Research Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–13

DOI: 10.1177/0956797617747090
www.psychologicalscience.org/PS
SAGE

Is Implicit Theory of Mind a Real and Robust Phenomenon? Results From a Systematic Replication Study

Louisa Kulke[1,2,3], Britta von Duhn[1], Dana Schneider[4], and Hannes Rakoczy[1,3]

[1]Department of Developmental Psychology, Institute of Psychology, University of Göttingen; [2]Department of Affective Neuroscience and Psychophysiology, Institute of Psychology, University of Göttingen; [3]Leibniz-ScienceCampus Primate Cognition, Göttingen, Germany; and [4]Institute of Psychology, Friedrich Schiller University Jena

Abstract

Recently, theory-of-mind research has been revolutionized by findings from novel implicit tasks suggesting that at least some aspects of false-belief reasoning develop earlier in ontogeny than previously assumed and operate automatically throughout adulthood. Although these findings are the empirical basis for far-reaching theories, systematic replications are still missing. This article reports a preregistered large-scale attempt to replicate four influential anticipatory-looking implicit theory-of-mind tasks using original stimuli and procedures. Results showed that only one of the four paradigms was reliably replicated. A second set of studies revealed, further, that this one paradigm was no longer replicated once confounds were removed, which calls its validity into question. There were also no correlations between paradigms, and thus, no evidence for their convergent validity. In conclusion, findings from anticipatory-looking false-belief paradigms seem less reliable and valid than previously assumed, thus limiting the conclusions that can be drawn from them.

Keywords

implicit theory of mind, replication, replication crisis, validity, reliability, open data, preregistered

Received 5/16/17; Revision accepted 11/19/17

Theory of mind, the ability to attribute mental states such as beliefs, desires, and intentions to other people and to ourselves, is fundamental to human nature and our social life. Traditionally, theory of mind has been considered to develop between the ages of 3 and 5 years, depending on experience, linguistic input, and central cognitive resources (Perner, 1991; Wimmer & Perner, 1983). Novel findings from the past 10 to 15 years, however, have fundamentally challenged this picture and instead suggested that some form of theory of mind may involve modular rather than central cognitive processes. These findings stem from implicit theory-of-mind tasks that operate without direct verbal measures. A growing body of evidence from studies with such tasks indicates that basic forms of theory of mind can be found even in infants. Further, they may operate in adults in largely spontaneous, automatic, and unconscious ways—even in situations in which there is no need or instruction to engage in theory of mind and no awareness of it (for a review, see Schneider, Slaughter, & Dux, 2017). Implicit tasks include interactive behavioral tasks (Buttelmann, Carpenter, & Tomasello, 2009; Southgate, Chevallier, & Csibra, 2010), violation-of-expectation looking-time paradigms (Kovács, Téglás, & Endress, 2010; Onishi & Baillargeon, 2005), priming (Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010; van der Wel, Sebanz, & Knoblich, 2014), and anticipatory-looking measures (e.g., Clements & Perner, 1994; Low & Watts, 2013; Schneider, Bayliss, Becker, & Dux, 2012; Senju, Southgate, White, & Frith, 2009; Southgate, Senju, & Csibra, 2007; Surian & Geraci, 2012).

A comprehensive pattern of evidence comes particularly from anticipatory-looking false-belief tasks. Such tasks build on standard explicit change-of-location false-belief tasks (Wimmer & Perner, 1983), in which

Corresponding Author:
Louisa Kulke, University of Göttingen, Institute of Psychology, Department of Developmental Psychology, Waldweg 26, 37073 Göttingen, Germany
E-mail: lkulke@uni-goettingen.de

an object is transferred in the absence or presence of a protagonist, with the subsequent test question being where the protagonist will then look for the object. Instead of directly asking the participant, researchers conducting implicit anticipatory-looking tasks capitalize on participants' spontaneous looking behavior: When the agent returns to the scene, will participants anticipate that the agent will search according to his or her true belief (i.e., where the object actually is) or false belief (i.e., where the agent incorrectly assumes the object to be) and look at the corresponding location (Clements & Perner, 1994)? Such tasks can and have been used from infancy to adulthood, suggesting that implicit theory of mind emerges early, remains in operation across the life span, and differs in subtle yet crucial ways between neurotypical and autistic adults (e.g., Senju et al., 2009; Southgate et al., 2007).

Far-reaching theoretical accounts build on these anticipatory-looking findings, including nativists' views (Baillargeon et al., 2015; Leslie, 2005; Scott & Baillargeon, 2017; Wang & Leslie, 2016) and various two-systems accounts (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013), both of which assume that there are early-developing, more or less modular, and automatic forms of theory of mind. From a theoretical point of view, it is currently debated which account best explains findings from anticipatory-looking tasks, given that these findings are robust and reliable.

From an empirical point of view, however, the more fundamental question is whether these implicit theory-of-mind findings are indeed robust and reliable. To date, existing positive evidence in anticipatory-looking tasks still comes from relatively few studies and labs, often with very small sample sizes (e.g., $n < 10$ per condition in Senju et al., 2009; Senju et al., 2010; and Southgate et al., 2007). Recent debates around the replication crisis in many areas of psychology highlight the dangers of publication biases and false-positive psychology. Thus, the robustness, replicability, and reliability of the existing findings from implicit theory-of-mind tasks need to be carefully examined—in particular given their far-reaching theoretical implications (Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011) and given that some studies failed to replicate original findings (e.g., Burnside, Ruel, Azar, & Poulin-Dubois, 2017; Kulke, Reiß, Krist, & Rakoczy, 2017; Powell, Hobbs, Bardis, Carey, & Saxe, 2017; Schuwerk, Vuori, & Sodian, 2015). Furthermore, existing research still does not provide sufficient information about the validity of anticipatory-looking and related implicit paradigms. For explicit theory-of-mind tasks, decades of research have produced convincing evidence for their validity: The convergent validity of various explicit paradigms has been established by correlation analyses (different and superficially dissimilar tasks that all tap theory of mind strongly correlate). Furthermore, for most individual tasks, stringent control conditions have been devised that rule out more parsimonious explanations (Gopnik & Astington, 1988; for a review, see Perner & Roessler, 2012).

For implicit tasks, in contrast, systematic tests of convergent validity are still almost nonexistent. And the only two existing studies that did investigate the convergent validity of various implicit tasks failed to find intertask correlations (Poulin-Dubois & Yott, 2017; Yott & Poulin-Dubois, 2016). In terms of control conditions, some recent studies have taken a closer look at priming paradigms supposedly tapping implicit theory of mind, and the researchers have concluded that once suitable control conditions are administered, the original findings can be explained in alternative, more parsimonious ways (Conway, Lee, Ojaghi, Catmur, & Bird, 2017; Phillips et al., 2015; Santiesteban, Catmur, Hopkins, Bird, & Heyes, 2014). To our knowledge, no such studies exist for anticipatory-looking measures yet.

