Do infants understand false beliefs? We don't know yet

– A commentary on Baillargeon, Buttelmann and Southgate's commentary---

Diane Poulin-Dubois*

Hannes Rakoczy*

Kimberly Burnside

Cristina Crivello

Sebastian Dörrenberg

Katheryn Edwards

Horst Krist

Louisa Kulke

Ulf Liszkowski

Jason Low

Josef Perner

Lindsey Powell

Beate Priewasser

Eva Rafetseder

Ted Ruffman

*** Shared first authorship**

Acknowledgements

We would like to thank Pamela Barone, Mike Frank, Charlotte Grosse Wiesmann, Jonas Hermes, Marina Proft, Paula Rubio-Fernandez, Rebecca Saxe, Dana Schneider, Britta Schünemann, Tobias Schuwerk and Lisa Wenzel for helpful comments.

Abstract

The commentary by Baillargeon, Buttelmann and Southgate raises a number of crucial issues concerning the replicability and validity of measures of false belief in infancy. Although we agree with some of their arguments, we believe that they underestimate the replication crisis in this area. In our response to their commentary, we first analyze the current empirical situation. The upshot is that, given the available evidence, it remains very much an open question whether infants possess a rich theory of mind. We then draw out more general conclusions for future collaborative studies that have the potential to address this open question.

1. Introductory remarks

Infant Theory of Mind (ToM) has been one of the hottest topic in developmental cognitive science for over a decade. Among the main reasons for this prominence has been a set of striking new findings based on ingenious non-explicit measures. They suggested a more precocious ToM (a meta-representational grasp of "belief" and other propositional attitude concepts) than previously assumed, questioned traditional dogma, and presented what may be the most impressive evidence for a rich, perhaps innate conceptual sophistication postulated in the nativist tradition. Currently, infant ToM research may be among the most hotly debated areas in developmental cognitive science because of a dawning replication crisis concerning this very evidence.

We –many of the authors of the (non-)replication studies published in the special issue—would like to thank the editors, Mark Sabbagh and Markus Paulus, for moving the field forward by bringing together, in a special issue of *Cognitive Development,* hitherto unpublished replication studies. We would like to thank Renée Baillargeon, David Buttelmann and Victoria Southgate (henceforth BBS) for their insightful commentary that pinpoints many critical issues and raises important questions for future research.

Here, we would like to take advantage of the opportunity to comment on their commentary in order to add some clarifications (and correct some minor mistakes) with regard to some of their arguments, as well as to complement their conclusions. We aim to offer a broader perspective on what we consider not just a set of anomalous findings but a serious replication crisis in infant ToM research, and draw out implications for future joint endeavors at collectively finding out the nature of infants' precocious socio-cognitive skills.

To foreshadow the general thrust of our comment: the current empirical situation in our field, exemplified by the papers in the special issue, is so serious that talking of a "replication crisis" is by no means exaggerated. In light of similar trends in other areas of psychology (e.g., Open Science Collaboration, 2015) a much more skeptical stance is warranted than acknowledged by BBS. Although there is now a large number of confirmatory published findings from a wide range of research paradigms, the bulk of the data has been collected from only a handful of labs. In addition, and more importantly, for all measures with positive findings in the original studies there is now a growing body of independent non-replications. Furthermore, from the replication crisis in other areas of psychology, and from the growing Open Science awareness, we have learned about the impact of publication bias, and file-drawer problems that should caution us against uncritically taking original findings at face value and remind us of the possibility of false positives (e.g., Simmons et al., 2011).

To avoid misunderstanding right away, we are *not* claiming that the original findings *are* false positives (and that recent non-replications are true negatives). What we are suggesting is that the original findings *might* be false positives and that there is no sound justification to give them priority over later replication failures. Although prior in temporal order they are by no means more valid than the present results and one cannot conclude that the latter must therefore be false negatives (a line of reasoning the commentary seems to imply at places). Rather, all current evidence considered, it remains an open question whether infants do operate with a (perhaps implicit) full-blown theory of mind. The only way to find out is for all of us researchers in the field to move from post-hoc speculations about existing incompatible findings to systematic investigation motivated by and designed on a priori grounds, by joining our efforts to conduct a large-scale and systematic collaborative inquiry.

The structure of our comment is as follows: in three sections (2.1-2.3.) corresponding to the three main types of measures – violation of expectation (VoE), interaction, anticipatory looking (AL)- we respond to specific comments made by BBS in their respective sections and add clarifications and corrections (detailed responses to more specific or technical issues regarding individual replication studies are to be found in separate Appendices). Then, in the final part (section 3), we draw out more general issues that we believe apply to the field as a whole, identifying central questions and directions for future collaborative studies.

2. Clarifications and responses regarding the different measures

2.1 Violation-of-expectation tasks

In 2005, a landmark study that changed the view on theory of mind development suggested that infants as young as 15 months understand false belief (Onishi & Baillargeon, 2005). This discovery was made by minimizing task demands, tapping into infants' looking patterns with the well-established violation of expectation (VoE) procedure. As pointed out by BBS, almost two decades later, there are over 30 published papers providing evidence for false belief understanding in children ages 6-36 months (Scott & Baillargeon, 2017). However, it is particularly unfortunate that in the case of the violation of expectation paradigm, out of 15 publications on false belief, 11 (73%) have been conducted by Baillargeon and her former students. This is obviously a less than desirable situation for the attainment of reliable cumulative knowledge. An additional problem is that these studies differed massively in all kinds of parameters, such as end-of-trial criteria, that ideally should be determined consistently and transparently in a priori ways – complicating the interpretability of the findings (Rubio-Fernandez, in press). Furthermore, although there is clearly a large set of studies that suggest that infants possess a mature concept of false belief that is revealed when task demands are reduced, it is also important to point out that there is evidence for a publication bias on this topic. Not all studies have found that infants have false belief understanding but, as one would expect, many of these studies have not been published. A subset of these null results is now published in a CD special issue and many others are mentioned in a recently published survey (Kulke & Rakozy, 2017). How should such non-replications be interpreted? In their commentary, BBS argue that the absence of VoE replications reported in the special issue are mainly due to methodological flaws.

Familiarization procedure

One of the methodological issues raised by BBS is that the authors of most failed replications changed elements of the original familiarization phase in ways that prevented infants from fully forming an expectation about the intentions of the agent (e.g., switching the box where she hides the toy across familiarization trials, or adding a blindfold on last trial). First, however, we note that there is no a priori explanation for why the original familiarization procedure (and only the original one) led to the formation of an expectation about the agent's intention, and why other procedures should not. Second, what BBS failed to discuss is that some of the successful replications among the 15 positive VoE findings they refer to have also introduced changes in the familiarization phase relative to the original procedure used by Onishi and Baillargeon. These include changes similar to those that BBS argue may account for null findings (e.g. familiarization events in Träuble et al., 2010 involved the object being placed in and transferred between each of the two hiding locations in an alternating pattern). If indeed changes in the familiarization procedure jeopardize task validity, does this mean that those findings may be false positives? It is also important to note that some of the changes that yielded negative findings were motivated by the need to provide more stringent tests of false belief (e.g., ensuring that reaches depicted during familiarization did not match only the expected and not

the unexpected test trial, as in Onishi and Baillargeon FB-Green, and many subsequent replications). Generally, in the absence of a sound basis for predicting a priori a functioning familiarization procedure, and without independent assessment of its working, there is no strong reason to believe that differences in outcome are a product of familiarization differences, rather than typical statistical variability around the true effect size.

Interference and order effects?

Another methodological change that BBS think might explain recent non replications is potential interference and order effects. It is true that studies that have tested infants on a battery of tasks have tended to generate null results. However, the tasks were, as one would expect, counterbalanced, and, more importantly, no order effects were reported. It is rather ironic that in the study frequently cited by BBS as a strict successful replication of the original experiment (Yott & Poulin-Dubois, 2012), the false belief task was always administered after a rule training task where the infant were trained to expect that the object was never at the last place they saw it. As for carry-over effects through multiple test sessions, the only study that examined this effect showed that infants do not attribute true beliefs to an agent who was previously showing misleading emotional cues (Chow & Poulin-Dubois, 2009). This is a potentially fruitful research line to investigate in future studies.

