked up a toy at the center of the apparatus floor, hid it in box-A, withdrew behind a curtain (task-2) or door (task-3) at the back of the apparatus for about 12 s, returned, retrieved the toy from box-A, hid it in box-B, and paused with her hand inside the box. In the second trial, the toy was back at the center of the apparatus, and the agent performed the same sequence of actions except that she first hid the toy in box-B. One factor that might have contributed to the tasks' negative results is that these events were somewhat confusing (e.g., why did the agent switch the toy to the opposite box when she returned?) and made it difficult for children to form a clear expectation about what the agent sought to accomplish.

2.1.2. No time to form an expectation

Negative findings will also be obtained in VOE false-belief tasks when children cannot form an expectation about an agent's actions because they are not given sufficient time to do so (Schulze & Buttelmann, 2017). This may be especially true when the agent's false belief is brought about by a novel event in the scene; children must have sufficient time to process this new information and work out its implications for the agent's actions.

In the task of Onishi and Baillargeon (2005), the familiarization trials were followed by a belief-induction trial. In one condition, for example, the toy moved in the agent's absence from box-A to box-B, and this event was followed by a paused scene that ended when children looked away and the trial ended. Because the toy moved for the first time in a self-propelled manner, children likely needed some time to process this information and work out its implications for the agent's actions. In their close replication of Onishi and Baillargeon (2005), Yott and Poulin-Dubois (2012) also used a belief-induction trial and obtained positive results. However, Powell et al. (2018) did not use a separate belief-induction trial; test trials were identical to the familiarization trials except that while the agent was briefly out of sight, the toy either moved from one box to the other on its own (task-2) or was moved by a bystander (task-3). In either case, the agent reached for a box immediately after she returned. Thus, one factor that might have contributed to these tasks' negative results is that children did not have sufficient time to process the novel information they had received and to form an expectation about what the agent would do before she completed her actions.

2.1.3. Interference effects

Negative findings will also be obtained in VOE tasks when extraneous factors interfere either with children's formation of an expectation or with the looking behavior that would normally index this expectation.

One such factor is *repeated testing with perceptually similar events*. If infants receive the same VOE task twice, even months apart, this can negatively affect their performance, because the task becomes a dual task tapping both their reasoning (e.g., what should happen next?) and their memory (e.g., I think I have seen this show before!). This might have contributed to the negative results of Yott and Poulin-Dubois (2016), who showed their subjects expected and unexpected events similar to those of Onishi and Baillargeon (2005) in two separate testing sessions held one to two weeks apart.

Another factor is *contamination* from prior tasks in the same testing session. For example, Yott and Poulin-Dubois (2016) used three different psychological-reasoning tasks, in counterbalanced order, in their first testing session. This meant that some portion of the children saw an experimenter behave irrationally in one task (e.g., reach for a toy inefficiently in a detour task) and then saw the same experimenter search for a toy incorrectly in a false-belief task. Given the documented effects of epistemic unreliability on children's expectations (e.g., children do not detect a violation when an agent who has previously behaved irrationally searches for a toy in the wrong location; Chow & Poulin-Dubois, 2009), it is important to avoid possible contamination effects when testing children in successive tasks.

The preceding comments make clear that several different extraneous factors can interfere with children's performance in VOE false-belief tasks. It might be argued that such factors are unlikely to have a substantial impact, because positive findings have been obtained under similar conditions in true-belief tasks (e.g., Yott & Poulin-Dubois, 2016). Such an argument would be mistaken, however: Being able to solve a simple problem (e.g., 2×4) under high processing demands does not mean that one will also be able to solve a harder problem (e.g., 17×19) under the same demands.

2.1.4. Non-standard criteria for ending each test trial

In a typical VOE task, each test trial ends when children look away for a set interval (ranging from about 0.5 to 2 consecutive seconds, depending on the task and age group). The rationale, as stated earlier, is that children will look away sooner when shown expected events; with unexpected events, they will keep inspecting and processing the events in an attempt to make sense of them.

In contrast to this standard approach, Dörrenberg et al. (2018) did not end each test trial when the child looked away for a set interval. Instead, after the toy was hidden, looking time was accumulated during (a) the 7-s phase in which the agent reached into a box and (b) the following 20-s still phase in which the agent paused with his hand inside the box (27 s in total). The problem with this approach is that if one allows children to keep looking back again and again in each trial, any difference between the trials will eventually dissipate.

With this in mind, we asked Dörrenberg and colleagues for their VOE false-belief data, which they kindly provided. Initial inspection suggested that children were much more attentive overall in the first 10 s of the still phase ($M = 7.26$, $SD = 2.68$) than in the second 10 s of the still phase ($M = 4.71$, $SD = 2.31$), $F(1, 25) = 46.23$, $p < .0001$. We therefore re-analyzed the data without these final 10 s (i.e., each trial was now 17 s, instead of 27 s). This analysis (which the authors confirmed) indicated that children looked significantly longer at the unexpected event ($M = 14.12$ s, $SD = 3.74$) than at the expected event ($M = 12.41$, $SD = 5.55$) overall, $F(1, 24) = 5.97$, $p = .022$. Moreover, 19/26 children showed this effect, $p = 0.014$ (cumulative binomial probability), including 8/10, or 80%, who saw the unexpected event first, and 11/16, or 69%, who saw the expected event first.[1] The main effect of order and the interaction of order and event were not significant, both Fs$(1, 24) < 1$, $p > .34$. These positive results thus provide a conceptual replication of Onishi and Baillargeon (2005).

2.2. Negative findings with adults

Low and Edwards (2018) presented adults with videos adapted from three VOE false-belief tasks with infants. Adults saw either an expected ($n = 36$) or an unexpected ($n = 36$) video from each task and were asked what they made of the ending. Low and Edwards reasoned that "if adults find it difficult to grasp the event sequences that make up complex VOE tasks, it would be difficult to sustain a rich mentalistic interpretation of infants' looking behavior on the same tasks" (p. 2). Adults' judgments about whether endings were expected or unexpected closely matched those of infants in the original reports for only one task, leading the authors to argue that "we should be cautious in drawing firm conclusions about mentalizing in infancy when adults' […] estimates of the expectedness of outcome events suggest that only certain VOE scenarios were interpreted in their intended fashion" (p. 1). However, there are many reasons why adults—with their greater knowledge, experience, and reasoning capacity—might not respond to VOE false-belief scenarios in the same manner as infants. Let us consider each scenario in turn.

When watching videos adapted from Onishi and Baillargeon (2005), 36/36 adults viewed it as expected when the agent searched for the toy according to his false belief about its location, and 29/36 viewed it as unexpected when he did not. In this case, adults' responses closely mirrored those of infants in the original report.