The rationale of the present study was therefore the following. First, to test the robustness and reliability of anticipatory-looking false-belief tasks, we implemented a systematic, preregistered replication study using original stimuli and procedures and sufficiently large sample sizes. Second, we investigated the validity of anticipatory-looking paradigms by determining their convergent validity across tasks and by testing for alternative explanations for those individual tasks that proved reliable. Studies 1 and 2 included large-scale replications of four different anticipatory-looking paradigms that yielded information concerning the replicability of each individual paradigm and of convergent validity across them. In these studies, there was no evidence for any convergent validation across tasks, and only one paradigm was robustly replicated. Studies 3a and 3b followed-up on these findings using the same paradigm and examining its validity by testing for alternative explanations.

Studies 1 and 2

Method

Studies 1 and 2 used four established anticipatory-looking change-of-location false-belief tasks that have previously been used with infants, children, and adults. The common denominator is that participants see short videos in which a target object of some relevance to a protagonist changes location. This change of location is witnessed or unwitnessed by the protagonist. When the protagonist is about to search for the object, participants' spontaneous belief attribution may manifest itself in anticipatory looking to the location where the protagonist believes the object to be (see Fig. 1).

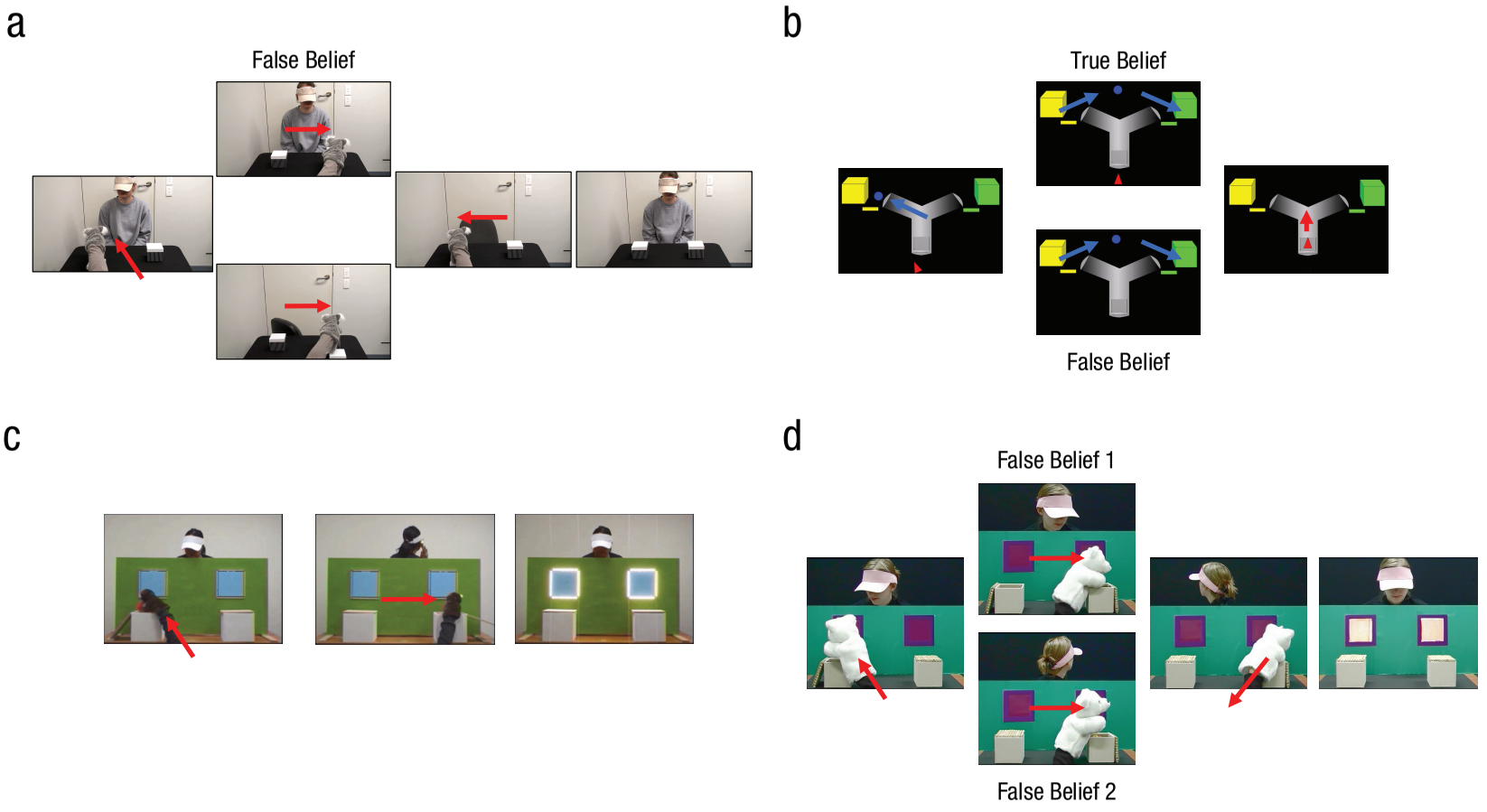

Fig. 1. Sequence of events in the test trials of the tasks of (a) Schneider, Bayliss, Becker, and Dux (2012); (b) Surian and Geraci (2012); (c) Low and Watts (2013); and (d) Southgate, Senju, and Csibra (2007) and Senju, Southgate, White, and Frith (2009). See the Method section for a description of the depicted tasks.

The tasks used by Schneider et al. (2012) and Surian and Geraci (2012) are the most stringent implicit versions of standard change-of-location false-belief tasks, in which an object (an animated ball in Surian and Geraci, 2012; a ball in Schneider et al., 2012) is transferred from Box 1 to Box 2. This is either witnessed by the protagonist (resulting in a true belief about the object's location) or is not (resulting in a false belief about the object's location; the protagonist was an animated triangle in Surian & Geraci, 2012, and a woman in Schneider et al., 2012). The Low and Watts (2013) task[1] is structurally analogous but includes only a false-belief condition. Finally, the task by Southgate et al. (2007) and Senju et al. (2009) is the simplest one because it also does not involve true-belief conditions either, and the target object is removed from the scene rather than transferred in the protagonist's absence. Thus, in the crucial anticipation phase, there is not the same conflict as in the other tasks between the object's real location and the location at which the protagonist believes it to be. This task has two false-belief conditions: False Belief 1 and False Belief 2. In both conditions, the target object is first placed in Box 1, then transferred to Box 2, and then removed from the scene. In the False Belief 1 condition, the protagonist witnesses the first two steps but not the third one. This creates the protagonist's false belief that the object is still located in Box 2, the object's last location before removal. In the False Belief 2 condition, the protagonist witnesses the first step but not the second and the third ones. Thus, now the protagonist should have the false belief that the object is located in Box 1 (which is different from the object's last location before removal).