Duration of the test phase & re-analyses

Concerning the null results by Dörrenberg et al. (2018) BBS argue that test trial duration may be the critical factor to explain the failed replication. In Dörrenberg et al. (2018), the duration of the test trial was fixed, not infant-controlled as in the original experiment in which the test trial ended when children looked away for two consecutive seconds. We agree with BBS that when test trials are too long, any condition differences will dissipate. However, the length of Dörrenberg et al.'s outcome phase was not a randomly chosen, overly long time interval. Instead, it was based on a literature review showing that infants typically attend in a still phase between 15 to 25 seconds before they look away for two consecutive seconds (see e.g., Onishi & Baillargeon 2005). Arguably, the cumulative looking time as provided by an eye-tracker might even be more accurate as a cognitive measure of attention than the latency to look away for two consecutive seconds as determined by live hand coding.

BBS report a re-analysis of the Dörrenberg et al. data in which they chose (arbitrarily) the first 10 seconds of the still-phase test time. This re-analysis reveals looking time patterns compatible with belief-tracking in the FB condition. However, apart from the problematic approach of the post-hoc, data-driven decision for the time window (based on "initial inspection"), the data Dörrenberg et al. provided showed that at least more than half of the children had not yet looked away for more than 2 seconds after the first 10 seconds of the still-phase, thus rendering it an invalid analysis by Baillargeon's own standard. In addition, BBS's re-analysis only presents half of the picture because it does not report the corresponding re-analysis for the TB condition. Looking time data in the FB condition can only be conclusively interpreted in conjunction and contrast with TB conditions. Interestingly, the same re-analysis as performed by BBS on the TB condition reveals that children's looking time at the unexpected event was not different from looking time at the expected event, with only 14/26 (54%) children looking longer at the unexpected event (for details, see the Appendix A). Thus, FB and TB conditions taken together do not constitute conclusive evidence for belief-tracking in the more limited 10 sec time window either. Curiously, BBS did not re-analyze the actually much more important replication analysis of performance on the first trial, between-subjects. Dörrenberg et al. used the

suggested 10 seconds time window and found no differences between congruent and incongruent belief-based action processing for FB and TB conditions (for details, see the Appendix A).

Validation of task analysis with adults

Finally, in response to Low and Edwards' (2018) contribution, BBS question the general strategy of validating controversial infant task analyses with adults. They argue that adults' responses to a given task may be quite irrelevant for determining what such a task taps in infants; tasks that are suitable for infants may simply not be suitable for adults. We respectfully disagree about this general line of argumentation in two respects (for details, see Appendix B). First, we disagree about the necessity of independent validation of task analyses. If we want to know what a given task measures we do need validation for any tasks analyses proposed. In many cases, such as standard FB tasks, this is trivial since the task analysis in terms of theory of mind reasoning is rather obvious. However, in cases where the task analysis offered by the authors is disputable (anecdotally, many of us and our students did not find Scott and Baillargeon's (2009) task analysis obvious or compelling at all), it is unclear what a task taps in the absence of an independent validation of the task analysis. The most obvious independent validation of Scott and Baillargeon's study is to ask adults or older children to describe what is happening in each scenario —as Low and Edwards did. Other possibilities would be to test convergent validation: Does performance in this task converge and correlate with performance in other FB tasks with less controversial task analyses? Second, we disagree about the lack of usefulness of testing adults with paradigms from infant studies more generally. Cognitive-developmental scientists have found it useful to gather converging data across different age groups and populations because similarities and/or differences in response profiles that persist despite differences in experiences can shed light how tasks are being interpreted and, further, help characterize diverse mental models of the psychological and physical worlds that influence human beings' attention and action (e.g., Dixson, Komugabe-Dixson, Dixson, & Low, 2017; Hinten, Labuschagne, Boden, & Scarf, 2018; Kovács, Téglás, & Endress, 2010; Xu, Carey, & Welch, 1999). Low and Edwards' findings are therefore relevant: even amongst adults who *did* interpret the belief-congruent of Scott and Baillargeon's (2009) study outcome as being expected, the explanations revolved around the agent making a decision based on the types of object present (penguin or grapefruit) rather than on object identity in the strict numerical sense. The findings give researchers a starting point for delineating infants' processing of the psychological world that could support tracking false-beliefs in a limited but useful range of situations.

Conclusion

Taken together, the current set of findings reported on VoE experiments on infant's FB understanding is complex and confusing. So far, independent replication attempts have largely yielded null findings, and exceptions are difficult to interpret: First, Yott and Poulin-Dubois (2012) did reproduce the original effects, but only after a rule training phase. Second, and quite ironically, a recent replication study with an analogous scenario as Onishi & Baillargeon's, approved by Renée Baillargeon (personal communication, October 2017) generated a "replication" in the sense that children produced the same pattern of looking times (Burnside, Severdija & Poulin-Dubois, 2018). Crucially, however, the scenarios differed from the original ones in that the human protagonist was replaced by a toy crane with minimal animacy properties. Since intuitively toy cranes are not holders of beliefs or other mental states (and indeed, adult participants in the study did no ascribe any such states to the crane), this suggests that the looking pattern in this and previous studies may not reflect belief ascription at all but rather some

simpler sub-mentalizing processes (alternatively, it may be that infants do operate with a concept of "belief" but initially ascribe this concept much more widely, in a sort of promiscuous" over-mentalizing").

Taken together, it thus remains unclear whether there is compelling evidence from VoE tasks for early belief ascription. More conceptual replications are needed with infants that clearly highlight the rationale for the methodological choices that are made by the researchers, including addressing potential confounds of the design, like Powell et al. (2018). Importantly, in such studies design decisions need to be made (and pre-registered) a priori in clear, theoretically motivated, and consistent ways. This will help to address yet another problem in interpreting published positive findings due to the fact (mentioned above) that the studies in question differed in various parameters such as end-of-trial criteria that ideally should be determined in a priori ways and kept constant across studies (Rubio-Fernandez, in press). Concerning validity, both independent validation tests of the material and task analyses with older children and adults and convergent validation studies using within-subjects designs are required. The few studies published that tested FB understanding across paradigms (e.g., VoE vs Interactive) with a within-subject design have yielded null results (Dörrenberg et al, 2018; Poulin-Dubois & Yott, 2017; Kulke et al, 2018; Powell et al, 2018).

2.2. Interaction tasks

Quite surprisingly, in their discussion of interactive measures BBS focus exclusively on one particular task by Buttelmann et al. (2009). This task has been the subject of a number of replication attempts, some of them unpublished (see Kulke & Rakoczy, 2017), some previously published (Fizke et al., 2017; Oktay-Gür et al., 2018; Poulin-Dubois & Yott, 2017), some published in the special issue (Crivello et al., 2018; Priewasser et al., 2018; Powell et al., 2018).

Replicability of the Buttelmann et al. (2009) task

Independent replication attempts so far have yielded partial replications or null results. Partial replication patterns in several studies reproduced a condition difference (children responded differently in a TB compared to a FB condition), but did not replicate the original above-chance performance within each condition (Fizke et al., 2017; Oktay-Gür et al., 2018; Priewasser et al., 2018). This is an interesting pattern since it is obviously ambiguous in terms of interpretation: it is compatible with (but provides no conclusive evidence for) belief-tracking; but it is equally compatible with much more parsimonious interpretations in terms of knowledge-ignorance distinction (rather than full-blown belief reasoning).

Null results were reported by Crivello and Poulin-Dubois (2018) in the special issue. To explain this lack of replication of the original experiment BBS argue that Crivello and Poulin-Dubois's results are due to critical methodological changes. More specifically, the much shorter distance between the child and the boxes might have led infants to produce impulsive responses. It is true that the success rate improved when the distance was slightly increased across the two experiments (from 37% to 58%) but children still did not perform at above chance levels. Furthermore, only a small proportion of children touched both boxes simultaneously (a clear sign of impulsiveness) during the experimental trials. Regarding the attrition rate, having the child to walk toward the boxes yielded an exceptionally large attrition rate in the original experiment (54% when all children excluded are counted, 40% if those who helped with parental prompting are included). This very high attrition rate motivated Crivello and

Poulin-Dubois to administer the task at a table. On the other hand, having the child walk toward the boxes as in the original paradigm (Buttelmann et al., 2009) also brought some risks, such as a change of decision *en route*. This did in fact occur, but the frequency was not reported in the original paper (Buttelmann, personal communication, July 2014). Another issue raised by the authors of the commentary concerns lower statistical power whereas in fact, Crivello and Poulin-Dubois' first experiment included a final sample size of 41 in the false-belief condition in contrast to 25 18-month-olds in the false belief condition in the original study by Buttelmann et al. Yet another potential factor that may have masked mindreading capacities that BBS mention is affiliation with the agent (who administered other tasks). This interpretation, although intriguing, clashes with the well-known evidence that infants are more likely to help a familiar adult (Spinrad & Stifner, 2006), and is inconsistent with the absence of an order effect (more exposure to the adult should hinder performance). Finally, cross-cultural differences are very unlikely to account for the null results given that many helping tasks that require some form of mindreading (e.g., intention) originally reported on German populations, have been replicated across many countries.