[1] In VOE tasks, it is not unusual to find both effects of event (looking longer at the unexpected event because it violates an expectation) and order (looking longer at whichever event is presented first). For children who see the unexpected event first, these two effects combine to yield longer looking at the unexpected than at the expected event; for children who see the expected event first, the two effects tend to cancel each other, resulting in approximately equal looking times at the two events (e.g., Baillargeon, 1987a, 1987b). Thus, it was less than ideal that in their false-belief condition Dörrenberg et al. (2018) tested 10 children with the unexpected event first and 16 children with the expected event first. Having an equal order distribution is critical in VOE tasks; an unequal distribution favoring either event can make it difficult to disentangle the effects of event and order.

When watching videos adapted from Song and Baillargeon (2008), 31/36 adults viewed it as expected when the agent searched for the toy skunk in the box from which a misleading skunk tail protruded; when the agent searched the plain box instead (where the skunk was actually hidden), 18 adults viewed this as unexpected, but another 16 adults viewed this as expected, mostly on the grounds that "the tail box was too obvious a trick to fall for" (p. 6).[2] This last result does not mean that these 16 adults were unable to grasp the events they were shown or failed to consider the agent's beliefs in interpreting his actions. Rather, it means that upon seeing the agent reach for the plain box, they were able to quickly generate an explanation for this unexpected outcome: The wily agent had not been deceived by the experimenter's conspicuous lure and had correctly guessed that the skunk was really hidden in the plain box. Although infants, too, can sometimes produce explanations for unexpected events (e.g., Aguiar & Baillargeon, 2002), it should not be surprising that in this case their responses resembled those of the 18 adults who viewed reaching for the plain box as unexpected, rather than those of the 16 adults who inferred that the wily agent had somehow not fallen for the experimenter's trick.

When watching the videos adapted from Scott and Baillargeon (2009), adults' responses diverged substantially from those of infants in the original report. However, it seems that adults drew different inferences than infants about the two toy penguins used in the events. For this particular task to assess false-belief understanding, subjects must assume that one penguin consists of a single unit that cannot come apart (the 1-piece penguin), whereas the other penguin consists of two units that, when stacked, look identical to the 1-piece penguin. However, in the task of Low and Edwards (2018), many adults apparently perceived *both* penguins as interchangeable containers with removable lids (e.g., "he can see one of the dolls and it's got its lid on", p. 6). For these adults, this meant that in the test trials (a) the agent could use either penguin to hide his key, and (b) it was more efficient for him to reach for the penguin that was visible under the transparent cover than for the penguin that was hidden under the opaque cover. In line with this analysis, 27/36 adults viewed it as expected when the agent reached for the transparent cover, and 16/36 adults viewed it as unexpected when the agent reached for the opaque cover. Moreover, many adults explained that the agent reached for the transparent cover because he could see the penguin under it: "It kind of makes sense that he'd go for that one, because you can see what's in there" (p. 6). Infants have been shown to respond in exactly the same way to events involving two interchangeable objects: When an agent saw a toy being placed under a transparent cover and an identical toy being placed under an opaque cover, infants expected the agent to reach for the toy that was visible to her, presumably because it required less mental effort and hence was more efficient (Scott & Baillargeon, 2013).

Why did the infants tested by Scott and Baillargeon (2009) assume that they were facing two different toys, a 1-piece and a 2-piece penguin? Several factors could have contributed to this assumption. First, the toys were not akin to empty containers with lids (e.g., the two pieces of the 2-piece penguin were more like the two halves of a grapefruit), so infants might have perceived the penguins as novel toys and not brought to bear their budding knowledge about containers and lids. Second, the toys were moved in ways designed to support these distinct construals. Thus, the 1-piece penguin was always held from the top when moved, making it clear that it moved as one unit, and the 2-piece penguin was always moved in two pieces and was not moved after it was stacked, so that infants never saw it move as one unit.

Together, the findings of Low and Edwards (2018) thus provide no evidence that their adults ever "failed to act in a belief-based manner" (p. 1). Rather, their findings make clear that adults' responses to false-belief scenarios may differ from those of infants for many reasons, including (a) adults are better able to generate explanations for unexpected outcomes (e.g., the agent was not deceived by the obvious lure), and (b) adults' greater knowledge about the world may lead them to perceive experimental displays differently (e.g., if one toy's lid can be removed, another identical toy's lid most likely can be removed too).

3. Interactive tasks

Interactive tasks, created to measure the behavioral consequences of participants' mindreading abilities, have received less attention with regard to replications compared to tasks measuring participants' gaze behavior. Still, evidence for the successful application of belief tracking in interactive contexts comes from a variety of studies using different paradigms and different behavioral measures. The first published study (Buttelmann, Carpenter, & Tomasello, 2009) was based on a change-of-location paradigm in which an agent placed her toy in box-A, and an assistant moved it to box-B. Based on the agent's presence (true-belief condition) or absence (false-belief condition) during the switch, she represented the toy either correctly at its current location B or incorrectly at its previous location A. In both conditions, the agent then unsuccessfully tried to open the empty box-A, and it was participants' turn to reason what might be the agent's goal. The rationale was that if the agent represented her toy in the box she tried to open, her goal was most likely that of retrieving her toy. In contrast, if she represented her toy in the other box but still tried to open the empty box, retrieving the toy was less likely to be her goal; instead, she most likely wanted the empty box opened for some reason. The 16-, 18-, and 30-month-old participants differed significantly between conditions in their likelihood to open box-B: While 77% of participants opened this box in the false-belief condition, only 29% of participants chose this box in the true-belief condition ($\chi^2(1, N = 124) = 29.15, p < .001$). Thus, the infants and toddlers (and, recently, great apes, Buttelmann, Buttelmann, Carpenter, Call, & Tomasello, 2017) were more likely to expect the toy to be the agent's goal in the false-belief than in the true-belief condition. In conceptual replications of this task, different laboratories demonstrated young children's predominant choice of box-B in the false-

[2] The skunk condition of Low and Edwards (2018) was actually adapted from the original *doll* condition of Song and Baillargeon (2008), as the original *skunk* condition was different: The agent preferred the skunk over the doll in the familiarization trials, and then faced a plain box and a box with a tuft of doll hair (not a box with a skunk tail) in the test trial. To succeed as they did, infants had to reason that the agent would falsely assume the doll was in the hair box and the skunk was in the plain box.

belief condition and their differentiating performance between the false- and the true-belief condition (Allen, 2015; Fizke, Butterfill, van de Loo, Reindl, & Rakoczy, 2017; Oktay-Gür, Schulz, & Rakoczy, 2018; Powell et al., 2018; Priewasser, Rafetseder, Gargitter, & Perner, 2018). Moreover, 17- to 18-month-old infants also passed other interactive tasks based on the change-of-location paradigm but including other measures or manipulations (e.g., Knudsen & Liszkowski, 2012a; Knudsen & Liszkowski, 2012b; Southgate, Chevallier, & Csibra, 2010). The results of one of these studies (Knudsen & Liszkowski, 2012a) has been replicated successfully in a study included in this special issue (Powell et al., 2018). Notably, infants' success in interactive false-belief tasks is not limited to the tracking of others' beliefs about the location of an object. They have also passed tasks in which an agent held a true or false belief about either the content of a deceptive box (Buttelmann, Over, Carpenter, & Tomasello, 2014) or the identity of an object (Buttelmann, Suhrke, & Buttelmann, 2015). Thus, infants have succeeded in interactive helping tasks based on all three paradigms invented originally as explicit verbal tasks and passed usually around 4 to 5 years of age (Gopnik & Astington, 1988; Hogrefe, Wimmer, & Perner, 1986; Wimmer & Perner, 1983).