In the original studies, 14-month-old children (Surian & Geraci, 2012), adults (Schneider et al., 2012), or both children from the age of 2 years and adults (Low & Watts, 2013; Senju et al., 2009; Southgate et al., 2007) revealed anticipatory-looking patterns suggestive of belief-based anticipation in both their first looks and overall looking time.

The current study exactly replicated the methods described in the original articles, using original stimuli and protocols, which were generously shared with us by the authors (for details, see Supplement A in the Supplemental Material available online). In Study 1, adult participants were tested on all four tasks within one session to test the replicability of each individual task as well as the convergent validity across tasks: If these tasks all tap the same underlying capacity, convergent validity of the tasks should be revealed by intertask correlations. Study 2 used only the most stringent of the four tasks, the Schneider et al. (2012) paradigm (which includes false belief and true-belief conditions, multiple trials, and parametric within-participants contrasts), in a single session to rule out the effects of multiple testing or order.

Study 1 was preregistered at the Open Science Framework (https://osf.io/dxb5n/). To reach the predetermined participant number (see Supplement A), we tested 119 neurotypical adults in total (age: M = 23.9 years, SD = 3.68, range = 18–35; 33 men), 43 (36%) of whom passed the original inclusion criteria of all paradigms (because excluded subjects differed between paradigms, only those participants who failed the criteria for each respective paradigm were excluded for the separate analyses of single paradigms; see Supplement A for exclusion specifics for each individual paradigm separately). Study 2 tested 91 neurotypical adults (age: M = 23.3 years, SD = 5.48, range = 18–47 years; 25 men), 9 of whom had to be excluded because of the original criteria. Participants took part in return for course credit or monetary compensation (Study 1: €8, Study 2: €6). All studies were approved by the University of Göttingen Ethics Review Board (Reference No. 143b) and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki.

The original cover task by Schneider et al. (2012) was used. Participants were instructed to closely watch the actress and record (by mouse click) whenever she waved. Participants received the four different paradigms in random order (Study 1) or only the Schneider et al. (2012) paradigm (Study 2). At the end of the session, participants completed a debriefing questionnaire designed by Schneider et al. (2012), which tested awareness about the aim of the experiment with six increasingly specific questions. Participants also completed a German translation of the Autism-Spectrum Quotient to determine their autistic traits (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001).

Anticipatory-looking behavior was measured using eye-tracking technology. Details of the eye-tracking procedures are described in Supplement A. To allow for comparability, we computed the same overall outcome measures for all paradigms, including proportion of looking time (computed for both eyes separately), differential looking score (DLS), and direction of first saccade (previously measured by all studies except Schneider et al., 2012). The DLS is the difference between looking time to the correct side and looking time to the incorrect side divided by the sum of looking time to both sides. To account for the true-belief control condition in Schneider et al.'s and Surian and Geraci's paradigms, we calculated the DLS for the false-belief and true-belief conditions and averaged to get an overall measure of belief-congruent looking (composite DLS; for details, see Supplement A). The time window and area of interest used for looking-time calculations differed between the original paradigms (1.750 s to the left or right side of the screen in Low and Watts, 2013; 6 s to the windows in Senju et al., 2009; 3.5 s to 12-cm × 9-cm

