Cognitive Development xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Cognitive Development

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

How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures

Sebastian Dörrenberg[a,b,*], Hannes Rakoczy[a], Ulf Liszkowski[b]

[a] Department of Developmental Psychology, University of Göttingen, Waldweg 26, 37073 Göttingen, Germany
[b] Department of Developmental Psychology, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany

ARTICLE INFO

Keywords:
Infancy
Cognitive development
Implicit Theory of Mind
False belief
Inter-task correlation
Pupil dilation

ABSTRACT

A growing body of infant studies with various implicit, non-verbal measures has suggested that Theory of Mind (ToM) may emerge much earlier than previously assumed. While explicit verbal ToM findings are highly replicable and show convergent validity, systematic replication studies of infant ToM, as well as convergent validations of these measures, are still missing. Here, we report a systematic study of the replicability and convergent validity of implicit ToM tasks using four different measures with 24-month-olds (N = 66): Anticipatory looking, looking times and pupil dilation in violation-of-expectation paradigms, and spontaneous communicative interaction. Results of anticipatory looking and interaction-based tasks did not replicate previous findings, suggesting that these tasks do not reliably measure ToM. Looking time and new pupil dilation measures revealed sensitivity to belief-incongruent outcomes which interacted with the presentation order of outcomes, indicating limited evidence for implicit ToM processes under certain conditions. There were no systematic correlations of false belief processing between the tasks, thus failing to provide convergent validity. The present results suggest that the robustness and validity of existing implicit ToM tasks needs to be treated with more caution than previously practiced, and that not all non-verbal tasks and measures are equally suited to tap into implicit ToM processing.

1. Introduction

How does our capacity to understand each other as rational agents with an inner life and subjective perspectives on the world, also known as "Theory of Mind" (ToM), develop? An enormous research program in developmental psychology has been devoted to this question over the last decades (Wellman, 2014). Recently, this research has been revolutionized by new studies with novel methods and surprising findings. In contrast to most tasks traditionally used in ToM research that relied heavily on verbal questions, these new studies have developed completely non-verbal and otherwise simplified, implicit tasks suited for testing even very young infants. The findings from these studies have been received as ground-breaking: They suggest that ToM, in particular the capacity to ascribe false beliefs (FB) to other agents – the litmus test for understanding subjectivity (Wimmer & Perner, 1983) – emerges much earlier than previously assumed in the first months of life (for an overview, see Baillargeon, Scott, & Bian, 2016; Baillargeon, Scott, & He, 2010; Scott & Baillargeon, 2017). A converging line of research suggests that these precocious ToM capacities may remain intact and largely automatic over the lifespan, as indicated by findings that adults often seem to engage in spontaneous yet utterly unconscious ToM processing (Kovács, Téglás, & Endress, 2010; Samson, Apperly, Braithwaite, Andrews, & Scott, 2010; Schneider,

* Corresponding author at: Department of Developmental Psychology, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany.
E-mail address: sebastian.doerrenberg@uni-hamburg.de (S. Dörrenberg).

https://doi.org/10.1016/j.cogdev.2018.01.001
Received 15 February 2017; Received in revised form 15 January 2018; Accepted 16 January 2018

Bayliss, Becker, & Dux, 2012; van der Wel, Sebanz, & Knoblich, 2014). Various kinds of such implicit measures have been used with infants, including looking time used with infants as an indicator of violations of expectation (Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007; Träuble, Marinović, & Pauen, 2010), anticipatory looking (Clements & Perner, 1994; Southgate, Senju, & Csibra, 2007; Surian & Geraci, 2012) and interactive measures such as spontaneous helping (Buttelmann, Carpenter, & Tomasello, 2009; Liszkowski, 2012a, 2012b; ; Southgate, Chevallier, & Csibra, 2010). These studies have produced evidence that in their spontaneous looking and interaction behavior, even very young infants seem capable of engaging in FB representation.

From a theoretical point of view, these findings have been taken as evidence for far-reaching theoretical accounts. According to nativist accounts, the findings suggest that ToM is a domain-specific, probably modular, capacity which is online very early in ontogeny and probably even inborn (e.g., Carruthers, 2013; Leslie, 2005). Standard verbal tasks have failed to uncover these early ToM competencies due to extraneous (linguistic and/or inhibitory) performance factors of the tests. According to recent two-systems accounts, the positive findings from the new implicit tasks reflect an early-developing, evolutionarily more ancient, largely automatic and efficient mindreading system. This system is distinct from and potentially the developmental basis for the later-developing, fully-fledged explicit and flexible ToM system tapped in classical verbal tasks (Apperly & Butterfill, 2009; Low, Apperly, Butterfill, & Rakoczy, 2016).