So what led authors in this special issue either to find somewhat different results than those obtained in original interactive studies with infants (Crivello & Poulin-Dubois, 2018) or to find results that support alternative interpretations (Priewasser et al., 2018, Study 2)? As in the previous section on VOE tasks, we will discuss some of the possible factors, focusing mainly on those relevant to the present studies (for other factors such as participants' age or motivation, see Buttelmann, 2017).

3.1. Possible factors for failures to replicate interactive tasks

3.1.1. Differences in set-up, materials, and procedure

Crivello and Poulin-Dubois (2018) introduced what they called "slight methodological changes" to the Buttelmann et al. (2009) study in their conceptual replication studies, which failed to replicate the original results. It seems possible that these differences in set-up and procedure were responsible for this failure. First, in order to decrease the drop-out rate, participants were seated at a table, thereby substantially decreasing the distance between participants and the test boxes (so much that ambiguous choices were possible). As mentioned above for VOE tasks (section 2.1.2), this meant that participants had less time to process what they had observed before they responded. The inability to process might then lead to chance performance between conditions. This is what Crivello and Poulin-Dubois found. Notably, when the authors increased the distance between the participants and the test boxes to a distance about half the distance used in Buttelmann et al.'s study, in their second experiment, performance increased significantly (see Experiment 2, result section).

The set-up was not the only difference between the Crivello and Poulin-Dubois (2018) task and that of Buttelmann and colleagues. While Buttelmann et al. (2009) (and Priewasser et al., 2018) trained infants only on how to unlock the test boxes, Crivello and Poulin-Dubois trained their participants to both lock and unlock the boxes (see procedure). This extended training on the boxes could have focused infants more on the boxes and less on other relevant aspects of the study (e.g., the toy or the agent's belief states). In support of this hypothesis, Crivello and Poulin-Dubois found that infants in the false-belief condition, a condition repeatedly replicated in previous studies (e.g., Powell et al., 2018; Priewasser et al., 2018), performed at chance level.

Priewasser et al. (2018) have to be acknowledged for the efforts they made to follow the original procedure as closely as possible in their direct replication study (i.e., Study 1) of the Buttelmann et al. (2009) task. However, in Study 2, which the authors conducted in order to challenge Buttelmann et al.'s interpretation of their findings, the set-up differed in at least two important aspects from the original set-up[3]. Firstly, the set-up was considerably more complex: Instead of two boxes, the authors presented participants with three boxes. Consequently, in Study 2 there were more features to process for participants, which might have influenced the results. Support for this possibility comes from two of the authors' findings. For example, the presence of the third box did distort the results insofar that children did not differ significantly in their performance between the two replication conditions (Old-FB and Old-TB). Further, the children had a strong preference for approaching the center box in the two control conditions (New-TB and New-FB). Although the removal of these center responses seemed to make the results more comparable to those obtained in the first study (with dramatically small numbers of participants in some of the cells), the participants still chose the box with the object equally often in both control conditions (Fisher exact test, $N = 20$, $p = .123$). This might be because the presence of the center box still influenced "left and right" responses at test.

A second aspect that differed fundamentally from the original study was that in the control conditions the agent tried to open a box that never contained the toy (Jacob, 2017). More specifically, this box did not play any role in the baiting or re-baiting process. It thus received noticeably less attention compared to the other two boxes during the procedure (i.e., children's attention was never directed towards the "always empty" box after the agent had placed her toy in box-A, see supplementary material). Due to this difference in involvement of boxes, it remains unclear whether it was unambiguous to participants how the agent represented the content of this third box (see also Jacob, 2017, for other problems due to the involvement of a third box).

3.1.2. Statistical power

In Buttelmann et al. (2009) Study 2, the infant data analyses included at least 25 individuals in each of the two conditions, resulting in at least 50 individuals per age group (16- and 18-month-olds). In their replication attempts, Priewasser et al. (2018) included only 14 individuals (Study 1) and 18 or 19 individuals (Study 2) in each condition. That is, the samples were noticeably

[3] In Study 2 there was another difference to the original study, that is participants opened each box only once (and not twice) in the training phase. This might focus participants 'attention less on the opening of boxes and more on the toy involved. We do not elaborate on this issue here.

smaller than those in the original study. With regard to the results the authors found, with 11 additional participants in each condition in their Study 1, their results could look even more comparable to the original ones. In their conceptual replication study, Crivello and Poulin-Dubois (2018, Study 1) also included fewer participants than did Buttelmann and colleagues.[4] Assuming that infants' and toddlers' mindreading abilities might still be less well developed than those of preschoolers, it seems important to include similar sample sizes when attempting to replicate the effects obtained in original studies. We are aware that the authors of the replication studies included larger samples in their second studies. However, these studies face other difficulties (see sections above and below).

3.1.3. Affiliation

Buttelmann et al. (2009) administered their interactive false-belief task as a single task, which kept the familiarity with the experimenter who acted as the agent holding the belief at test relatively low. In contrast, Crivello and Poulin-Dubois (2018) administered this task as part of a battery of four (Exp. 1) or three (Exp. 2) tasks in counterbalanced order. The experimenter who acted as the agent holding the belief at test was the same experimenter who conducted the other tasks (Poulin-Dubois, personal communication, July 2014). Consequently, the familiarity between infants and the agent was significantly higher in their study than in that of Buttelmann and colleagues (at least for the majority of infants tested). Previous studies have found that familiarity with an agent can have a debilitating effect on adults 'performance in false-belief and perspective-taking tasks (Savitsky, Keysar, Epley, Carter, & Swanson, 2011; Todd, Hanko, Galinsky, & Mussweiler, 2011). In addition, studies using brain imaging reported decreased activity in theory-of-mind related areas in response to familiar faces compared to unfamiliar faces (e.g., Bartels & Zeki, 2004). Thus, it seems possible that Crivello and Poulin-Dubois' failure to replicate Buttelmann et al.'s results is due to less mindreading activity in their participants due to increased familiarity with the agent. This hypothesis is partly supported by the data (which the authors kindly provided). In Crivello and Poulin-Dubois' first experiment, correct performance decreased from 66.7% of infants in the group who received this task as their first task to 12.5% of infants in the group who received this task as their fourth task (with 50.0% and 23.1% of infants performing correctly in the two middle groups, respectively). Infants who received this task in the first half of their test session significantly outperformed those who received this task in the second half of their test session ($\chi^2(1, N = 41) = 5.71$, $p = .017$). Surprisingly (with respect to this result), no effect of order was obtained in the false-belief condition of Crivello and Poulin-Dubois' second experiment, in which - irrespective of order - infants always chose the box with object at chance level.