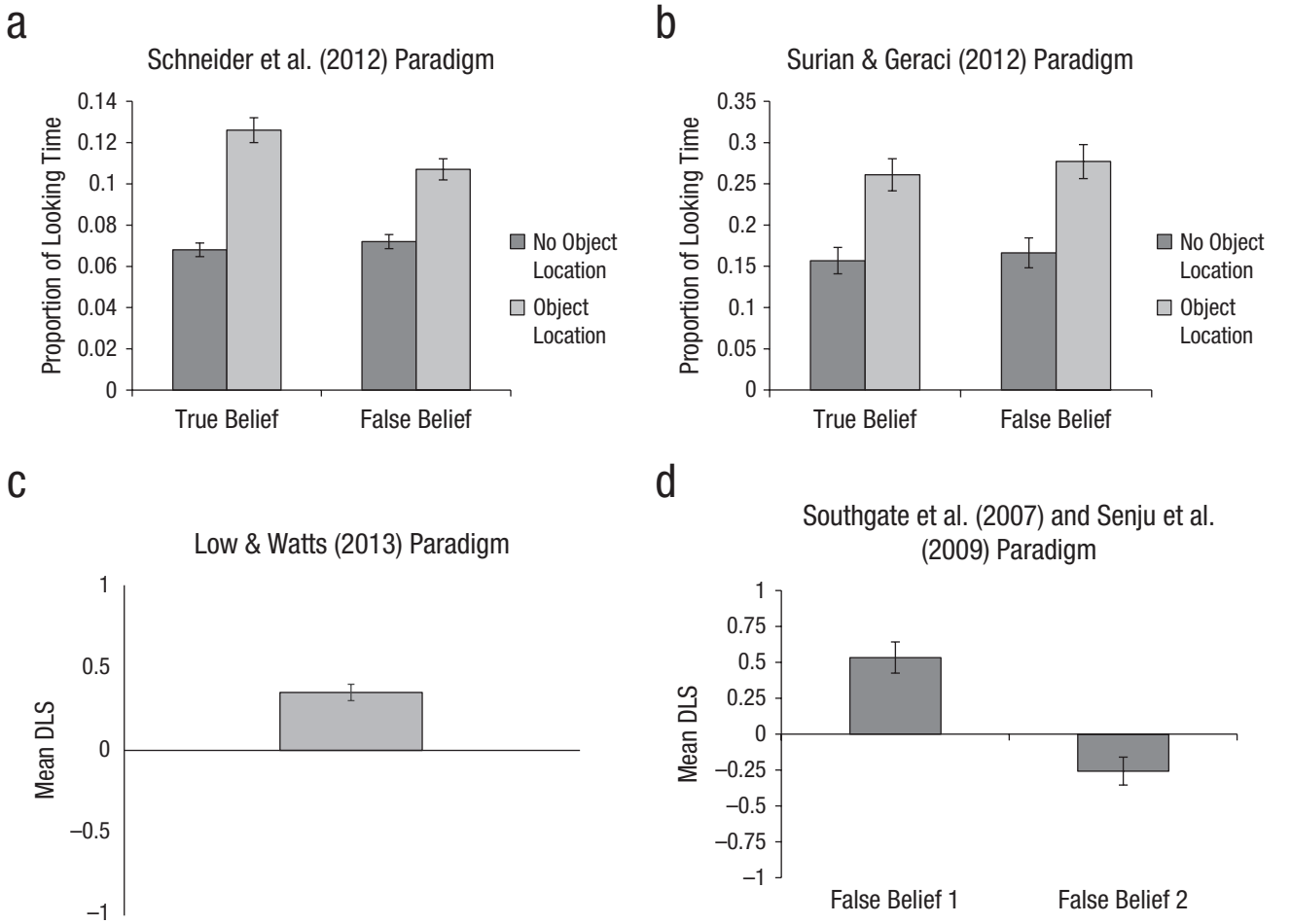

Fig. 2. Results from Study 1, using the tasks of (a) Schneider, Bayliss, Becker, and Dux (2012); (b) Surian and Geraci (2012); (c) Low and Watts (2013); and (d) Southgate, Senju, and Csibra (2007) and Senju, Southgate, White, and Frith (2009). The mean proportion of looking time (a, b) is shown as a function of belief condition and object location. The mean differential looking score (DLS; c, d) is shown as a function of belief condition. Error bars show ±1 *SE*.

areas around the boxes in Surian and Geraci, 2012; and 5 s to the left and right box and arm areas in Schneider et al., 2012). Therefore, looking times are depicted as the proportion of looking toward the area of interest in their originally defined time windows in the following analyses. Follow-up Bayesian analyses were conducted using the *BayesFactor* package (Morey, Rouder, & Tamil, 2015) in the R programming environment (R Core Team, 2012) using Cauchy priors as described by Liang, Paulo, Molina, Clyde, and Berger (2012).

Results of Study 1

Replicability of the individual paradigms. Analyses were based on the preregistration of the current study, unless otherwise noted (for details, see https://osf.io/dxb5n/). As some results deviated in direction from the original findings, two-tailed tests are reported (unlike initially planned in the preregistration). Full data sets can be accessed at https://osf.io/2bvt8/. In the following, for each individual paradigm, we first summarize the original findings and then report the results from the corresponding analyses in the present study (see Fig. 2). Additional preregistered analyses for all paradigms are reported in Supplement B in the Supplemental Material and show results similar to the main findings reported in the following.

Schneider et al. (2012) paradigm. The main results of the original Schneider et al. (2012) study were as follows. There was a significant interaction effect of belief (false belief vs. true belief) and location (object vs. no object) on looking times, with participants looking significantly longer at the no-object location in the false-belief than in the true-belief condition. In the present study (n = 108), the corresponding 2 × 2 analysis of variance (ANOVA) on

proportion of looking time showed a significant effect of object location, $F(1, 431) = 78.69$, $p < .001$, $d = 0.86$; a significant effect of belief, $F(1, 431) = 8.27$, $p = .004$, $d = 0.28$; and an interaction of object location and belief, $F(1, 431) = 6.98$, $p = .009$, $d = 0.26$. Planned t tests showed that there was no significant difference between false-belief and true-belief trials at the no-object location, $t(431) = -0.79$, $p = .429$, $d = -0.08$, Bayes factor in support of the null over the alternative hypothesis (BF_{10}) = 0.074; however, looking times at the object location were significantly higher in the true-belief condition ($M = .13$, $SD = .12$, 95% confidence interval, or CI = [.11, .14]), than in the false-belief condition ($M = .11$, $SD = .12$, 95% CI = [.10, .12]), $t(431) = 3.53$, $p < .001$, $d = 0.34$. Looking times were longer to the object location than to the no-object location in the true-belief condition—object: $M = .13$, $SD = .12$, 95% CI = [.11, .14]; no object: $M = .07$, $SD = .08$, 95% CI = [.06, .08], $t(431) = -8.20$, $p < .001$, $d = -0.79$—and also in the false-belief condition—object: $M = .11$, $SD = .12$, 95% CI = [.10, .12]; no object: $M = .07$, $SD = .08$, 95% CI = [.06, .08], $t(431) = -5.65$, $p < .001$, $d = -0.54$. The original results were thus only partially replicated: As in the original study, there was a significant Belief × Location interaction on looking times. But unlike in the original study, this was not due to the crucial difference in looking to the no-object location (participants looked longer to this location in the false-belief than in the true-belief condition).

Surian and Geraci (2012) paradigm. The main findings of the original Surian and Geraci (2012) study were as follows. There was no interaction effect of belief (false belief vs. true belief) and location (object vs. no object) on looking time. First saccades, however, were more often directed to the no-object location than to the object location in the false-belief condition, and vice versa in the true-belief condition. In the present study ($n = 102$), a 2 × 2 ANOVA on the proportion of looking time also showed no significant effects of belief, $F(1, 201) = 2.02$, $p = .157$, $d = 0.20$, $BF_{10} = 0.079$, and no interaction, $F(1, 201) = 0.03$, $p = .869$, $d = 0.02$, $BF_{10} = 0.116$, but a significant effect of object location, $F(1, 201) = 28.20$, $p < .001$, $d = 0.75$; specifically, participants looked significantly longer at the object location ($M = .27$, $SD = .29$, 95% CI = [.24, .30]), than the no-object location ($M = .16$, $SD = .24$, 95% CI = [.14, .19]). Analyses of first saccades, however, did not reproduce the original findings: Participants looked more often at the object location ($n = 56$) than at the no-object location ($n = 30$) only in the true-belief condition, $p = .007$ (binomial test), $d = 0.34$, but did not look differentially in the false-belief condition—no object: $n = 33$, object: $n = 50$; $p = .078$ (binomial test), $d = -0.20$ (but note a tendency to the belief-incongruent location according to Bayesian analyses: $BF_{10} = 1.335$). The present study thus replicated only the original negative findings but not the positive ones.

Low and Watts (2013) paradigm. The main finding of the original Low and Watts (2013) study was that participants looked proportionally more to the belief-congruent (no-object) location than to the belief-incongruent (object) location, as indicated by a significantly positive DLS. In the present study ($n = 119$), participants also looked proportionally more to the belief-congruent than to the belief-incongruent location, as indicated by a significantly positive DLS ($M = .35$, $SD = .54$, 95% CI = [.25, .45]), $t(118) = 7.16$, $p < .001$, $d = 0.65$. The original findings were fully replicated.