Validity of the Buttelmann et al. (2009) task

In another contribution to the special issue, Priewasser et al. (2018) are concerned about the validity of the Buttelmann et al. (2009) task. Their main points are the following (for details, see Appendix C): Methodologically, the task suffers from confounds between TB and FB conditions. Most dramatically: The conditions do not only differ in epistemic respects (such that the agent knows about the object's location in the TB condition, but holds a false belief in the FB condition), but also in motivational respects. For example, the agent expressed interest for the object throughout the FB condition, but witnessed in a detached manner the object's transfer in the TB condition and may thus be taken to be less interested in the object on his return to the boxes. In a new 3-box version of the original study, Priewasser et al (2018) investigated whether some such factor other than the agent's false belief may be responsible for the difference in helping behavior between conditions. In a new-FB condition the relevance of the agent's false belief for explaining his behavior was neutralized (the agent attempting to open a neutral box). Yet the same difference in helping behavior emerged as in the original study. Since the agent's false belief was irrelevant for understanding his intentions the difference in behavior must be based on some other factor. There is no obvious reason why the same factor would not also have been operative in the original study and thus the results strongly suggest that this factor—and not the agent's false belief—is responsible for the difference in helping behavior in all versions. Important to note, this evidence against mentalism cannot be due to a false negative produced by the greater complexity of the 3- than the 2-box version, as intimated in the commentary. For, the 3-box version did show a very strong and reliable effect between conditions.

At this point the results are strong evidence that in the helping paradigm by Buttelmann et al (2009) no understanding of belief need be involved. To contest this conclusion one needs to show that the switch from the 2-box to the 3-box version made children approach the task completely differently (see Appendix C).

Conclusion

Interestingly, there is another interactive FB task that does not involve confounds between FB and TB conditions and is thus much more stringently interpretable than Buttelmann et al.'s task. This task

by Southgate et al. (2010) which, strangely enough, BBS ignored in their comment, has recently also been subject to replication attempts (one of them published in the special issue). These replication studies yielded consistently negative results and no evidence for convergent validation (Grosse Wiesmann et al., 2017; Supplement; Dörrenberg et al., 2018). So, taking the findings from these different types of tasks together, there is currently no robust evidence from interaction tasks for early belief ascription.

2.3. Anticipatory looking tasks

When it comes to the interpretation of non-replications, AL measures are special compared to VoE and interaction measures. Since they can be implemented with video-stimuli and eye-tracking systems in completely automated ways, strict replications can be realized by using the original stimuli and procedures. Such direct replications are particularly conclusive. They leave little (if any) room for attributing non-replication findings to procedural deviations from original methods. Many such direct replications have recently been conducted, four of them reported in the special issue (Burnside et al., 2018; Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Kulke et al., 2018a). Most of these direct replications, as well as less direct yet still very stringent conceptual replications converge in failing to replicate the original findings.

Familiarization phase and inclusion rates

The overview of the replicability of familiarization trials in Table 1 of BBS's comment is very helpful. Unfortunately, however, it contains some factual mistakes: Most relevant for current purposes, 65% (33/51) rather than only 58% subjects (30/51) anticipated correctly in the second familiarization trial of Dörrenberg et al. (2018) – a rate much like in the original study[1]. Relatedly, BBS state that based on high exclusion rates due to failed familiarization trials, these replication studies are based on smaller sample sizes. It should be noted, however, that this is not the case, as the sample sizes, even after exclusion of subjects failing the familiarization, were considerably larger or identical to those tested by Southgate et al. (2007) in 9 out of the 12 replication studies summarized in Table 1 in BBS's commentary.

More generally, it should be emphasized that the directness of replications of AL measures extends to the familiarization trials and inclusion criteria. Again, the very same stimuli and procedures as in original studies can be used – and have been used in the studies reported here. That is, whether familiarization performance matches that of original studies is itself a question of replicability. This is very different from designs in which different familiarizations are implemented or different inclusion criteria adopted. In the latter case, the informativeness of the results on the critical test trials may be a matter of justified dispute, for example, when inclusion criteria were different and questionable. In the present case, however, problematic inclusion rates do not put into question the appropriateness of the replication study, but the replicability and usefulness of original procedures. Conceptually, if the familiarization performance did indeed inform about intention-tracking behavior, participants who pass the familiarization criterion should also show improved performance in the test trial, at least in the TB condition that does not involve false belief processing but simple ascription of an intention to the agent. Empirically, however, several studies have reported statistical analyses showing that whether or not participants pass the familiarization did not significantly affect the results in the test trial (e.g. Grosse Wiesmann et al., 2018; Kulke et al., 2018a). Therefore, the most likely explanation for

[1] The other mistakes are: In Kulke et al. (2018a, Study 2b) 41/64 rather than 40/64 children passed the inclusion trials, in Kulke et al. (2018b) the correct number is 89/163 rather than 80/163 (the latter refer to those who passed ALL familiarization criteria of all four tasks).

unsuccessful familiarization trials is that the task is not very appropriate and suitable in general to induce action-anticipation and intention-tracking looking behavior.

Put constructively and prospectively, one of the major challenges for future replication studies with AL measures will be to devise better familiarization methods that ensure that participants fully engage in spontaneous action anticipation that produce lower exclusion rates.

The emerging picture

Across many replication attempts, most of them with larger sample sizes than, and with comparable inclusion rates to the original studies, there is thus currently no conclusive evidence for robust replicability for infant ToM as measured with AL tasks (nor for spontaneous forms of ToM as measured by AL in adults; Kulke et al., 2018b): None of the replication studies of Southgate et al. (2007), for example, constitutes a full-replication[2]. Some fail to replicate any effects (Kulke et al., 2018a, Study 1, subsample of 2-year-olds; Grosse Wiesmann et al., 2018; Schuwerk et al., 2018[3]). Many replicate only the effects observed in the ambiguous FB1 condition (that by itself is difficult to interpret since it is as compatible with belief-tracking as it is with simpler processes such as object-tracking; Kulke et al., 2018a, Study 1, Study 2a); and of those some find chance performance in the less ambiguous FB2 condition (Kulke et al., 2018a, Study 1, Study 2a), whereas others find below-chance performance (Kulke et al., 2018a, Study 2b). Yet other studies yield mixed results in FB1 but find the opposite effect from the original study (below-chance looking) in FB2 (Dörrenberg et al., 2018).

As BBS note in their commentary, it may be particularly challenging to interpret these non-replications given that they do not exemplify one consistent and overarching pattern. However, such complex and diverging patterns of findings are not only unproblematic; they are exactly what one would expect in cases of original false positive findings where the task may not tap what it was supposed to tap, where original findings reflect statistical fluctuations (rather than any systematic distortions) and happened to make it into print against the background of given file-drawer problems and publication biases. Again, we wish to emphasize that we are *not* suggesting that this is necessarily the case, but that the empirical pattern of (non-) replication results is highly compatible with such a possibility.

All things considered, there is thus currently no robust and unambiguous evidence from AL tasks for an early rich ToM (belief ascription). In particular, there seems to be no AL task that is appropriate and sensitive regarding familiarization and inclusion, and that delivers replicable and valid evidence for infant belief attribution. The tasks reported in the special issue suffer from lack of appropriateness/sensitivity and do not robustly replicate. The one task that proved robustly replicable in a recent large-scale direct replication study with adults (the location condition of Low and Watts, 2013) turned out to lack construct validity (effects disappeared once crucial confounds were controlled for; Kulke et al., 2018b, Study 3).