3.1.4. Differences in populations studied

While Buttelmann et al. (2009) tested German children, Priewasser et al. (2018) tested Austrian and Scottish children, and Crivello and Poulin-Dubois (2018) tested Canadian children. Although all these populations can be considered WEIRD populations (Henrich, Heine, & Norenzayan, 2010), it is not unlikely that differences might occur with respect to the participants' mastering of Buttelmann et al. (2009) interactive helping task (for such an effect in explicit tasks see Wellman, Cross, & Watson, 2001). For example, the populations might differ in their reactions towards the relevant cues in the test situation (e.g., the agent's attempt to open a box, the object being located in one of the boxes). In fact, a closer inspection of Priewasser et al.'s data reveals that in Study 1, Austrian and Scottish children performed similarly in the false-belief condition, with a majority of participants choosing the box with the object (86% and 100%, respectively), in line with the findings of Buttelmann et al. In contrast, while Austrian children chose the expected empty box in the true-belief condition at relatively high rates (i.e., 60% of participants), only a minority of Scottish children did so (i.e., 33%). Since the testing environment also differed for these two groups (i.e., laboratory in Austria and playgroup in Scotland), it remains unclear which factor caused this difference in performance.

3.2. An evaluation of alternative interpretations and the supporting evidence

Priewasser et al. (2018) suggested that participants' different performances in the true- and false-belief conditions of the Buttelmann et al. (2009) task might be due to differences in *social context* between the two conditions. More specifically, while in the false-belief condition the assistant sneakily transferred the toy from one location to the other, in the true-belief condition no conspiratorial context was involved (Priewasser et al., 2018; see also Allen, 2015; but see Buttelmann, 2017). Although this difference between conditions could have contributed to the findings in the Buttelmann et al. (2009) task, infants have also passed a number of published interactive tasks where no such conspiratorial context was included in any of the conditions (e.g., Buttelmann et al., 2014; Buttelmann et al., 2015; Knudsen & Liszkowski, 2012b). Counter-evidence for this argument also comes from explicit false-belief tasks: Here the expected response is that participants have to point at the empty box in order to indicate where a mistaken agent will look for her object. It is known that performance in these tasks is increased with the more conspiratorial context included (Wellman

[4] The numbers Crivello and Poulin-Dubois (2018) reported for excluded infants in Buttelmann et al.ös (2009) task are misleading. In order to have a final sample of $n = 132$ 16- and 18-month-olds, $n = 43$ infants were tested but had to be excluded from the analysis for several reasons (see Buttelmann et al., 2009, p. 340). An additional $n = 26$ 16-month-olds and $n = 18$ 18-month-olds did not help the experimenter during the test phase. Consequently, their response could not be analyzed (but they were not "excluded from the analyses" in the strict sense). Buttelmann and colleagues report separate results for the infants who did help at any point of the response phase ($n = 132$) and those who helped without parental assistance ($n = 100$). Thus, the exclusion rate - in the strict sense - was about 20% (43 out of 219 infants who entered the testing room). Further, to say that eight participants informed the experimenter about the new location of the toy *before* he acted on the empty box is incorrect. Although some of the toddlers tested by Buttelmann et al. did inform the experimenter before he acted on the box, the authors did not report any numbers about how many toddlers actually did this. Additionally, note that infants, that is the age group tested by Crivello and Poulin-Dubois, informed the experimenter in the Buttelmann et al. study only "as (or right after) E2 tried to open the box" (p. 340).

et al., 2001). Thus, if the conspiratorial context made the agent's goal (in contrast to the agent's belief) more salient, one would expect the opposite in these tasks: Participants' attention should be focused more on the location containing the object, resulting in failing the explicit task.

Another suggestion Priewasser et al. (2018) alluded to is that the conditions differed regarding the agent's *ownership* of and *interest in* the toy. We think it is important to note that in both the false-belief and true-belief conditions the agent put her toy into the box stating, "I will now put this toy away" at the end of the play phase, right before the assistant switched the location of the toy. The agent therefore interrupted her play with the toy in a comparable manner in both conditions (keeping objective facts more similar than supposed by Priewasser et al.).

4. Anticipatory-looking tasks

The special issue includes three papers (Dörrenberg et al., 2018; Grosse Wiesmann, Friederici, Disla, Steinbeis, & Singer, 2018; Kulke, Reiß, Krist, & Rakoczy, 2018a), reporting non-replications of an anticipatory-looking (AL) paradigm originally implemented by Southgate, Senju, & Csibra, 2007, and extended in subsequent studies (Senju et al., 2010; Senju, Southgate, Snape, Leonard, & Csibra, 2011; Senju, Southgate, White, & Frith, 2009). In addition, a further two papers published recently elsewhere have reported failed replications of the same paradigm, in both adults (Kulke, von Duhn, Schneider, & Rakoczy, 2018b) and toddlers (Schuwerk, Priewasser, Sodian, & Perner, 2018). Failed replications are also reported by Kulke et al. for other AL studies with adults (Schneider, Bayliss, Becker, & Dux, 2012a) and infants (Surian & Geraci, 2012), and by Burnside and colleagues of the Schuwerk, Jarvers, Vuori, & Sodian, 2016 study (Burnside, Ruel, Azar, & Poulin-Dubois, 2018). Considered together, these failed replications are troubling and reveal limitations on the utility of AL as a measure of spontaneous Theory of Mind, in both children and adults.

The aim of this part of the commentary is to try to make some sense of these non-replications. Interpretation is difficult because there is no clear pattern of results. Unlike for VOE and interactive helping paradigms, it is less clear what procedural differences might impact AL, since some of the non-replications had conditions using the original stimuli. Part of the absence of any clear pattern arising from the non-replications might be due to the wide variety of parameters used, including age of participants, stimuli, and analysis criteria, only a few of which matched the original study. In fact, Experiment 2 (with sound) by Grosse Wiesmann et al. (2018) is the only direct replication attempt of Southgate et al. (2007), having used two familiarization trials and a single test trial with the original stimuli and the original age group, and not in conjunction with multiple AL paradigms. But, even when considering only those studies that employed the original age, stimuli, and analysis criteria, some report above-chance performance in one condition and below-chance performance in another, whereas others find at-chance performance across conditions. However, the most consistent and striking finding from these non-replication reports is the apparent failure of most of the various paradigms to elicit anticipatory saccades towards the correct window in *familiarization* trials in the majority of participants, where no mental-state attribution is required. The remainder of this section will recap on the aim of the original AL false-belief paradigm design, before considering what these non-replications might mean for the paradigm and for theories of early false-belief understanding more generally.