Southgate et al. (2007) and Senju et al. (2009) paradigm. In the original studies of Southgate et al. (2007) and Senju et al. (2009), the main finding was that participants generally showed more anticipatory looking to the belief-congruent than to the belief-incongruent location, with no significant differences between the False Belief 1 and False Belief 2 conditions. In contrast, in the present study ($n = 54$), the overall DLS did not significantly differ from zero ($M = .16$, $SD = .65$, 95% CI = [−.02, .35]), $t(50) = 1.77$, $p = .083$, $d = 0.25$, $BF_{10} = 0.649$. When considering the two false-belief conditions separately, we found that DLSs were significantly positive in the False Belief 1 condition ($M = .53$, $SD = .56$, 95% CI = [.31, .76]), $t(26) = 4.95$, $d = 0.95$, $p < .001$, but significantly negative in the False Belief 2 condition ($M = -.26$, $SD = .48$, 95% CI = [−.46, −.06]), $t(23) = -2.64$, $p = .015$, $d = -0.54$. The overall original findings were not replicated, whereas a differential analysis of the False Belief 1 and False Belief 2 conditions suggests that only the False Belief 1 condition could be replicated.

Relations between the paradigms. Correlations between the DLSs of all paradigms were computed to test for convergent validity. For each correlation between two tasks, the data for a given participant were included if that participant fulfilled the inclusion criteria for each of the two paradigms. Results revealed that there were no significant correlations between mean DLS and composite DLS in any two of the paradigms (see Table 1 for correlation coefficients and Supplement C in the Supplemental Material for detailed analyses and Bayesian statistics).

Relations of anticipatory looking to covariates. Eight percent of participants guessed the aim of the study during the debriefing procedure. Anticipatory looking in the different tasks did not differ between participants as a function of whether they did or did not guess the aim of the study (indicated by the debriefing questionnaire), $F(1, 565) = 2.54$, $p = .111$, $d = 0.13$, $BF_{10} = 0.450$, nor was

Table 1. Overall Correlations Between the Mean Differential Looking Scores in All Paradigms

Study	1	2	3	4	5	6	7	8	9
1. DLS Schneider, Bayliss, Becker, & Dux (2012) false-belief condition	$r = 1.0$ ($N = 103$)								
2. DLS Schneider et al. (2012) true-belief condition	$r = -.14$, $p = .179$ ($N = 101$)	$r = 1.0$ ($N = 102$)							
3. DLS Southgate, Senju, & Csibra (2007) and Senju, Southgate, White, & Frith (2009)	$r = -.03$, $p = .835$ ($N = 43$)	$r = -.16$, $p = .320$ ($N = 43$)	$r = 1.0$ ($N = 51$)						
4. DLS Surian & Geraci (2012) false-belief condition	$r = .00$, $p = .989$ ($N = 74$)	$r = -.05$, $p = .679$ ($N = 73$)	$r = -.23$, $p = .163$ ($N = 38$)	$r = 1.0$ ($N = 88$)					
5. DLS Surian & Geraci (2012) true-belief condition	$r = -.13$, $p = .262$ ($N = 76$)	$r = -.09$, $p = .499$ ($N = 76$)	$r = -.01$, $p = .956$ ($N = 40$)	$r = -.02$, $p = .850$ ($N = 80$)	$r = 1.0$ ($N = 90$)				
6. DLS Low & Watts (2013)	$r = .19$, $p = .053$ ($N = 103$)	$r = .05$, $p = .614$ ($N = 102$)	$r = .22$, $p = .130$ ($N = 51$)	$r = .10$, $p = .349$ ($N = 88$)	$r = -.11$, $p = .319$ ($N = 90$)	$r = 1.0$ ($N = 119$)			
7. Composite DLS Schneider et al. (2012)	$r = .63$, $p = .000$ ($N = 101$)	$r = .69$, $p = .000$ ($N = 101$)	$r = -.20$, $p = .204$ ($N = 42$)	$r = -.04$, $p = .738$ ($N = 72$)	$r = -.18$, $p = .128$ ($N = 75$)	$r = .19$, $p = .054$ ($N = 101$)	$r = 1.0$ ($N = 101$)		
8. Composite DLS Surian & Geraci (2012)	$r = -.10$, $p = .444$ ($N = 66$)	$r = -.06$, $p = .615$ ($N = 66$)	$r = -.24$, $p = .162$ ($N = 35$)	$r = .72$, $p = .000$ ($N = 80$)	$r = .68$, $p = .000$ ($N = 80$)	$r = .02$, $p = .835$ ($N = 80$)	$r = -.13$, $p = .300$ ($N = 65$)	$r = 1.0$ ($N = 80$)	
9. Autism-Spectrum Quotient value	$r = .04$, $p = .703$ ($N = 103$)	$r = -.18$, $p = .067$ ($N = 102$)	$r = -.22$, $p = .127$ ($N = 51$)	$r = -.03$, $p = .811$ ($N = 88$)	$r = .03$, $p = .818$ ($N = 90$)	$r = .04$, $p = .693$ ($N = 119$)	$r = -.12$, $p = .253$ ($N = 101$)	$r = .00$, $p = .995$ ($N = 80$)	$r = 1.0$ ($N = 119$)

Note: DLS = differential looking score.

it related to interindividual differences in autistic traits (for the full analyses, see Supplement B).

Test for order effects. One potential concern about the current study is that it presented all four paradigms to each participant in one test session, with potential effects of order or multiple testing. To rule out such effects of trial order or fatigue, we conducted three exploratory (nonpreregistered) analyses. First, for each task, a mixed model was computed investigating the effect of the position of the task in the testing sequence (1, 2, 3, 4) as well as the interaction effect of position with other factors on DLSs. Results revealed that there were no main or interaction effects of position on DLSs in any of the paradigms (see Supplement D in the Supplemental Material).

Second, for each paradigm, the preregistered analyses were repeated for only those participants who completed this paradigm first, so that any order effects or other kinds of influence of other paradigms on performance in this target paradigm could be ruled out (see Supplement D). The pattern of results was comparable with those found in the analyses on the full participant sample. Taken together, these two analyses suggest that there were no significant effects of paradigm position, justifying the inclusion of all participants, independent of the position in which they completed the paradigm.