Absence of evidence is, of course, no evidence of absence. The only way to move research forward in this area is to join forces and implement collaborative, large-scale, multi-lab conceptual replication studies of AL measures of infant belief ascription. Such projects will need to develop tasks that are appropriate and sensitive regarding familiarization and inclusion, and have compelling face-validity as

[2] With the possible exception of Wang & Leslie, 2016. This study replicated only one of the original conditions (FB1), however. And see also Schuwerk et al. (2018) for critical concerns about the analyses that do deviate considerably from the original ones.

[3] It should be noted, though, that only the FB2 condition of Southgate et al. (2007) was tested here.

conclusive indicators of ToM. With such tasks, the conceptual replicability of AL-effects can be probed on grounds that are theoretically and methodologically motivated in a priori ways. Such projects may need to address specific questions regarding the interpretation of ambiguous patterns of partial replications. For example, in the AL tasks modeled on Southgate et al. (2007) in which the target object is removed from the scene, how do we interpret certain patterns of results, for example, a replication of effects in the ambiguous FB1, but not in the less ambiguous FB2 condition (see Kulke et al., 2018a, Study 2a; or Grosse Wiesmann et al., 2018 in older children)? On the one hand, such a pattern could be evidence for belief-tracking that is masked by task complexity in FB2. On the other hand, it may reflect low-level cognitive processes such as object-tracking. Novel control conditions are required to tease these alternative interpretations apart. Similarly, in AL tasks that are more directly modeled on change-of-location scenarios (Wimmer & Perner, 1983) like those of Surian and Geraci (2012) or Schneider et al. (2011), how do we interpret patterns of results (found, for example, in adults in some studies of Kulke et al., 2018b) such that participants do anticipate differently in FB and TB conditions, but anticipate correctly only in TB while looking randomly in FB? Such a pattern could be evidence for belief-tracking (masked by inhibitory demands in FB), for merely tracking knowledge-ignorance, or for even simpler processes. Again, new control conditions are required to reach firm conclusions.

3. Conclusion & Outlook

In conclusion, we would like to thank BBS for their detailed commentary. We believe that it raises many interesting questions and issues, and we find ourselves in agreement with many of their suggestions regarding future directions.

Looking now: A more skeptical stance

As a friendly amendment to their analysis, however, we believe that a more skeptical stance than theirs is warranted in light of the current published and unpublished empirical evidence. The stance we are advocating is not skeptical in the sense of *denying* early rich ToM, but in the sense of considering it an open empirical question whether we have good reason to belief in such a thing. Neither should this stance be confused with a-theoretical, unconstrained merely empirical exploration. On the contrary, it is firmly theoretically grounded and motivated. The theoretical possibilities, ranging from a very rich nativist via intermediate conceptual change and dual-processing to parsimonious sub-mentalizing accounts are on the table and lay the basis for deriving clear and competing predictions. One can test theories against each other without prior commitments (and thus the potential pitfalls of confirmations biases) to any one of them. And this is exactly the stance, we think, the current evidence warrants: Non-committed, yet theoretically motivated curiosity. On the basis of current evidence, we simply do not know (yet) whether the infant truly has a rich theory of mind.

But are we over-reacting? Are the more than 30 papers with various methods and positive findings referred to by BBS not strong and conclusive evidence that infants indeed have a rich ToM? In our view, one of the lessons from the recent replication crisis in other areas of psychology and of the rise of Open Science awareness and practice is that it will not suffice as a compelling argument to simply mention the mere number of published positive findings. In the absence of more detailed and fine-grained information, in particular about the diversity of labs sampled from, file-drawer problems (how many non-published negative findings match the published positive ones?) and thus potential publication biases, the number of published positive findings per se is of limited epistemic value (e.g. Ioannadis, 2005; Simmons et al., 2011). Not surprisingly, there is evidence for file-drawer problems and publication biases in infant ToM research (Kulke & Rakoczy, 2017), and researchers are now beginning to empty their file-drawers – the special issue being one example.

It will also not suffice as a compelling argument to mention that positive published findings come from studies with various dependent measures. Because for each of these measures for which there have been independent replication attempts, the largely negative results of the latter raise serious doubts about the replicability (and/or validity) of each of the former. For details regarding VoE, interaction and AL measures, see the sections above. Regarding priming and altercentric interference measures (which were not subject of any paper in the special issue), see, for example Conway et al. (2017) for doubts about replicability, and Santiesteban et al. (2014) and Phillips et al. (2015) for doubts about validity.

Finally, it is premature to conclude that we have any convergent evidence, in a relevantly strong sense, for rich infant ToM (a claim made by BBS). In this context, it is essential to keep apart two notions of "convergent evidence". According to a weaker notion, two independent studies both supposed to tap a similar phenomenon with similar findings constitute converging evidence. According to a stronger notion, though, convergent evidence requires cross-validation and thus intra-individual consistency (correlation) across various, superficially different, measures supposed to tap the same underlying phenomenon. Now, regarding explicit ToM, there is much converging evidence not only in the weak, but also in the strong sense: Across superficially very diverse tasks that all share a conceptual deep structure (all require meta-representation), performance converges and correlates (for an overview, see Perner & Roessler, 2012). With regard to research on infant ToM, in contrast, published positive findings only supply convergent evidence in the weak sense. The few studies that have tested for convergent evidence in the strong sense so far have all yielded negative findings (e.g., Dörrenberg et al., 2018; Kulke et al., 2018; Poulin-Dubois & Yott, 2017).

Looking ahead: A constructive and collaborative stance

All in all, we thus disagree with BBS about the seriousness of the replication crisis and its implications for the question of whether infants possess a rich ToM. Luckily, however, this disagreement is only a transient condition, and has the potential to stimulate collaboration and progress. We completely agree with BBS that it is important to speculate about potential differences between positive original and negative replication results (although we do not agree with BBS's implicit assumption that, as a default, original findings enjoy more than merely temporal priority), and to do so in post-hoc ways. Well, how else should one do it now other than post-hoc? But "after-the-fact", if one needs a slogan, is merely *before the next study*.

Post-hoc speculations are potentially helpful as steps towards systematic empirical tests. These can and should be done in two ways. First, meta-analyses on existing published and unpublished findings can explore whether there is indeed evidence that a potential moderating factor speculated about (say, the familiarity of the experimenter) can explain the divergence between positive and negative findings. This approach was adopted in a recent meta-analysis on neonatal imitation that systematically tested for the potential role of many moderating factors that had been suggested in a long-standing debate (results failed to reveal any evidence for an influence of any of these factors; Slaughter, 2018). Second, the silver bullet clearly is large-scale, collaborative, pre-registered, multi-lab replication studies in which potential factors can be implemented as independent variables. Fortunately, such a large-scale collaborative replication project is currently in the planning stage under the umbrella of the *ManyBabies* framework (Frank et al., 2017) and includes many researchers in the

field as participating scientists, including BBS and us (and hopefully soon may more[4]). In several waves of this project (termed "ManyBabies2"), the replicability and validity of different measures of infant ToM such as VoE, AL and interactive measures will be investigated in concerted ways in dozens of labs around the world. In the AL wave, for example, currently in preparation, conceptual replication studies of various existing AL measures with varying complexity (Southgate et al., 2007, Surian & Geraci, 2012) will be implemented in a first step. Depending on the patterns of findings of these replication studies, in a second step there will then be follow-up validation and control studies in order to clarify which kind of implicit ToM processes, if any, these tasks tap. Given this procedure, and given that the project involves researchers that cover the whole ideological spectrum, *ManyBabies2* goes way beyond mere replication research. By designing conceptual replication experiments approved by all participants – in a priori task analysis– as face-valid tests of infant ToM, and by implementing potential moderating factors as variables, the project opens new avenues of "adversarial collaboration" (Mellers, Hertwig & Kahnemann, 2001). Researchers who may hold different theoretical views about infants' social-cognitive repertoire devise empirical test cases the potential implications of which they agree about. In this way, theoretical disputes can be ideally resolved, in collaborative, rational and a priori ways that go way beyond the often unconstrained, unproductive and tiring back and forth between studies and defensive, counter post-hoc speculations from alternating sides.

[4] Researchers interested in participating are very welcome and should email the first authors. A more general call for participation will be issued over the relevant mailing lists once data collection for the replication studies is about to begin.