4.1. The original study by Southgate et al. (2007)

The original study was motivated by the findings of Onishi and Baillargeon (2005), and had the aim of investigating whether preverbal infants could demonstrate not only 'after-the-event' but also 'online' epistemic-state tracking. The study of Southgate et al. (2007) had two false-belief scenarios that included elements that controlled for one another. First, in the FB1 scenario, the false-belief-congruent box was the box where the infant and the agent last saw the object placed; in the FB2 scenario, the object was placed in another box after the agent turned away and so the false-belief-congruent box was not the same box that the infant observer last saw the object placed. This was intended to control for a possible tendency to look in anticipation towards the window above the last location of the ball. Second, in the FB1 scenario, the box that was last attended by the agent was not the box where the agent last saw the object, in order to control for any potential tendency to focus on the box last fixated by the agent. In the absence of evidence suggesting that children possess either kind of bias, both were equally possible, and so FB1 and FB2 both contained important controls. As stated in the original 2007 paper, two false-belief scenarios were included instead of the more usual true-belief control because (a) together, these two conditions made opposing predictions and (b) because true-belief controls are themselves difficult to interpret and were originally included in verbal false-belief studies not because they offer any unique insight into belief reasoning, but simply because, together, true- and false-belief conditions make opposing predictions. The reasoning was that if we found a different pattern of looking between FB1and FB2, then success across these conditions could not be explained by the use of any particular non-mentalizing strategy.

While (Kulke et al., 2018a; Kulke et al., 2018b) argue that the Southgate et al. paradigm is a less stringent test of false-belief understanding than other paradigms because the object is removed from the scene rather than being transferred, it is important to point out that this is irrelevant if the question we are interested in is whether observers can generate belief-based action predictions. Change-of-location false-belief reasoning entails a prediction about where an agent will search for an object based on her (now obsolete) knowledge; it has nothing to do with the actual location of the object. Thus, whether the object is transferred to another box, or removed from the scene, the belief-reasoning requirements are the same. Our design involved removal of the object from the scene for the simple reason that our measure was post-cue saccades and, if the object remained in the scene, we thought it likely that an observer would consider the true location of the object (e.g., see Keysar, Barr, Balin, & Brauner, 2000 for an analogous result). This could be a logical part of arriving at the correct prediction and not wrong, but it would likely create patterns of eye movements that

were difficult to interpret as some would reflect an expectation of the agent's search and some could reflect the observer's consideration of the object's real location en route to making an accurate prediction. Thus, the object was removed from the scene not to make belief reasoning easier, but because doing so allowed for clearer predictions. Indeed, Wang and Leslie (2016) subsequently demonstrated that when the object is transferred rather than removed, eye movements do not reveal above-chance anticipatory looking to the belief-congruent location even in adults, despite the fact that adults can of course pass a verbal false-belief task of the same structure.

4.2. Performance on familiarization trials

Making a prediction on the basis of an agent's false belief requires that one also understand the agent's goal. In the original Sally-Anne paradigm, the test question ("Where will Sally look for her marble?") makes Sally's goal explicit; in non-verbal paradigms, we often provide familiarization trials to try to make the agent's goal clear. Thus, in the original study of Southgate et al. (2007), we included two familiarization trials which were aimed at showing infants that (a) the agent would reach for an object she saw placed in a box and (b) her reaching would follow a light and sound cue. We reasoned that infants would have no way of knowing this on the first familiarization trial, but such an expectation should be present by the second familiarization trial if infants were (a) motivated to predict the agent's action and (b) had understood that the light and sound cue predicted the agent's action. The familiarization trial was thus a true-belief trial in the sense that the agent knew the true state of affairs and the observer should predict where she would reach accordingly. Infants who did not evidence correct anticipatory looking by the second familiarization trial were excluded from further analysis because failure to accurately predict where the agent would reach on the false-belief test trial would be ambiguous. It could indicate that they did not understand that the agent had a false belief, or it could indicate that they had not understood the agent's goal or the cue-reach association or that they were not motivated to predict her action. Such criteria are also present in traditional verbal and non-verbal false-belief tasks in the form of pre-checks (e.g. Call & Tomasello, 1999) or control questions (Wellman et al., 2001).

Table 1 reveals clear differences between the rate at which participants (infants, older children, and adults) met this inclusion criterion between the different versions of the paradigm. For example, whereas 64% of 25-month-olds exhibiting an anticipatory saccade in Southgate et al. (2007) correctly predicted the location of the agent's search by the second familiarization trial, only 25% and 34% of infants did this in Schuwerk et al. (2018) and Grosse Wiesmann et al. (2018), personal communication), respectively. Similarly, Senju et al. (2009) and Low and Watts (2013) both reported that 100% of their adult participants (and for Low & Watts, also 3- and 4-year-olds) showed correct prediction by the last familiarization trial, whereas only 45% of adult participants reached this criterion in the Kulke et al. (2018b) study. These low inclusion rates meant that in order to reproduce the original analyses, these papers had to make inferences based on samples that were of a comparably small size to the original study (and in one case from 9 infants, Kulke et al., 2018a).

To what extent these differences have implications for the interpretation of test trial failures is unknown. When all (Low & Watts, 2013; Senju et al., 2009)) or most (Senju et al., 2010; Senju et al., 2011; Southgate et al., 2007) participants pass the second familiarization trial, one could conclude that as a group, they understood the task and were motivated to make a prediction. When only half the sample exhibit a correct anticipatory look on the second familiarization trial, it is legitimate to ask whether what looks like success in these 50% is actually success, or rather just random looking. The lights in the window will draw participants' attention to one of the windows so there is a 50% chance of being 'correct' without anticipating action. Thus, it is not straightforward to use equivalent test trial failure in those who passed the second familiarization trial and those who did not as justification for abandoning the original inclusion criteria (Dörrenberg et al., 2018). To compare, in their meta-analysis of the traditional false-belief task, Wellman et al. (2001) excluded from their analysis any studies in which fewer than 60% of children answered the comprehension-check control questions correctly.

There is no obvious explanation for these differences in familiarization success. For several of the studies, participants completed multiple tasks with a similar structure, and this may have contributed to a reduced interest in anticipating the agent's actions. Although several of the studies analyzed effects of task order on test trial performance, it is not clear that any did so for familiarization performance. The failure of the replication attempts to elicit evidence of a basic understanding of the task and/or motivation to predict the agent's action makes it difficult to draw any conclusions about the failures on the test trial, designed to measure belief reasoning. To be clear, the conclusion that test trial performance is difficult to interpret in the context of such a high proportion of participants failing to reach familiarization criterion is not a defense of the paradigm. On the contrary, that ostensibly equivalent versions of the task can elicit correct anticipation on familiarization trials at rates ranging from as few as 45% (Kulke et al., 2018b) to 100% (Low & Watts, 2013; Senju et al., 2009) suggests that spontaneous action prediction, the behavior which drives anticipatory looking is not reliably elicited by this task, and subject to factors that remain unknown. If true, this means not only that the task does not consistently reveal belief attributions, but it does not even consistently elicit spontaneous goal-directed action prediction.