Results of Study 2

As an additional test for order effects, Study 2 ($n = 82$) was designed to replicate the Schneider et al. (2012) paradigm in a single session to rule out more stringent multitest or order effects. The findings were comparable with the nonreplication described above (see Fig. 3). A univariate 2 (false belief vs. true belief) × 2 (ball vs. no-object location) ANOVA on the proportion of looking revealed a significant effect of object location, $F(1, 327) = 50.77$, $p < .001$, $d = 0.79$; a significant effect of belief, $F(1, 327) = 10.63$, $p = .001$, $d = 0.36$, $BF_{10} = 0.145$; and an interaction, $F(1, 327) = 7.27$, $p = .007$, $d = 0.30$, $BF_{10} = 0.247$. Follow-up t tests showed that the looking time to the no-object location did not significantly differ between the true-belief condition ($M = .06$, $SD = .07$, 95% CI = [.06, .07]), and the false-belief condition ($M = .06$, $SD = .07$, 95% CI = [.06, .07]), $t(327) = 0.40$, $p = .690$, $d = 0.07$, $BF_{10} = 0.086$, but there was a significant difference between the true-belief ($M = .10$, $SD = .10$, 95% CI = [.09, .11]) and false-belief ($M = .09$, $SD = .11$,

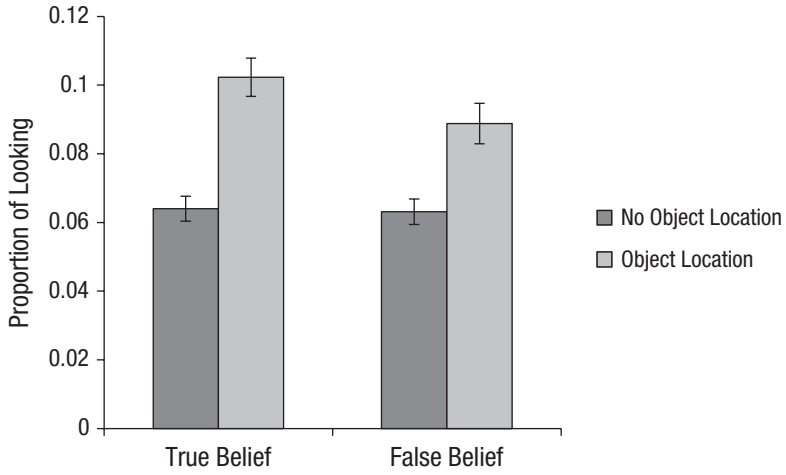

Fig. 3. Mean proportion of looking as a function of belief condition and object location in Study 2. Error bars show ±1 *SE*.

95% CI = [.08, .10]) conditions at the object location, $t(327) = 3.70$, $p < .001$, $d = 0.21$. Further comparisons of looking times toward the different areas of interest showed that participants looked significantly longer at the object location than the no-object location in both the true-belief condition, $t(327) = -8.98$, $p < .001$, $d = -0.33$, and the false-belief condition, $t(327) = -4.53$, $p < .001$, $d = -0.24$. Detailed analyses are reported in Supplement E in the Supplemental Material.

Summary of results

Table 2 summarizes the current results and contrasts them with the original findings. The current studies failed to replicate the previously observed full pattern of the studies of Southgate et al. (2007) and Senju et al. (2009), Surian and Geraci (2012), and Schneider et al. (2012). The only original results fully replicated here were those from the Low and Watts (2013) paradigm.

But how is this pattern of differential replicability to be explained? There are two broad possibilities why only the Low and Watts (2013) paradigm was robustly replicated. One possibility is that this paradigm is particularly valid (perhaps because of lower processing demands or other relevant task factors) and therefore the most sensitive and suitable one to tap implicit theory of mind. The contrary possibility is that this task may be particularly prone to alternative explanations because of potential confounds. To investigate this second possibility, we conducted Studies 3a and 3b.

Studies 3a and 3b

Method

Study 3a (https://osf.io/sy328/) and Study 3b (https://osf.io/3b8tq/) were preregistered on the Open Science Framework. Sample size was predetermined to be 13 participants per condition on the basis of the original study by Low and Watts (2013). Healthy adult participants between 18 and 35 years of age were recruited via posters and leaflets. Participants who did not show first looks in either direction were not included in the first-look analysis; thus, 36 participants were tested for Study 3a (age: $M = 22.4$ years, $SD = 3.09$; 17 women), and 19 participants were tested for Study 3b (age: $M = 23.3$ years, $SD = 1.95$; 9 women) to reach the predetermined sample size for all measures. No participants needed to be excluded because of technical difficulties or lack of attention, and no additional exclusion criteria were applied on the basis of the original study.

Studies 3a and 3b included replications of the methods described in the original article, using and extending original stimuli, which were generously shared with us by the authors. Closer inspection of the stimuli revealed two potential material confounds. First, in the original videos in the two familiarization trials, the object was always placed on the same side. That side was later also the correct (i.e., belief-congruent) side in the test trial (see Fig. 4). Second, in the original stimulus videos, the actress always turned toward the belief-congruent location at the end of the trial, right before the beginning of the time window in which anticipatory looking was measured. Both facts may cue participants' looking behavior toward the belief-congruent location. Study 3a introduced an alternative version of the familiarization trials. Specifically, original videos were cut and remerged to ensure that in one familiarization, the object was placed on the right side, and in the other, it was placed on the left side (Control 1). In the test trial, the belief-congruent side was always on the left. Study 3b introduced a further alternative version of the test trial (Control 2), in which the actress additionally turned toward the belief-incongruent

Table 2. Comparison of Original Findings With Replication Findings

Outcome	Original findings	Findings: Study 1	Findings: Study 2	Success of replication in current study
Southgate, Senju, & Csibra (2007) and Senju, Southgate, White, & Frith (2009)				
First saccade	Correct > incorrect; False Belief 1 = False Belief 2	Correct = incorrect; False Belief 1 = False Belief 2		Not replicated
Looking time	Correct > incorrect; no interaction with condition	Correct = incorrect; significant interaction with condition		Not replicated
Differential looking score (DLS)				
False Belief 1 condition	DLS > 0	DLS > 0		Replicated
False Belief 2 condition	DLS > 0	DLS < 0		Not replicated
Surian & Geraci (2012)				
First saccade				
True belief	Correct > incorrect	Correct > incorrect		Replicated
False belief	Correct > incorrect	Correct = incorrect		Not replicated
Looking time	No interaction of Belief × Location	No interaction of Belief × Location[a]		Replication of null effect
Low & Watts (2013)				
First saccade	Correct > incorrect	Correct > incorrect		Replicated
DLS	DLS > 0	DLS > 0		
Schneider, Bayliss, Becker, & Dux (2012)				
Looking time	Interaction of Belief × Ball Location	Interaction of Belief × Ball Location	Interaction of Belief × Ball Location	Mixed replication/ not replicated
No ball location	False belief > true belief	False belief = true belief	False belief = true belief	Mixed replication/ not replicated
Ball location	False belief = true belief	False belief < true belief	False belief < true belief	Mixed replication/ not replicated

Note: An equal sign (=) means that there was no difference between conditions, a less-than sign (<) means that the first condition had smaller values, and a greater-than sign (>) means that the first condition had larger values in the specified outcome measure than the second one. DLS = differential looking score.