References

Blakemore, S. J., Sarfati, Y., Bazin, N., & Decety, J. (2003). The detection of intentional contingencies in simple animations in patients with delusions of persecution. *Psychological Medicine, 33*(8), 1433-1441. doi: 10.1017/s0033291703008341

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. doi:10.1016/j.cognition.2009.05.006

Burnside, K., Severdija, V., & Poulin-Dubois, D. (2018). Do infants attribute FB to a toy crane? Manuscript submitted for publication.

Conway, J. R., Lee, D., Ojaghi, M., Catmur, C., & Bird, G. (2017). Submentalizing or mentalizing in a Level 1 perspective-taking task: A cloak and goggles test. *Journal of Experimental Psychology: Human Perception and Performance, 43*(3), 454-465. doi: 10.1037/xhp0000319

Crivello, C., & Poulin-Dubois, D. (2018). Infants' false belief understanding: A non-replication of the helping task. *Cognitive Development*, *46*, 51–57. doi:10.1016/j.cogdev.2017.10.003

Dixson, H.G.W., Komugabe-Dixson, A.F., Dixson, B.J., & Low, J. (2017). Scaling theory of mind in a small-scale society: A case from Vanuatu. *Child Development*. Advanced online publication. doi.org/10.1111/cdev.12919

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12–30. doi.org/10.1016/j.cogdev.2018.01.001

Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental science*, *20*(5), e12445. doi:10.1111/desc.12445

He, Z., Bolz, M., & Baillargeon, R. (2011). False-belief understanding in 2.5-year-olds: evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Developmental Science*, *14*(2), 292–305. doi:10.1111/j.1467-7687.2010.00980.x

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. doi:10.1016/j.jecp.2017.05.005

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... & Lew-Williams, C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. doi:10.1111/infa.12182

Frith, C. D. (2004). Schizophrenia and theory of mind. *Psychological Medicine, 34*(3), 385-389. doi: 10.1017/s0033291703001326

Hinten, A.E., Labuschagne, L.G., Boden, H., & Scarf, D. (2018). Preschool children and young adults' preferences and expectations for helpers and hinderers. *Infant and Child Development*. Advanced online publication. doi:10.1002/icd.2093

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine, 2*(8), e124. doi: 10.1371/journal.pmed.0020124

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental*

Science, *12*(4), 670–679. doi:10.1111/j.1467-7687.2008.00805.x

Köster, M., Ohmer, X., Nguyen, T. D., & Kärtner, J. (2016). Infants understand others' needs. *Psychological Science*, *27*(4), 542–548. doi:10.1177/0956797615627426

Kovács, A.M., Téglás, E., & Endress, A.D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. Science, 330 (6012), 1830–1834. doi:10.1126/science.1190792

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development, 46,* 97–111. doi:10.1016/j.cogdev.2017.09.001

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888-900. doi:10.1177/0956797617747090

Kulke, L., & Rakoczy, H. (2018). *Implicit Theory of Mind – An overview of current replications and non-replications*. *Data in Brief*, *16,* 101–104. doi:10.1016/j.dib.2017.11.016

Low, J., & Edwards, K. (2018). The curious case of adults' interpretations of violation-of-expectation false belief scenarios. *Cognitive Development*, *46*, 86–96. doi: 10.1016/j.cogdev.2017.07.004

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. doi: 10.1177/0956797612451469

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275. doi:10.1111/1467-9280.00350

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. doi: 10.1126/science.1107621

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in cognitive sciences*, *16*(10), 519–525. doi:10.1016/j.tics.2012.08.004

Phillips, J., Ong, D. C., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A Second Look at Automatic Theory of Mind. *Psychological Science, 26*(9), 1353-1367. doi: 10.1177/0956797614558717

Poulin-Dubois, D., & Chow, V. (2009). The effect of a looker's past reliability on infants' reasoning about beliefs. *Developmental Psychology*, *45*(6), 1576–1582. doi:10.1037/a0016715

Poulin-Dubois, D., & Yott, J. (2017). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental science*, *21*(4). Advanced online publication. doi:10.1111/desc.12600

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development, 46*, 40–50. doi:10.1016/j.cogdev.2017.10.004

Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or Teleology? *Cognitive Development*, *46*, 69–78. doi:10.1016/j.cogdev.2017.08.002

Rubio-Fernández, P. (in press). Publication standards in infancy research: Three ways to make violation-of-expectation studies more reliable. *Infant Behavior & Development*.

Santiesteban, I., Shah, P., White, S., Bird, G., & Heyes, C. (2014). Mentalizing or submentalizing in a communication task? Evidence from autism and a camera control. *Psychonomic Bulletin & Review*.

Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological science*, *23*(8), 842–847. doi:10.1177/0956797612439070

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*, *5*(5), 172273. doi:10.1098/rsos.172273

Scott, R. M., & Baillargeon, R. (2017). Early False-Belief Understanding. *Trends in Cognitive Sciences*, 21 (4), 237–249. doi:10.1016/j.tics.2017.01.012

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology:Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366. doi: 10.1177/0956797611417632

Slaughter, V. (2018). *Neonatal imitation: Does it exist?* Paper presented at the Asia-Pacific BabyLab Constellation Conference Singapore.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. doi:10.1111/j.1467-9280.2007.01944.x

Spinrad, T. L., & Stifter, C. A. (2006). Toddlers' empathy-related responding to distress: Predictions from negative emotionality and maternal behavior in infancy. *Infancy*, *10*(2), 97–121. doi:10.1207/s15327078in1002_1

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*(1), 30–44. doi:10.1111/j.2044-835X.2011.02046.x

Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy*, *15*(4), 434–444. doi:10.1111/j.1532-7078.2009.00025.x

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. doi:10.1111/mila.12099

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. doi:10.1016/0010-0277(83)90004-5

Xu, F., Carey, S., & Welch, J. (1999). Infants' ability to use object kind information for object individuation. *Cognition, 70* (2), 137–166. doi:10.1016/S0010-0277(99)00007-4

Yeung, H. H., Denison, S., & Johnson, S. P. (2016). Infants' looking to surprising events: When eye-tracking reveals more than looking time. *Plos One*, *11*(12), e0164277. doi:10.1371/journal.pone.0164277

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. doi:10.1111/j.2044-835X.2011.02060.x

Appendix

More specific and detailed responses to more specific issues in BBS's comment

Appendix A: How to interpret the VoE data by Dörrenberg et al. (2018)

Written by Sebastian Dörrenberg, Hannes Rakoczy & Ulf Liszkowski

In their section on negative findings with violation-of-expectation (VoE) false belief tasks (paragraph 2.1.4.), BBS report re-analyses of our findings from a conceptual VoE replication (Dörrenberg, Rakoczy, & Liszkowski, 2018; Anticipation + Outcome task). Here we clarify that (1) BBS used an arbitrary, unjustified time window for their re-analyses of our data, and (2) did not analyze the full set of data which we provided to her.

BBS correctly remark that we did not use an infant-controlled method of ending the test trial when children looked away for two consecutive seconds. Instead, we accumulated looking time over the whole outcome phase (7 seconds with reaching movement of agent, and 20 seconds with still frame of agent with hand in box). However, the length of our outcome phase was not a randomly chosen time interval, but based on literature review. That is, in most relevant VoE studies, looking time measurement starts after the actor remains still, and children then typically watch incongruent test trials for 15 to 25 seconds (or even longer, up to about 50 seconds) before they look away for two consecutive seconds (e.g., Onishi & Baillargeon, 2005; Scott, 2017; Scott, Richman, & Baillargeon, 2015; Surian, Caldi, & Sperber, 2007; Träuble, Marinović, & Pauen, 2010; Woodward, 1998). Therefore, we chose an adequate but moderate still phase length of 20 seconds, assuming to get neither ceiling nor floor effects in cumulative looking time (which we did not get). BBS's concern that condition differences "dissipate" after a too long test time window, while principally correct, thus does not apply to our design. This is also apparent from our data.