4.3. Test trial performance

Although familiarization performance raises doubts about our ability to interpret test trial performance, there is one noteworthy pattern from the test trials of these studies. While several of the studies did reveal above-chance performance in the FB1 condition in infants (Dörrenberg et al., 2018), older children (Grosse Wiesmann et al., 2018; Kulke et al., 2018a), and adults (Kulke et al., 2018a; Kulke et al., 2018b), only one study with adults reports above-chance performance in the FB2 condition (Schuwerk et al., 2018), whereas in the other studies, performance ranged from at-chance (Grosse Wiesmann et al., 2018; Kulke et al., 2018a) to below-chance

Table 1
Participants in each of the versions of the Southgate et al. (2007) AL paradigms who met the original inclusion criterion and where this data was available. The total n's are those participants remaining after exclusions for other reasons.

Study	Age range	Stimuli	number of familiarisation trials	Participants meeting original criteria/ total	% Participants meeting original criteria
Southgate et al. (2007)	25-month-olds	original	2	20/31	64%
Senju et al. (2009)	adult controls	original	4	17/17	100%
Senju et al. (2010)	6-8-year-olds	original	4	17/19	89%
Senju et al. (2011)	18-month-olds	new	4	36/40	90%
Low and Watts (2013) (location test)	3-4 year olds, adults	new	2	52/52	100%
Non-replications					
Dorrenberg et al. (2018)	24-month-olds	original	4	30/51	58%
Grosse Wiesmann et al. (2018) (Exp.2)	25-month-olds	original	2	17/50	34%
Kulke et al. (2018a) (study 1 subset)	25-month-olds	new	4	9/23	39%
Schuwerk et al. (2018) (Exp.2)	25-month-olds	new	2	17/68	25%
Grosse Wiesmann et al. (2018) (Exp.1)	2-year-olds	original	2	27/46	59%
Kulke et al. (2018a) (study 1)	2-6-year-olds	new	4	240/460	52%
Kulke et al. (2018a) (study 2a)	2-10-year-olds	original	4	28/52	54%
Kulke et al. (2018a) (study 2b)	4-5-year-olds	original	4	40/64	63%
Schuwerk et al. (2018) (Exp. 1)	2-3-year-olds	new	2	20/47	42%
Kulke et al. (2018a) (study 2b)	adults	original	4	80/163	49%
Kulke et al. (2018b) (study 1)	adults	original	4	54/119	45%
Schuwerk et al. (2018) (Exp.2)	adults	new	2	54/83	65%

(Dörrenberg et al., 2018; Kulke et al., 2018b). In contrast, the original studies found no difference between performance on FB1 and FB2 in 25-month-olds (Southgate et al., 2007), and adults (Senju et al., 2009), and above-chance performance on FB1 and FB2 in 6–8 year olds (Senju et al., 2010). Furthermore, above-chance performance was recently reported across two FB2 conditions in apes (Krupenye, Kano, Hirata, Call, & Tomasello, 2016).

Given that the original design of the study by Southgate et al. (2007) required success on both FB1 and FB2 in order to rule out as far as possible that one or other of the conditions in isolation could be explained by the use of alternative strategies, the general failure to replicate FB2 makes it possible that the participants who did pass FB1 in these non-replication studies did so by simply looking at the window above-which they had last seen the ball (Dörrenberg et al., 2018; Kulke et al., 2018a; Kulke et al., 2018b). However, the pattern of data from the non-replications argues against this interpretation. First, if participants were doing this, why would they not also do this on familiarization trials? If they had shown this tendency, one would have expected a high level of success on the second familiarization trial when in fact success was low. Second, this strategy predicts below-chance performance on the FB2 condition, but in the majority of the non-replications, this was not the case. This pattern of results is more consistent with FB2 being more challenging in terms of memory or attention demands. In fact, Senju et al. (2010) reported that performance on FB1 was significantly better than on FB2, and speculated that the additional time that participants must maintain the representation of the agent's epistemic state during FB2 compared to FB1 is likely to make this condition more challenging. Thus, while the non-replications of FB2 certainly suggest that it is not a reliable condition for revealing epistemic-state tracking, the fact that the non-replications do not find evidence that looking behaviour is driven by any of the non-mentalistic biases that our controls were designed to reveal, suggests that successful replications of FB1 cannot be explained in this way.

Relating also to the interpretation of performance on test trials, several of the replication attempts searched for correlations between the various AL false-belief tasks administered, both in children and adults; the reasoning was that if the different AL tasks are indeed all measuring a sensitivity to others' mental states, then we should observe convergent validity between them (Dörrenberg et al., 2018; Kulke et al., 2018b). While tasks that have the same demands and tap the same cognitive processes should, in principle, show a relationship, no relationship would be likely if the tasks failed to replicate on an individual task basis, as was the case in these replication attempts. Furthermore, if, as Kulke et al. and Dörrenberg et al. speculate, success on FB1 is more likely due to behavior reading, then one could expect inter-task correlations on this basis too. Thus, it is not necessarily meaningful to interpret either the presence or absence of correlations among tasks in this instance.

4.4. Theoretical implications

Based on their non-replications, several authors (rightly) question the utility of the original AL paradigm (Kulke et al., 2018b; Dörrenberg et al., 2018; Kulke et al., 2018a; Schuwerk et al., 2018). It seems fairly clear, given the failure to elicit basic action prediction on familiarization trials, that the original paradigm does not consistently elicit or reveal spontaneous action prediction or epistemic-state tracking across multiple age ranges and populations. To what extent a revised version of the AL paradigm that increases motivation to predict might improve its success remains to be seen. But, to what extent should these non-replications lead us to question the existence of early false-belief understanding? Do these non-replications indeed "shake the empirical foundations of ambitious theories" (p. 13, Kulke et al., 2018a; Kulke et al., 2018b)?

The theory that preverbal infants have an understanding of others' beliefs is ambitious to the extent that it challenged many years of research suggesting otherwise. In assessing the impact of the AL non-replications on this theory, two things should be kept in mind. First, as was noted in the Introduction, the AL paradigm is only one of about 10 different methods that have yielded positive evidence of early false-belief understanding. Second, more recent versions of the AL paradigm, which do not suffer from some of the limitations of the original paradigm, have been developed. These, of course, also require replication, but they need to be considered in the bigger picture.