[a]The interaction was determined to be significant if participants were preselected on the basis of behavior that conformed to the study hypothesis in the first-saccade outcome measure.

direction. Study 3a tested, in a between-participants design, the original condition and Control 1 condition, and Study 3b tested the Control 2 condition.

Results and discussion

The full data set can be accessed at https://osf.io/zp76h/. Mean DLSs as a function of condition are depicted in Figure 5. One-sample *t* tests showed that the DLS was significantly positive in the original condition ($M = .29$, $SD = .44$, 95% CI = [.08, .49]), $t(19) = 2.90$, $p = .009$, $d = 0.66$, $BF_{10} = 5.487$, but did not differ from zero in the Control 1 condition ($M = .20$, $SD = .50$, 95% CI = [−.07, .47]), $t(15) = 1.58$, $p = .135$, $d = 0.4$, $BF_{10} = 0.716$, or the Control 2 condition ($M = -.08$, $SD = .77$), 95% CI = [−.47, .30], $t(17) = -0.46$, $p = .652$, $d = -0.10$, $BF_{10} = 0.267$. Planned independent-samples *t* tests showed no significant difference between the original and the Control 1 conditions, $t(34) = -0.56$, $p = .578$, $d = -0.19$, $BF_{10} = 0.366$, and no difference between the Control 1 and the Control 2 conditions, $t(30) = -1.28$, $p = .211$, $d = -0.44$, $BF_{10} = 0.597$, but a marginal difference between the original and the Control 2 conditions, $t(27) = -1.79$, $p = .084$, $d = -0.58$, $BF_{10} = 1.166$ (for additional preregistered analyses, see Supplement F in the Supplemental Material). In sum, the original pattern of belief-congruent looking could be reproduced only under conditions in which the belief congruency of the locations is confounded with additional factors, and therefore, this pattern might not reflect belief-based anticipation.

Discussion

This research investigated the replicability and validity of four major anticipatory-looking false-belief paradigms. Regarding replicability, the results of Studies 1

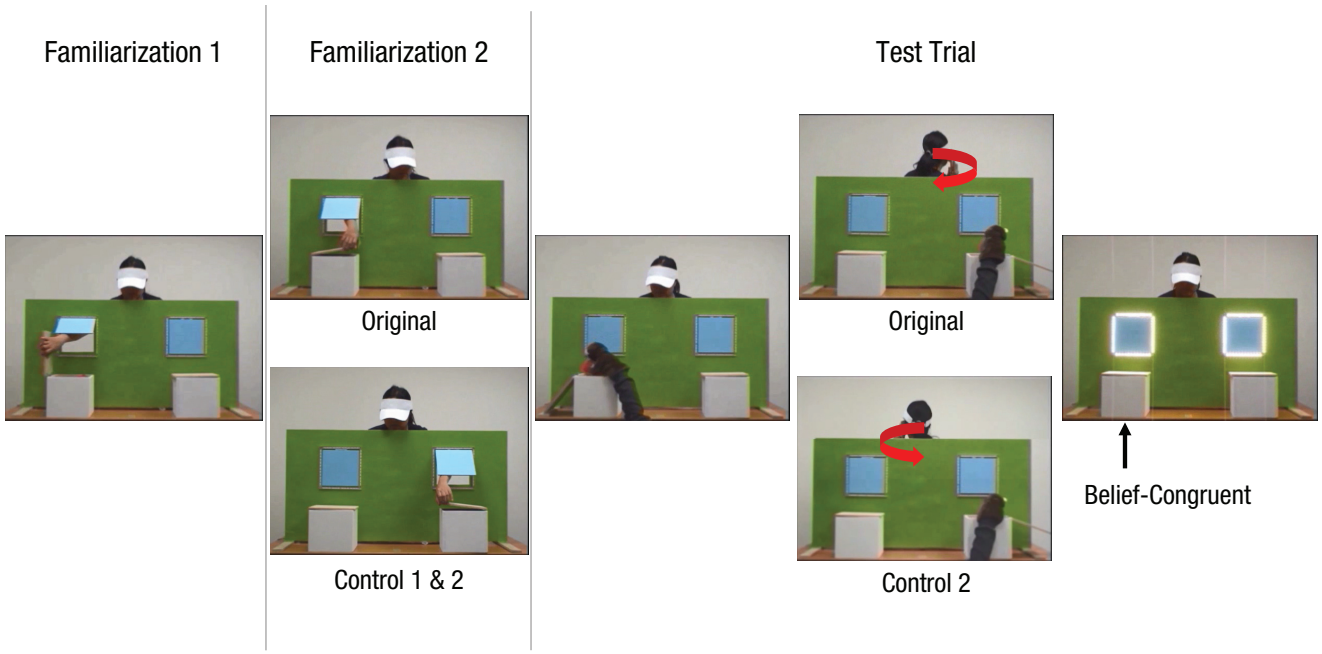

Fig. 4. Sequence of events in Studies 3a and 3b. In the original study (Low & Watts, 2013), both familiarizations and the turn of the actress were directed toward the belief-congruent side. In the Control 1 condition, the second familiarization was to the other side, and in the Control 2 condition, the actress furthermore turned toward the belief-incongruent side.

and 2 show that only one of the four paradigms (Low & Watts, 2013) was fully replicable. Study 1 also investigated performance in the four paradigms in relation to each other and failed to find any evidence for correlations and thus convergent validity. Studies 3a and 3b more carefully and critically investigated the only paradigm that proved robust in Study 1 (Low & Watts, 2013). Results suggest that once potential material confounds are removed, this task no longer reveals anticipatory looking suggestive of implicit theory of mind. Thus, it is possible that the original and replication findings are subject to alternative, more parsimonious explanations, for example, because of salient visual features rather than belief processing driving the observed gaze pattern. Taken together, these findings indicate that anticipatory-looking false-belief tasks are not as reliable and valid as previously assumed.

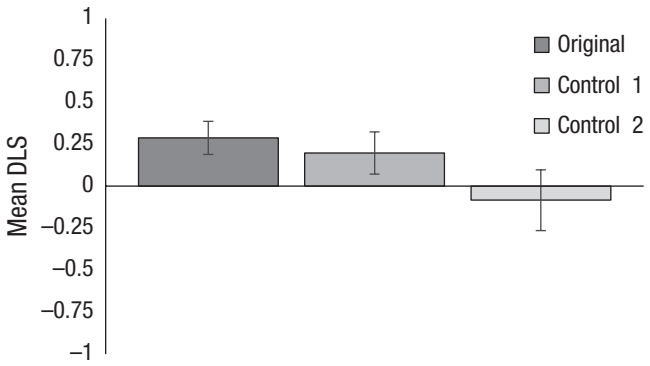

Fig. 5. Mean differential looking score (DLS) as a function of condition in Studies 3a and 3b. The original and Control 1 conditions were tested in Study 3a, and the Control 2 condition was tested in Study 3b. Error bars show ±1 *SE*.