BBS analyzed the first 10 seconds of the still phase. Why not the first 8, 11, or 13 seconds? From the spread-sheet of BBS's re-analyses we gathered that differences were calculated for looking times in congruent and incongruent trials for each two-second-bin of the outcome phase. This data-driven "initial inspection" (p. 114) might be a method of choice when one is in the dark about the timing of an effect, but it appears to us less than ideal when one knows about the typical time window of the effect (which is around 20 seconds), and even less ideal when one can actually check in the data whether infants likely looked away for 2 seconds after 10 seconds had elapsed. Given the two-second-bin data structure, we checked whether across two two-second-bins infants looked away for at least 2 seconds. On this conservative estimate (because the 2 seconds needed not be consecutive), more than half of the children (at least 54% in incongruent as well as in congruent trials) had not looked away for 2 seconds after the first 10 seconds of the still phase, and indeed kept looking much longer into the second 10 seconds period. Thus, splitting the still phase in two halves and only analyzing the first ten seconds is not just arbitrary, it also conflicts with the standard VoE method of ending a trial only when children look away for two consecutive seconds. It is this post-hoc approach that bears the danger of subsequent replication failures.

Still, if one accepts BBS's analyses, there is a looking time difference between incongruent and congruent false belief-based action processing in the first 10 seconds of the still phase, which we can confirm. However, the analysis does not present the full picture, since BBS analyzed only the false belief condition. Looking time data in VoE false belief tasks must be compared to true belief control conditions (e.g., Onishi & Baillargeon, 2005; Träuble et al., 2010) to rule out low-level explanation such as, for instance, heightened attention due to the agent reaching into the full container (unexpected in FB, but expected in TB). Thus, an effect in only one of these two conditions speaks against a high-level interpretation such as false belief processing. Note that BBS claim in their comment that the true belief condition is "a *simple problem ... under high processing demands*" compared to the false belief

condition that is a "*harder problem ... under the same demands*" (p. 114), making it more likely -in their view- that children pass true belief than false belief conditions.

BBS found 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*" (p. 114). This finding does also apply to the true belief condition, where children were more attentive in the first half compared to the second half of the still phase (F(1, 25)=15.89, p=.001). This seems to be what would be expected in VoE tasks, i.e. attention decreases over time until children finally look away for two consecutive seconds. Maybe this pattern could also be found in other VoE studies, e.g. in that of Onishi and Baillargeon (2005) if cumulative looking time was analyzed. Baillargeon further found in her re-analysis *"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"* (p. 114). The corresponding analysis on the true belief condition, however, shows that children's looking time at the unexpected event (M=13.04, SD=4.00) was not different from looking time at the expected event (M=13.00, SD=3.66) overall (F(1, 25)=0.00, p=.963). BBS further reported that "*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*" (p. 114). Again, the corresponding analysis on the true belief condition needs to be reported to complement the picture. Here, only 14/26 (54%) children looked longer at the unexpected event (binomial test, p=.423, one-tailed), 7/12 (58%) who saw the unexpected event first, and 7/14 (50%) who saw the expected event first.

To provide the omnibus test of our analyses on the data of both conditions, now without the last 10 seconds of the outcome phase, a 2x2x2 ANOVA with congruency (incongruent, congruent) as within-subject factor, order (incongruent first, congruent first) and condition (TB, FB) as between-subject factors revealed a pattern that resembled our initially reported findings. There was a significant effect for congruency (F(1, 48)=4.03, p=.050, η_p^2=0.077), a marginally significant interaction between congruency and order (F(1, 48)=3.86, p=.055, η_p^2=0.074), and congruency tended to interact with condition (F(1, 48)=3.05, p=.087, η_p^2=0.060). Testing our hypotheses for each condition separately, we found that in the FB condition infants looked longer during incongruent trials when the incongruent trial was presented first (F(1, 24)=5.17, p=.032, η_p^2=0.177), but not when the congruent trial was presented first (F(1, 24)=1.73, p=.202, η_p^2=0.067). Crucially, in contrast to our original analyses on the whole outcome phase, there were no effects in the TB condition (all ps>.182).

In our original analyses we were careful to report first trial between-subject analyses because the Onishi & Baillargeon (2005) study was a between-subject study (see also BBS's concerns about contamination and order effects). BBS did not report that 'replication' analysis. When we ran the first trial between-subject analysis without the last 10 seconds of the outcome phase, there were no significant effects for either false or true belief conditions (all ps>.161).

References

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited Commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112–124. https://doi.org/10.1016/j.cogdev.2018.06.001

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12–

30. https://doi.org/10.1016/j.cogdev.2018.01.001

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670–679. https://doi.org/10.1111/j.1467-7687.2008.00805.x

Köster, M., Ohmer, X., Nguyen, T. D., & Kärtner, J. (2016). Infants understand others' needs. *Psychological Science*, *27*(4), 542–548. https://doi.org/10.1177/0956797615627426

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Scott, R. M. (2017). Surprise! 20-month-old infants understand the emotional consequences of false beliefs. *Cognition*, *159*, 33–47. https://doi.org/10.1016/j.cognition.2016.11.005

Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology*, *82*, 32–56. https://doi.org/10.1016/j.cogpsych.2015.08.003

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586. https://doi.org/10.1111/j.1467-9280.2007.01943.x

Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy*, *15*(4), 434–444. https://doi.org/10.1111/j.1532-7078.2009.00025.x

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Yeung, H. H., Denison, S., & Johnson, S. P. (2016). Infants' looking to surprising events: When eye-tracking reveals more than looking time. *Plos One*, *11*(12), e0164277. https://doi.org/10.1371/journal.pone.0164277

Appendix B: How to interpret validation studies on VoE stimuli with adults

Written by Jason Low & Katheryn Edwards

Low and Edwards (2018) found that adults perceived the event sequence of Onishi and Baillargeon's (2005) object-location violation-of-expectation (VOE) scenario to be meaningful. However, regardless of whether or not adults were told to track beliefs, participants found Scott and Baillargeon's (2009) object-identity VOE scenario difficult to interpret. Contrary to what has been reported of infants' responses, adults judged the unexpected outcome of Scott and Baillargeon's VOE scenario as being expected, explaining that it was sensible for the agent to reach first towards the penguin toy which was visible. Baillargeon, Buttelmann and Southgate (2018) raised two criticisms.

First, Baillargeon et al. (2018) claimed that adults judged the ending where the agent reached for the transparent cover as being expected because participants interpreted both penguin toys as being "interchangeable containers with removable lids". They claimed that if the event involved interchangeable objects, there would be "less mental effort" for adults to conclude that the agent will reach for the object that was visible to him. It is not clear, however, what it is about that VOE event context that would lead adults—but not infants—to base expectations on simple relational states that require less mental effort. The toy penguins Low and Edwards (2018) used were not designed as containers with lids, and the researchers even ensured that the objects afforded distinct construals (the 1-piece penguin was held by its top when moved and the 2-piece penguin was moved in pieces). Baillargeon et al. revealed that the components of the 2-piece toy in Scott and Baillargeon's study "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". There are two ways to frame Baillargeon et al.'s dismissal of Low and Edwards' choice of penguin toys. On the less positive side, if the VOE effect is an artefact of the penguin toy merely needing to look more like a grapefruit instead of a lidded-container, then researchers will need to be even more skeptical about the claim that an abstract understanding of false-beliefs about object identity develops early in life. On the more positive side, Baillargeon et al.'s dismissal could turn out to be an unexpected victory for Low and Edwards' study in raising attention to a promissory boundary condition that could modulate perception of outcomes in the penguin VOE paradigm. A direction for future research might be to manipulate the two penguin toys' spatial cavities to determine whether and to what extent infants would even track beliefs involving both features and locations of the two toys (e.g., the stacked 2-piece penguin that affords containment is in the transparent box). Future research could test whether infants' belief reasoning would show signature limits in these or similar ways.

Second, Baillargeon et al. (2018) dismissed Low and Edwards' (2018) approach and findings as being relevant to understanding infants' responses to VOE false-belief scenarios because adults have "greater knowledge, experience and reasoning capacity". Baillargeon et al. missed the message of the study. Cognitive-developmental scientists have found it useful to gather converging data across different age groups and populations because similarities and/or differences in response profiles that persist over vast differences in experiences can illuminate how tasks are being interpreted and, further, help characterize diverse mental models of the psychological and physical worlds that influence human beings' attention and action (e.g., Dixson, Komugabe-Dixson, Dixson, & Low, 2017; Hinten,

Labuschagne, Boden, & Scarf, 2018; Kovács, Téglás, & Endress, 2010; Xu, Carey, & Welch, 1999). In lieu of data, researchers often assume how mature mindreaders would behave when faced with highly complex VOE paradigms and then discuss infants' looking expectations as if actual data of people's interpretations existed (see also Silva, Ten Hope, & Tucker, 2014). The popular assumption is that the penguin VOE study showcases infants' ascriptions of false-beliefs about object identity; however, the set-up could just as well be processed by tracking false-beliefs about object types (Butterfill & Apperly, 2013). Without converging data from adults, the field is mired in an unending debate about the meaning of Scott and Baillargeon's VOE paradigm. Low and Edwards' findings are therefore relevant: Even amongst adults who did interpret the expected outcome as being expected, the explanations revolved around the agent making a decision based on the types of object present rather than on object identity in the strict numerical sense. The findings give researchers a starting point for delineating infants' processing of the psychological world that could support tracking false-beliefs in a limited but useful range of situations.