For example, it was, in part, because of dissatisfaction with the structural differences between FB1 and FB2 that Senju et al. (2011) adapted the original design into a study in which all participants (in that case 18-month-old infants) saw the same simplified FB1 condition, but infants' understanding of the agent's visual experience differed between groups. In this study, instead of turning away from the scene during the transfer, the agent put on a blindfold. Based on a clever study by Meltzoff and Brooks (2008), infants were given prior experience of a blindfold; though it appeared identical before looking through it, one group of infants discovered that it was opaque (i.e., it must prevent the agent from seeing the transfer), while the other group discovered that it was transparent (i.e., the agent wearing the blindfold must still see the transfer). This study showed that infants who experienced the opaque blindfold tended to anticipate that the agent would search in the location where she (and the infant) had last seen the object, but those infants who experienced the transparent blindfold had no particular expectation about where the agent would search and looked significantly less to the location where the agent had last seen the object.

Motivated partly by the difficulty of interpreting eye-movement data, another study (Southgate & Vernetti, 2014), with both 6-month-old infants and adults, employed a different measure of anticipation (EEG alpha suppression of motor cortex) in a within-subjects design where two false-belief conditions made opposing predictions but were matched in length and complexity. Not only did this study demonstrate that 6-month-olds made differing predictions depending on the agent's false belief (which always conflicted with the infant's own knowledge), but it also showed that adults (a) exhibited the same pattern of belief-congruent prediction and (b) this pattern of EEG-indexed prediction was supported by eye-tracking data.

The original Southgate et al. (2007) and Senju et al. (2009) studies were based on a single paradigm that was subsequently adopted by many authors because of its ease of application. However, it is only one paradigm among many, and many other paradigms not included in this special issue have provided converging evidence for early epistemic-state tracking.

4.5. Implicit Theory of Mind

Although neither Southgate et al. (2007) nor Senju et al. (Senju et al., 2009; Senju et al., 2010; Senju et al., 2011) made any claim that what was revealed on the AL task was an implicit processing of others' beliefs (instead describing the behavior as 'spontaneous'), each of the non-replication papers describe the task as an *implicit* false-belief task measuring *implicit* Theory of Mind. To the extent that the task does not explicitly ask the participant for any judgment, then it is an implicit task, but it does not follow that the process that is elicited by the implicit task is necessarily also implicit. In fact, the evidence for nonverbal paradigms revealing implicit processes is mixed. Several studies have demonstrated that adults observing a Sally-Anne scenario while engaged in a cover task, exhibit belief-processing-consistent anticipatory eye movements without apparently being aware that they have been engaged in mentalizing (Schneider et al., 2012a, b; Schneider, Slaughter, Becker, & Dux, 2014), suggesting that anticipatory looking can reflect an implicit process. On the other hand, one of these studies also suggests that anticipatory looking is not implicit in the sense of being encapsulated from extra-modular influences (Schneider, Lam, Bayliss, & Dux, 2012b), as was proposed by the two-systems view (Apperly, 2010). Ultimately, we do not know whether what we measured in the original studies, where the participants were (unlike in those in the Schneider et al. tasks) given no cover or alternative task at all, was an implicit process. We do not know whether the paradigm might elicit implicit mindreading in infants, children, and adults equally, and we do not know how the context in which the paradigm is run – as part of a battery of tasks, alongside explicit tasks, or with specific task instructions – might influence how participants construe the task. In the absence of this knowledge, the mechanism recruited during the AL tasks could just as well have been the same mechanism recruited for elicited-response tasks. For example, both adults and infants observing a version of the original Southgate et al. AL task (Hyde et al., 2015; Hyde et al., 2018; Schneider et al., 2014) appear to recruit the same core brain regions that are recruited during explicit responding to false-belief scenarios (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011), and in a much earlier task, Call and Tomasello (1999) found a correlation between 4- and 5-year-olds' performance on a nonverbal (implicit) false-belief task and a traditional verbal false-belief task.

One of the outcomes of these AL non-replications has been to question whether *implicit* Theory of Mind is a real phenomenon (Kulke et al., 2018b). However, how we construe the cognitive processes supporting mentalizing on the AL task influences how we interpret these non-replications. If one thinks of the task as measuring implicit Theory of Mind, then failure may be interpreted differently (i.e., as an absence of implicit Theory of Mind) than if it is conceptualized as just Theory of Mind. In the latter case, failure would be less likely to be interpreted as an absence of Theory of Mind (given that we know adults have Theory of Mind abilities), but rather as the result of a task that does not adequately elicit evidence for mentalizing across different populations. Given that we do not know what processes underlie AL in most of the AL paradigms, it seems premature to speculate on the reality of a phenomenon from the results of these studies.

5. Conclusions

Many of the non-replications discussed in this commentary can plausibly be attributed to four broad factors. One has to do with differences in procedure from the original tasks: Children are less likely to succeed if the experimental situation is made too complex or if they are given too little time to process the relevant information. Another factor has to do with data coding: Deviations in how responses are measured can result in negative findings. A third factor has to do with interference or contamination from other tasks in the same testing session or in prior testing sessions. If the focus of a project is to determine whether participants can demonstrate false-belief understanding in a particular task, then embedding this task in a long battery of tasks (especially when these make use of the same experimenters, stimuli, or both) may be counterproductive, because negative findings become difficult to interpret. Lastly, the fourth factor is that children are less likely to succeed if extraneous responses become equipotent in the experimental situation. In the case of interactive tasks, for example, children are less likely to succeed if the training inadvertently renders playing with locking and unlocking the boxes a salient part of the experimental situation. AL tasks may be particularly vulnerable to this difficulty. If an agent's toy is moved from box-A to box-B in her absence, participants might plausibly (a) look towards box-A in anticipation that she will search there for her toy, given her false belief, but also (b) first look towards box-B to remind themselves of the toy's current location or (c) first look towards box-B because that is where she searched last (e.g., she went to this box last time, but this time she needs to go to the other box). Because several different responses may be plausible, evidence for belief tracking may be obtained only when everything about the task conspires to make participants highly engaged by the agent's actions, so that they are entirely focused on predicting what she will do next.

Where do these non-replications lead us? We do not agree with claims in some of the special-issue papers that these negative findings cast doubt on the conclusion that some capacity for belief understanding is already present in infants and toddlers. First, the non-replications stand in contrast to a large body of positive and convergent findings: As was mentioned earlier, over 30 published reports, using 11 different methods, have now provided evidence of false-belief understanding in children under 3 years of age. Second, as this commentary makes clear, many of the non-replications are open to alternative interpretations. Third, and relatedly, several of the non-replications have yielded mutually inconsistent findings and, as such, do not support a coherent alternative account of early Theory-of-Mind abilities. Finally, some studies classified as non-replications may actually be closer to replications: Failing to find a significant difference between two conditions at an alpha level of .05 does not necessarily mean that the original finding was not replicated, given the natural range of p values observed in repeated testing (Cumming, 2008).

Nevertheless, it goes without saying that our theories must rest on strong empirical foundations. The non-replication studies have brought to light limitations—some minor, others more major—in some of the paradigms whose findings have been important in building current theories, and we must build on this new awareness to ensure that these and other methods achieve high replicability in the future.