1.1. Robustness, reliability and replicability of implicit ToM tasks

From an empirical point of view, however, it is still unclear how robust, reliable and replicable these results from the novel implicit measures really are. Questions of reliability and replicability of experimental findings have recently taken center-stage in methodological debates about the evidential status of psychological research (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Makel, Plucker, & Hegarty, 2012; Simmons, Nelson, & Simonsohn, 2011; Simonsohn, Nelson, & Simmons, 2014). In this context, systematic replication attempts across many labs often yield negative results such that existing, often classical, effects cannot be robustly reproduced in independent labs (Open Science Collaboration, 2015). As a consequence, the value and necessity of large-scale and systematic replication studies are now virtually ubiquitously acknowledged in cognitive psychology. In research on automatic ToM in adults, questions of replicability and interpretation of existing results with implicit tasks have recently begun to be addressed (Heyes, 2014; Kovács, Téglás, & Endress, 2016; Phillips et al., 2015; Schneider, Slaughter, & Dux, 2017).

Surprisingly, however, hardly anything is known to date about the robustness, reliability and replicability of implicit ToM findings in infants. This is surprising since reliability issues may be particularly pressing in this area of research: First of all, there are still relatively few established infant studies from implicit measures with positive findings, and most of the published studies have used rather small sample sizes and single trial designs, making them vulnerable to spurious findings (but see Scott & Baillargeon, 2017). Second, to date there are no meta-analyses and we currently do not know about the potential body of unpublished failed replication attempts (the so-called *file-drawer problem*). Third, for most of the published studies there have not been any published replications in independent labs. Fourth, in the few exceptional cases where there are published replication attempts (though they are mostly conceptual, and not direct replications, and often administer multiple within-subject conditions), results are often negative (Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2017; Poulin-Dubois & Yott, 2017; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012; Yott & Poulin-Dubois, 2016; Zmyj, Prinz, & Daum, 2015).

1.2. Convergent validity of implicit ToM tasks

A second fundamental question regarding implicit ToM findings in infants concerns their interpretation and validity. Even if individual implicit ToM tasks turned out to be reliable, this would still not settle issues of validity. What is needed are tests of the convergent validity of individual paradigms. If different tasks are in fact all tapping the same underlying cognitive phenomenon – implicit ToM – then they should converge and correlate. Such correlational patterns of superficially different tasks all designed to tap the same underlying phenomenon have been amply documented for explicit ToM (Astington & Gopnik, 1988; Hamilton, Brindley, & Frith, 2009; Perner & Roessler, 2012; Rakoczy, Bergfeld, Schwarz, & Fizke, 2015). For implicit ToM, however, there hardly have been any analogous studies of convergent validation by correlation. One recent study has investigated diachronic correlations between infant implicit and later explicit ToM measures in a longitudinal design (Thoermer et al., 2012). In this study, an implicit measure (anticipatory looking) in a very specific type of FB task (change-of-location) predicted performance in later explicit FB tasks, but only in superficially analogous (change-of-location) ones and not in other FB tasks. Given the very local nature of this correlation, however, this finding leaves open different interpretations in rich (implicit tasks tap the same kind of ToM processes as later explicit ones) or lean terms (the shared variance between the tasks is reducible to commonalities in the surface features).

With regard to studies of synchronic correlations of various implicit ToM tasks at a given time, to our knowledge there are so far only two studies from one lab. One study (Yott & Poulin-Dubois, 2016) tested infants in a VoE FB task (conceptually after Onishi & Baillargeon, 2005) and in other implicit tasks of their understanding of desires and intentions. Results revealed that – in addition to not replicating the original FB task finding – there was no systematic pattern of inter-task correlations comparable to those found in explicit ToM tasks. Another study (Poulin-Dubois & Yott, 2017) examined 18-month-olds' performances between different ToM constructs. These included a VoE FB task (conceptually after Onishi & Baillargeon, 2005) and an interactive FB task (conceptually after Buttelmann et al., 2009), which both could not be replicated and failed to show any correlations. However, given the diverse, yet un-validated and un-replicated, set of further infant FB tasks, more studies are required that use different tasks to test for convergent validity and the robustness of findings.

1.3. Rationale of the present study

Against this background, the rationale of the present study was to test for the reliability and validity of implicit FB measures in infants more systematically and comprehensively. First, in order to examine the robustness and replicability of individual measures, we implemented direct and conceptual replications of structurally very similar implicit ToM tasks, using three different kinds of dependent measures: Anticipatory looking (with the stimuli from Southgate et al., 2007), communicative interaction (Southgate et al., 2010), looking time in a new eye-tracking-based VoE task (conceptually after other VoE studies, Onishi & Baillargeon, 2005; Surian et al., 2007; Träuble et al., 2010). We tested 24-month-olds because (i) this is the youngest age group to perform proficiently in the Southgate et al. (2007) anticipatory looking task, (ii) 20–31-month-olds have been shown to succeed in different VoE tasks (He, Bolz, & Baillargeon, 2011; Scott, 2017; Scott, He, Baillargeon, & Cummins, 2012), and (iii) children have been successfully tested in interactive FB tasks at 2 and 3-years of age (Király, Oláh, Kovács, & Csibra, 2016; Knudsen & Liszkowski, 2012b; Rhodes & Brandone, 2014).