Overall, given that adults find it challenging to interpret VOE false-belief scenarios, researchers should take seriously how extremely challenging it can be to constrain which aspects of a situation may be deemed relevant when interpreting an agent's belief-based behaviour. The main message from Low and Edwards (2018) is clear: Future VOE studies of false-belief understanding should at least seek converging conclusions across multiple participant groups.

References

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited Commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*. Advanced online publication. https://doi.org/10.1016/j.cogdev.2018.06.001

Butterfill, S.A., & Apperly, I.A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28* (5), 606-637. https://doi.org/10.1111/mila.12036

Dixson, H.G.W., Komugabe-Dixson, A.F., Dixson, B.J., & Low, J. (2017). Scaling theory of mind in a small-scale society: A case from Vanuatu. *Child Development*. Advanced online publication. https://doi.org/10.1111/cdev.12919

Hinten, A.E., Labuschagne, L.G., Boden, H., & Scarf, D. (2018). Preschool children and young adults' preferences and expectations for helpers and hinderers. *Infant and Child Development*. Advanced online publication. https://doi.org/10.1002/icd.2093

Kovács, A.M., Téglás, E., & Endress, A.D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. Science, 330 (6012), 1830-1834. https://doi.org/10.1126/science.1190792

Low, J., & Edwards, K. (2018). The curious case of adults' interpretations of violation-of-expectation false belief scenarios. *Cognitive Development*. Advanced online publication. https://doi.org/10.1016/j.cogdev.2017.07.004

Onishi, K.H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308, 255–258. https://doi.org/10.1126/science.1107621

Scott, R.M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80* (4), 1172-1190. https://doi.org/10.1111/j.1467-8624.2009.01324.x

Silva, F.J., Ten Hope, M.I., & Tucker, A.L. (2014). Adult humans' understanding of support relations: An up-linkage replication. *Learning & Behavior, 42* (4), 337-347. https://doi.org/10.3758/s13420-014-0151-0

Xu, F., Carey, S., & Welch, J. (1999). Infants' ability to use object kind information for object individuation. *Cognition, 70* (2), 137-166. https://doi.org/10.1016/S0010-0277(99)00007-4

Appendix C: How to interpret helping False Belief tasks?

written by Josef Perner, Beate Priewasser & Eva Rafetseder

In order to not get embroiled too deeply in the very specific but often undeterminable issues we try to make our general research logic explicit and arrive at three points, which have to be addressed by any successful objections.

Aims of our study:

(1) Replicate the findings of BCT in Experiment 1

(2) Investigate whether representation of the agent's false belief is necessary for more frequent target (box B directed) responses in the FB than the TB condition in Experiment 2.

(3) Make a case for children's use of teleology.

Aim 1 Replication

Despite a comparatively small sample size we managed (though barely) to replicate the finding in its essential feature, which is the difference in target responses between conditions, indispensable for claiming that children are sensitive to the agent's belief.

The difference was not replicated in the old conditions of Experiment 2, plausibly due to the even smaller sample size and an increased error rate due to the middle object distraction. Nevertheless the results do not differ from the results of Experiment 1. Hence there is no reliable evidence that the effect could not be obtained with three boxes.

Aim 2 Mentalism

The, most central point of our investigation was to determine whether BCT's results establish the mentalist claim that infants adjust their intentional (helping) actions to an agent's false belief. To restate our conclusion that their results do not achieve their aim it helps to reflect explicitly on our research strategy.

In general, to show that children of a certain age possess a certain *target ability* we need to show that they display a certain *target response* that could not be displayed unless they possessed the ability in question. To convincingly claim that the target ability is necessary for the target response and that the response could not arise for different reasons, control conditions are needed. Control conditions should ideally be the same as the experimental condition except, of course, for lack of those features that are needed to make the target ability indispensable for the target response to emerge. If the target response occurs more frequently in the experimental than the control conditions one has grounds to claim that children possess this ability. This is classical experimental practice.

In particular, Buttelmann, Carpenter, and Tomasello, et al. (2009: BCT) looked for children's ability to ascribe a false belief to an agent. They contrasted two conditions, each of which consisted of a manipulation and an action phase. In the manipulation phase of both conditions an agent played with a toy and then put it inside one (box A) of two boxes. For the false belief (FB) condition the manipulation consisted of having the agent leave the room, while the toy was being transferred to box B, and have him return afterwards. In the true belief (TB) condition the agent saw the object being transferred, then he briefly walked away, and returned. The action phase was the same in both conditions: upon his return the agent approached the empty box A and tried to open it. The child, who was familiar with the locking mechanism, was asked to help. The majority of children in the TB condition helped the agent with opening the empty box, while in the FB condition they helpfully

directed the agent to the toy's new location. BCT claim that directing the agent to box B more often in the FB than in the TB condition (let us call it the *target effect*) comes about because (some) children have taken the agent's false belief into account. This is a sensible claim from a mentalist's point of view since an agent trying to open a box, where he thinks an interesting toy is, gives a strong indication that he is looking for the toy. The helpful children will, therefore, redirect him to the toy's new location.

To maintain BCT's mentalist claim we need to be able to infer from the presence of the target effect the target ability, which is to take the agent's belief into account. This is possible, given the context of the experiment, only if other reasons for the target effect can be excluded. In our Experiment 2 we checked whether the target effect still obtains when the action relevance of the belief is neutralized, by having the agent open a box in both conditions, which he knows to be empty. This leaves the child without a clue that the agent is looking for the toy. Hence children's tendency to help by directing the agent to the toy's location in the FB condition should not be different from the TB condition. If the target effect should still show up despite these measures then it could not be connected to children attending to the agent's belief.

To implement this strategy we used three boxes instead of two as in BCT. Otherwise the manipulations for creating the true and the false belief were left identical to the original. In the action phase, however, the agent tried open the third box C (new conditions), known to be empty, instead of box A, where he put his toy initially. The results taken from Priewasser et al (2018, Table 3) are shown in the bottom two rows of Table 1. The data show clearly that children in the new FB condition still respond with the target behavior (directing the agent to box B) more often than in the new TB condition. The results are also similar to our replication of the original 2-box conditions of Experiment 1 (first two rows in Table 1). The comparability is also shown by the odds ratios of similar magnitude (last column) and the fact that a cross experiment comparison was not significant ($\chi^2(3) = .44$, $p = .93$). This result implies that under the conditions of the 3-box experiment the target effect can be obtained even when the action relevance of the agent's belief has been muted. Hence, some other factors than consideration of the agent's action relevant belief must have been responsible for the target effect.

Table 1. Number of children's responses directed at one of two or three boxes (data from Tables 2 and 3 of Priewasser et al (2018)

Condition		Agent tries to open	Response directed at box …			Total n	Odds ratio box B vs. (B + box tried)
			agent tries to open	containing toy	third box		
Original 2 boxes	FB	Box now empty	1	13	n.a.	14	9.75
	TB		6	8	n.a.	14	
New 3 boxes	FB	Box always empty	6	18	2	26	7.87
	TB		12	8	7	27	

The next question is whether these alternative factors were also responsible for producing the target effect in BCT's 2-box study. Since the manipulations used to produce the beliefs were the same in the two experiments it seems very plausible that the same alternative factors were active in both studies. Hence we cannot infer from the presence of the target effect in these studies that infants must have represented the agent's belief. Consequently, BCT's study fails to demonstrate an early sensitivity to belief. This conclusion can be avoided if we can show—or plausibly argue—for one of three

possibilities: (1) the use of three boxes instead of two created a difference between conditions in the three box version that was not part of the 2 box original. The agent's trying to open the hitherto untouched box C (2) produced the target effect or (3) left the agent's belief relevant for his action. We find it difficult to come up with plausible arguments for any of the three possibilities. Fortunately, in their commentaries Baillargeon, Buttelmann, and Southgate (2018) and Jacob (2018a,b), who the former cite, made pertinent claims.