What, in a broader perspective, do the present findings suggest about the status of implicit theory of mind? There are two possibilities. First, implicit theory of mind may be a real phenomenon but fragile and thus difficult to tap. Perhaps anticipatory-looking measures yield informative results only under limited circumstances, for example, when the stimuli are ecologically relevant and the need for anticipation is strong (see Krupenye, Kano, Hirata, Call, & Tomasello, 2016, for a recent argument and study along such lines regarding nonhuman primates). Outside these limited circumstances, participants might simply not be motivated enough to anticipate and thus look back and forth between areas of interest in unsystematic ways. In addition, studies using anticipatory-looking measures have high dropout rates, which further questions their suitability to assess implicit false-belief reasoning.

Alternatively, and more radically, there may be no such thing as a separate and implicit, perhaps even modular form of theory of mind (Heyes, 2014). Rather, there may be only one form of theory of mind that develops as traditionally assumed, in a relatively protracted fashion, building on linguistic experiences and drawing on central cognitive resources (Heyes, 2014).

Some recent findings indeed suggest parallel developments in different kinds of anticipatory-looking and verbal theory-of-mind tasks (Grosse Wiesmann, Friederici, Disla, Steinbeis, & Singer, 2017). According to this second option, what looks like earlier forms of implicit theory of mind in ontogeny, or unconscious and automatic theory of mind across the life span, may in fact be explained more parsimoniously. Evidence along such lines comes from several recent studies (Conway et al., 2017; Heyes, 2014; Phillips et al., 2015; Santiesteban et al., 2014).

By itself, the current study cannot provide enough information to allow us to decide between these two possibilities. Although it failed to find evidence for the reliability and validity of some implicit theory-of-mind tasks, it does not present conclusive evidence that implicit theory of mind does not exist. First, only adult participants were tested in the present studies because previous research showed no differences across the life span in anticipatory-looking tasks (e.g., Senju et al., 2009; Southgate et al., 2007). However, anticipatory-looking tasks were originally designed for children and may thus not be particularly suitable and sensitive measures for adult participants. To draw more comprehensive conclusions, researchers should test further age groups. Second, the current study focused on anticipatory-looking tasks only. Anticipation as a measure compared with other implicit measures, such as violation of expectation or priming, may involve more extraneous task demands that mask participants' competence. Tasks with these other implicit measures thus need to be revisited in independent and systematic replication studies.

In sum, the current study casts doubts on whether the far-reaching theories regarding implicit theory of mind rest on firm and reliable foundations. Thus, there is a strong need for systematic, large-scale, collaborative, and preregistered multilab replication and validation studies to explore more systematically whether implicit theory of mind is a real and robust phenomenon and under which conditions and in which age groups it can be measured.

Action Editor

Rebecca Treiman served as action editor for this article.

Author Contributions

L. Kulke and H. Rakoczy developed the study concept. All authors contributed to the study design. Testing, data collection, and data analysis were performed by L. Kulke and B. von Duhn. L. Kulke and H. Rakoczy interpreted the data. L. Kulke drafted the manuscript, and H. Rakoczy and D. Schneider provided critical revisions. All the authors approved the final manuscript for submission.

Acknowledgments

We thank Luca Surian, Victoria Southgate, Atsushi Senju, and Jason Low for sharing their original stimuli with us. We would like to thank Holger Sennhenn-Reulen for computing the sample sizes required for this project. We also thank the students and research assistants involved in the testing and data processing for this project, particularly Margarita Martens, Luisa Hofberger, Max Hinrichs, and Isabel Ganter.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the German Science Foundation, research unit "Crossing the Borders: The Interplay of Language, Cognition, and the Brain in Early Human Development" (Grant RA 2155/4-1). D. Schneider was supported by Young Researcher Support Grant DRM/2014-02 from Friedrich Schiller University and German Research Foundation (DFG) Network Grant 387279900 (SCHN 1481/2-1).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617747090

Open Practices

All data have been made publicly available via the Open Science Framework (Studies 1 and 2: https://osf.io/2bvt8/; Studies 3a and 3b: https://osf.io/zp76h/). The design and analysis plans for the studies were preregistered at the Open Science Framework (Study 1: https://osf.io/dxb5n/; Study 3a: https://osf.io/sy328/; Study 3b: https://osf.io/3b8tq/). The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617747090. This article has received badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Note

1. This is the location condition of Low and Watts (2013).

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*, 953–970.

Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K.-S., . . . Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In E. Borgida & J. A. Bargh (Eds.), *APA Handbook of Personality and Social Psychology* (Vol. 1, pp. 79–150). Washington, DC: American Psychological Association.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17.

Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2017). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*. Advance online publication. doi:10.1016/j.cogdev.2017.08.006

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*, 337–342.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*, 606–637.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*, 377–395.

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*, 454–465.

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*, 26–37.

Grosse Wiesmann, C., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2017). Longitudinal evidence for 4-year-olds' but not 2- and 3-year-olds' false belief-related action anticipation. *Cognitive Development*. Advance online publication. doi:10.1016/j.cogdev.2017.08.007

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science*, *9*, 131–143.

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

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*, 110–114.

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2017). How robust are anticipatory looking measures of theory of mind? Replication attempts across the life span. *Cognitive Development*. Advance online publication. doi:10.1016/j.cogdev.2017.09.001

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, *9*, 459–462. doi:10.1016/j.tics.2005.08.002

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

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*, 305–311.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes Factors for common designs (R package Version 0.9.12-2) [Computer software]. Retrieved from http://CRAN.R-project.org/package=BayesFactor

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, Article aac4716. doi:10.1126/science.aac4716

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

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

Phillips, J., Ong, D. C., Surtees, A. D., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science*, *26*, 1353–1367.

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

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2017). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*. Advance online publication. doi:10.1016/j.cogdev.2017.10.004

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*, 1255–1266.

Santiesteban, I., Catmur, C., Hopkins, S. C., Bird, G., & Heyes, C. (2014). Avatars and arrows: Implicit mentalizing or domain-general processing? *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 929–937.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*, 433–438.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic theory of mind processing in adults. *Cognition*, *162*, 27–31.

Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism*, *19*, 459–468.

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

Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., . . . Csibra, G. (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., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*, 883–885.

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*, 1359–1366.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*, 907–912.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*, 587–592.

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.

van der Wel, R. P., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition*, *130*, 128–133.

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*, 147–176.

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. (2016). Are infants' theory-of-mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development*, *17*, 683–698.