Second, we aimed to test for validation of these measures. In a first step, in exploratory ways, we reasoned that if these different tasks all tap the same underlying phenomenon (implicit ToM), then this phenomenon should be measurable in novel ways as well. Much like, for example, infant categorization processing can be tapped in analogous ways by various behavioral and physiological measures (e.g., Elsner, Pauen, & Jeschonek, 2006), infant implicit ToM should reveal itself in various novel behavioral and physiological parameters. In a first step in this direction, a recent study found a novel neurophysiological signature of belief-based action prediction, i.e. mu-desynchronization measured with EEG revealed that infants predicted an agent, who wanted an object from a box but held a false belief about the content of the box, to reach into the empty box, or not to reach into the full box (Southgate & Vernetti, 2014). Here, we took a complementary approach by using a novel pupillometrical measure, in addition to looking times, in a VoE FB task. Pupil dilation measures have recently begun to be used in developmental research as another window into the infant mind (Hepach & Westermann, 2016). Increase in pupil size, if not due to luminance, typically indicates arousal and heightened levels of attention. We thus reasoned that violations of expectation should lead to heightened levels of attention and an increase in pupil size, which should correlate with the pattern of looking times (Gredebäck & Melinder, 2011; Jackson & Sirois, 2009). Following this logic, we created a new eye-tracking-based VoE FB task, where we showed infants scenarios in which an agent had a false belief and then acted either in belief-congruent or belief-incongruent ways (looking for the object where he falsely believed it to be or where it really was, respectively) and measured both looking times and pupil dilation. If infants really engage in spontaneous implicit FB processing and thus expect the agent to search in belief-based ways, they should look longer, and show increased pupil dilation in response to the unexpected outcome. Measuring looking times with an eye-tracker has been established in various VoE studies (e.g., Jackson & Sirois, 2009; Köster, Ohmer, Nguyen, & Kärtner, 2016; Yeung, Denison, & Johnson, 2016). We also measured pupil size changes in response to the induction of true beliefs (TB) versus false beliefs, because witnessing the induction of a false belief might already lead to heightened arousal, as has been shown in the emotional expressions of slightly older children (Moll, Kane, & Mcgowan, 2016; Moll, Khalulyan, & Moffett, 2017).

In a second step to test the validity of different implicit FB tasks, we examined their convergent validity by testing for correlations between the four different measures more generally: Anticipatory looking (Southgate et al., 2007), looking times as an indicator of violation-of-expectation as well as pupil dilation (in our new eye-tracking-based VoE task), and spontaneous communicative interaction (Southgate et al., 2010). Since these are all implicit ToM tasks that were specifically designed to reduce processing load and were mastered by the majority of infants in previous studies, we assumed that differences in task demands across tasks should be minimal. Thus, if these tasks indeed all conceptually measure the same, namely implicit forms of representing an agent's belief, as assumed by early mindreading accounts and to some extent by two-systems accounts, then they should converge and correlate.

To replicate previous findings, we were careful to include single trial analyses in all our tasks and measures, as in the original studies. Therefore, we made sure our within-subject sample was sufficiently large to allow for between-subject analyses of the first trial of each task. In addition, we ran within-subject analyses across several trials, because these analyses are based on the larger sample and have more power. Further, we made sure to test for predicted effects directly with planned comparisons, and report one-tailed results for the planned comparisons when appropriate. While this is a more lenient procedure, it would make it more likely to replicate previous findings. To validate previous findings, we looked for correlations across the different measures, and in a more exploratory step, at selective measures and composite scores that were most relevant given the pattern of findings.

2. Material and methods

2.1. Participants

66 German 24-month-olds (median age = 24 months; 16 days; age range = 24;4–25;0; 36 girls and 30 boys) from mixed, mostly middle-class, socioeconomic backgrounds in the metropolitan city Hamburg were recruited from a databank of children whose parents had previously agreed to participate in infant studies. All infants participated in the Anticipatory Looking (AL) task; 35 of the infants were tested in the false belief conditions of the Anticipation + Outcome (A + O) task and the Interaction task; and 31 of the infants were tested in the true belief conditions of the A + O task and the Interaction task.