Are large-scale, multi-lab replication projects targeting particular false-belief methods the appropriate next step for our field? Scientific progress is always a cumulative effort, and typically researchers in different labs build upon prior findings in new directions, thereby both extending these findings and providing convergent validity for them. In our view, the 30+ reports of belief understanding in young children provide a good illustration of the typical scientific process; our theories have emerged from the convergent validity of many studies. However, these are troubled times, with much attention now being focused on failed replications in the social sciences. In this context, it is clear that the non-replications published here and elsewhere have shaken the confidence of many researchers in the field (including, in some cases, our own), making large-scale, multi-lab replication projects targeting false-belief paradigms a constructive and valuable next step. We hope that the issues we have raised in this commentary will contribute in a useful way to the design of these projects.

Acknowledgments

We would like to thank Francesco Antilici, Frances Buttelmann, Cindy Fisher, Dora Kampis, Dan Hyde, Rose Scott, and Charlotte Grosse Wiesmann for helpful comments on various portions of this commentary. We also thank Sebastian Dörrenberg, Diane Poulin-Dubois, and Beate Priewasser for providing us with the data of their studies.

References

Aguiar, A., & Baillargeon, R. (2002). Developments in young infants' reasoning about occluded objects. *Cognitive Psychology, 45*, 267–336.
Allen, J. W. (2015). How to help: Can more active behavioral measures help transcend the infant false-belief debate? *New Ideas in Psychology, 39*, 63–72.
Apperly, I. (2010). *Mindreaders: The cognitive basis of "Theory of mind"*. Hove: Psychology Press.
Baillargeon, R. (1987a). Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology, 23*, 655–664.
Baillargeon, R. (1987b). Young infants' reasoning about the physical and spatial characteristics of a hidden object. *Cognitive Development, 2*, 179–200.
Bartels, A., & Zeki, S. (2004). The neural correlates of maternal and romantic love. *Neuroimage, 21*(3), 1155–1166.
Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*.
Buttelmann, D. (2017). Calling for careful designs for the evaluation of interactive behavioral measures on early false-belief reasoning. A commentary on Allen (2015). *Frontiers in Psychology, 8*, 1302.
Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLOS ONE, 12*(4), e0173793.
Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*, 337–342. http://dx.doi.org/10.1016/j.cognition.2009.05.006.
Buttelmann, D., Over, H., Carpenter, M., & Tomasello, M. (2014). Eighteen-month-olds understand false beliefs in an unexpected-contents task. *Journal of Experimental Child Psychology, 119*, 120–126.
Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *Journal of Experimental Child Psychology, 131*, 94–103.
Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development, 70*(2), 381–395.
Chow, V., & Poulin-Dubois, D. (2009). The effect of a looker's past reliability on infants' reasoning about beliefs. *Developmental Psychology, 45*, 1576–1582.
Crivello, C., & Poulin-Dubois, D. (2018). Infants' false belief understanding: A non-replication of the helping task. *Cognitive Development*.
Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300.
Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage, 55*, 705–712.
Dörrenberg, S., Rakoczy, H., & Liskowski, U. (2018). How (not) to measure infant theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*.
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.
Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*(1), 26–37.
Grosse Wiesmann, C., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2- and 3-year-olds' false belief-related action anticipation. *Cognitive Development*.
Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development, 57*(3), 567–582.
Hyde, D. C., Aparicio Betancourt, M., & Simon, C. E. (2015). Human temporo-parietal junction spontaneously tracks others' beliefs: A functional near-infrared spectroscopy study. *Human Brain Mapping, 36*, 4831–4846.
Hyde, D. C., Simon, C. E., Ting, T., & Nikolaeva, J. I. (2018). Functional organization of the temporal-parietal junction for theory of mind in preverbal infants: A near-infrared spectroscopy study. *Journal of Neuroscience, 38*(18), 4264–4274.
Jacob, P. (2017). *Can teleology explain why very young children help a mistaken agent? [Blog post]*. November 02 Retrieved fromhttp://cognitionandculture.net/blog/pierre-jacobs-blog/can-teleology-explain-why-very-young-children-help-a-mistaken-agent/.
Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32–38.
Knudsen, B., & Liszkowski, U. (2012a). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science, 15*, 113–122.
Knudsen, B., & Liszkowski, U. (2012b). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy, 17*, 672–691.
Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110–114.
Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018a). How robust are anticipatory looking measures of theory of mind? Replication attempts across the life span. *Cognitive Development*.
Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018b). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science.*.
Low, J., & Edwards, K. (2018). The curious case of adults' interpretations of ciolation-of-expectation false-belief scenarios. *Cognitive Development*.
Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science,*

24(3), 305–311.
Meltzoff, A. N., & Brooks, R. (2008). Self-experience as a mechanism for learning about others: A training study in social cognition. *Developmental Psychology, 44*(5), 1257–1265.
Oktay-Gür, N., Schulz, A., & Rakoczy, H. (2018). Children exhibit different performance patterns in explicit and implicit theory of mind tasks. *Cognition, 173*, 60–74.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.
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.
Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind with tasks varying representational demands. *Cognitive Development*.
Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or teleology? *Cognitive Development*.
Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology, 47*(1), 269–273.
Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage, 101*, 268–275.
Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012a). Eye movements reveal sustained implicit processing of others' mental States. *Journal of Experimental Psychology: General, 141*, 433–438.
Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012b). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science, 23*(8), 842–847.
Schulze, C., & Buttelmann, D. (2017). Toddlers' updating of belief attribution through indirect communicative acts. March *Paper presented at the International Convention of Psychological Science*.
Schuwerk, T., Jarvers, I., Vuori, M., & Sodian, B. (2016). Implicit mentalizing persists beyond early childhood and is profoundly impaired in children with autism spectrum condition. *Frontiers in Psychology, 7*, 1696.
Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*.
Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80*, 1172–1196.
Scott, R. M., & Baillargeon, R. (2013). Do infants really expect others to act efficiently? A critical test of the rationality principle. *Psychological Science, 24*, 466–474.
Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249.
Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in asperger syndrome. *Science, 325*, 883–885.
Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., et al. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology, 22*, 353–360.
Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental States to others? A critical test. *Psychological Science, 22*(7), 878–880.
Song, H., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology, 44*, 1789–1795.
Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition, 130*, 1–10.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by two-year-olds. *Psychological Science, 18*, 587–592.
Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science, 13*(6), 907–912.
Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30*, 30–44.
Todd, A. R., Hanko, K., Galinsky, A. D., & Mussweiler, T. (2011). When focusing on differences leads to similar perspectives. *Psychological Science, 22*(1), 134–141.
Wang, L., & Leslie, A. M. (2016). Is implicit theory of mind the 'Real Deal'? The own-belief/True-belief default in adults and young preschoolers. *Mind & Language, 31*(2), 147–176.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.
Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology, 30*, 156–171.
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.