Baillargeon, et al. mention four possible features of the 3-box study thought to invalidate our conclusion. (i) The greater complexity of the 3-box study provided "more features to process for participants, which might have influenced the results." (ii) The third box created a tendency to respond with the centre box. (iii) Use of participants from different countries and different testing environments leaves it unclear which factor led to fewer box A responses than in the original BCT study. (iv) Since "children's attention was never directed towards the "always empty" box [C] …, it remains unclear whether it was unambiguous to participants how the agent represented the content of this third box." We find it difficult to see how increased complexity, the attraction of the centre box, or the participation of children from different countries could have (re 1) created a difference between the conditions that wasn't present in the 2 box version, (re 2) could make the agent's search in box C create the target effect, or (re 3) left the action relevance of the agent's belief intact. This is our principled defense for why the commentators' objections do not affect the validity of our conclusion. In addition we can back up our defense with specific answers.

In (i) the commentators argue that the added third box in Experiment 2 increased the processing load and any interpretation of the observed behaviour cannot be applied to the original experiment. In support of their argument they highlight that in the 'old' conditions of Experiment 2 (the agent searches in box A) responses in the TB and the FB conditions should have differed significantly. Although this was not found, response frequencies of the old conditions in Experiment 2 closely mirrored those of Experiment 1 (as outlined above). Moreover, and relevant to argument (ii), children practically ignored the added box as a response option those conditions (only one child chose that box), indicating a systematic choice behaviour rather than processing overload (which would have most likely produced chance behaviour). In (iii) the commentators argue that use of participants from different countries and different testing environments could have undermined replication efforts in our Experiment 1. Since replication only failed for the TB (but not the FB) condition, it remains to be explained, why these factors affected conditions selectively. They further argued (iv) that children's attention was never directed to box C which could have led children to be uncertain about how the agent represented the content of that box. If this was indeed the case, one would predict this ambiguity to affect both new conditions FB and TB in a similar fashion. This would mean that, under the mentalistic approach, no difference was expected, a prediction not confirmed in our data.

Unfolding their arguments Baillargeon, et al., repeatedly allude to the relatively small sample size in our studies and that this would have weakened our conclusions. Although small sample sizes are a good argument against failures to replicate, i.e., fail to achieve significant results due to lack of power, it does not invalidate the significant target effect in the 3-box study, on which our conclusions are based. Furthermore, the fact that our children showed less help with box A (Experiment 1) or box C (Experiment 2) in the TB conditions than in the original study does not invalidate our conclusion. For, we still find the target effect of a significant difference between box B responses in the FB and in the TB condition.

Jacob (2018a) raised–under the heading of "refutation of the mentalistic interpretation"–several further objections:

> (v) "I see no convincing reason why the mentalistic prediction should accept the burden of assuming that children in the new FB condition should behave as they did in the old[5] TB condition."

Since our results show that there are other reasons for the target effect operative in our 3-box version than the agent's false belief, these reasons could not have been miraculously absent in the original 2-box study, unless one can argue that the changes caused by introducing the third box and the choice of box C by the agent could have invalidated these reasons. Reassuringly, Jacob did seem to accept the burden after all by listing three further objections:

> (vi) "When the agent unsuccessfully tries to open box A while knowing that her toy is in box B, young children may assume that she has some reason or other for trying to open box A based on her prior selection of box A over box B for placing her toy, although they do not know her reason. By contrast, in the new FB condition, when the mistaken agent tries unsuccessfully to open box C, children cannot draw on the fact that the agent earlier placed her toy in box A (not C), in order to infer that the agent must have some unknown reason to deal with box C."

This argument does not strike us as very self-evident. One obvious reason for going to an empty box, be it A or C, would be to retrieve the toy but having *forgotten* where it was, or to *place a new object* into an empty box. It is not apparent why such obvious reasons could not be found if the empty box is used for the first time.

> (vii) "In light of the second difference between the old TB and the new FB condition, the mentalistic account is likely to predict that the children will be baffled by the fact that the agent's attempt at opening box C cannot be justified by her false belief that her toy is in box A (as the agent's action was in the old FB condition). They will also be more baffled by the agent's action in the new FB condition than by the agent's action in the old TB condition. In light of the fact that in the new FB condition (but not in the old TB condition), the agent holds a false belief about her toy's location and is also naturally construed as eager to find the toy that she owns, the children are likely to reason that if their goal is to help the agent, then the most efficient means at their disposal is to provide her with her toy (about whose location she has a false belief)."

If we understand the mentalistic predictions correctly, the account would also predict children to help more often in the old FB condition, where the agent tries to open the box where he thinks the toy is, than in the new FB condition, where he tries to open box C, known to be empty. Table 3 in Priewasser et al (2018) however shows no trace of that: 63% B-responses in the old FB condition and 69% in the new FB condition.

To end, we indulge in the unnecessary luxury of speculating about the definitive causes of the target effect despite our insistence that we do not need to be specific for our argument: Our results demand the existence of an alternative cause of the target effect in the 3-box study and it is hard to see how this cause could not also be operative in the original 2-box procedure. Priewasser et al (2018) and Allen (2015) identified three ways in which conditions differed in the manipulation phase: (a) The agent shows clear interest in the toy, (b) the toy seems to belong to the agent, and (c) the toy is transferred under secrecy triggering a hide and seek schema in the FB conditions but not (or less so) in the TB conditions.

The structural analysis of our research strategy, expanded above, made us aware of a fourth very plausible difference that might, to some degree, also appeal to mentalists. (d) Although the

[5] We suspect he meant the original 2-box condition, although we used "old-TB" in the 3-box version of Experiment 2 in which the agent tried to open box A, as in the original, instead of box C as in the new-TB condition.

agent's attempt to open box C instead of A in the new FB condition does cancel the action relevance of his false belief, it does not cancel the relevance of his ignorance of the toy's actual location as action relevant: Someone who knows does not need help; someone who fails to know needs help. Importantly, this difference does not help establish the mentalist account that children's helping shows their ability to represent the agent's false belief. Moreover, although the distinction between knowledge and ignorance suggests some mentalist competence, it does not amount to the ability to represent propositional attitudes, and can also be claimed by teleology, to which we now turn.

Aim 3 Teleology

Since we could show that the response pattern in BCT and our replication was not based on children's mentalising we wondered whether teleology would suffice to account for it. The commentators and also Jacob (2018) raised objections against Allen's (2016) and our intuition that the conspiratorial cue in the FB condition suggested a hide and seek routine to children. Although we once used the term sneakily and spoke of playing a trick we did not mean to suggest that children understood this as acts of deception (as interpreted by Jacob 2018) but simply as a cue to a hide and seek scenario. Children this age experience such a scenario as something where one "hides" (without understanding to suppress all cues to where that is) and then waits until the seeker looks in a different place before being successful. After his first visit to an empty place the children already giggle audibly and often help the seeker to find them. Playing hide and seek in the FB condition, the agent's attempt to look inside the empty box A conforms to the first step in such a scenario followed by helpful cues to where the object really is. The commentators correctly pointed out that in a false belief task such conspiratorial cues increase the likelihood of predicting that the agent will look in the empty place. This, however, does not contradict our application in BCT's helping task, where the agent has already tried to look in the wrong place and children subsequently try to help correct his error.

In our defense against our commentators above we pointed out that the agent's attempt to open box C in the new 3-box conditions made the agent's belief irrelevant for explaining his attempt, but it did not make his ignorance about the location irrelevant. So mentalists can claim that the results may show the involvement of infants understanding of some psychological states like knowledge. In relation to teleology it is important to state that the behavioral impact of certain psychological states does not require an understanding of these states as propositional attitudes. For instance intention action can be understood as action for a good reason (Anscombe 1957; Perner & Roessler, 2010) and knowledge can be understood as a relation between the agent and the event of which he has or does not have knowledge (AGENT knows/does not know EVENT). And knowledge of the relevant facts can be seen as an enabling condition for intentional action. This contrasts with false belief, because one cannot relate an agent to a non-existing event. Instead, one has to conceive of the belief's content as a proposition, i.e., the agent believes that the event has happened.