When gaze samples in the video-based tasks were below 70%, we rated the quality based on the gaze replay to reduce the drop-out rate. We included those participants only if we had gaze data during all relevant events (e.g., Teddy changing ball locations, agent reaching through door in outcome phase). After these data reduction steps, our participants had mean weighted gaze samples of 85% (SD = 14) in the AL task and of 80% (SD = 17) in the A + O task. To rule out effects of the amount of tracked gaze samples, we report additional analyses where we only included participants with higher amounts of gaze samples (see Table A1, Appendix B).

Nine infants were excluded from the AL task because of poor gaze data quality (6) or because they showed no anticipatory looks (3), which resulted in an N of 57 infants. In the A + O task, nine infants were excluded from the false belief condition and five from the true belief condition because of poor gaze data quality (11), fussiness (2) or experimenter error (1), which resulted in an N of 26 infants per condition. In the Interaction task, three infants were excluded from the false belief condition (resulting in an N of 32 infants) and three from the true belief condition (resulting in an N of 28 infants) because they refused to participate.

2.2. Design

All infants were tested in three different non-verbal change-of-location paradigms in the following order: a video-based AL task with two false belief versions (original stimuli of Southgate et al., 2007), an interaction-based task with a true belief condition or a false belief condition (adopted from Southgate et al., 2010) and a new eye-tracking-based VoE task that included an anticipation phase comparable to Southgate et al. (2007) but in addition a belief-congruent and a belief-incongruent outcome and a true belief condition or a false belief condition (A + O task). True and false belief conditions were administered between subjects to avoid confusions and longer testing. We chose the least exhausting and biasing task order: first, the AL task was the shortest task and never showed a belief-based outcome, so that it could not reduce belief-based action expectations for the next tasks; second, because calibration ceases over time, we could not administer the AL and A + O tasks back-to-back. The Interaction task was more like a fun game with interesting toys and thus served as a natural break between the eye-tracking tasks; third, the A + O task included belief-incongruent outcomes which could confuse infants or affect their expectations of the agent's actions in later tasks and therefore needed to be administered last. Keeping the task order the same also reduces irrelevant variation which is advisable for correlational analyses.

2.3. Set-up and procedure

2.3.1. Video-based tasks

2.3.1.1. Eye-tracking set-up. Infants were fastened in a car seat with a headrest to minimize mobility, and watched film clips (25 fps, 1280 × 1024 pixel) on a 24 in. screen (Dell U2412M) from a viewing distance of approximately 65 cm. Display resolution was set at 1920 × 1200 pixel. The screen was surrounded by 2.5 m high walls made of black stage cloth. The room had no windows, and room luminance (emitted from the ceiling) was kept constant across all tasks and participants. Sound was played via powered speakers that were hidden behind the screen. Parents were seated centrally behind the infants and were instructed not to interact. Twelve infants sat on their parents' laps because they refused to sit in the car seat. A Tobii (Stockholm, Sweden) X120 eye-tracker was installed underneath the screen and recorded infants' eye movements with a sampling rate of 120 Hz. Stimulus presentation and recording were controlled via a Dell Latitude E6530 notebook using Tobii Studio software. We used a 5-point infant calibration. Between the test trials, we showed infants 6 s long attention getter videos depicting fun cartoons, e.g. a train or cute bugs, which emitted sounds, to keep attention on the screen.

2.3.1.2. Stimuli and procedure

2.3.1.2.1. Anticipatory Looking task. For the Anticipatory Looking task, we used the original video clips of Southgate et al. (2007). For a detailed description of this task see Appendix A. Infants watched an agent repeatedly reach through one of two windows for a toy hidden in one of two boxes. In the FB1 condition, the agent then witnessed a teddy changing the toy from Box 1 to Box 2, but did not witness the teddy thereafter removing the toy from the scene (the agent falsely believed the object to be in Box 2). In the FB2 condition, the agent did not witness the teddy changing the toy from Box 1 to Box 2 and then remove it from the scene (the agent falsely believed the object to be in Box 1). If infants' anticipatory looking was belief-based, in the FB1 condition infants should anticipate the agent to reach for the location were the object was last ("box 2"); and in the FB2 condition, they should anticipate the agent to reach for the location were the object was first ("box 1"). Since the FB1 condition could simply elicit looks to the ball's last location, FB2 controls for this issue, because the ball's last location is different from the agent's belief of the ball location.

Different to the original study, we mirrored the video clips in order to counterbalance between the target sides. Mirroring had no influence on infants' performances in both test conditions (see Appendix B for analyses). Further, before the two familiarization trials, we showed infants two warm-up trials, in which infants saw the agent reaching through the door for a red whale toy that was sitting on a box (one trial on each box), a procedure that has been used by the authors for the same task in follow-up studies (Senju et al., 2010; Senju, Southgate, Snape, Leonard, & Csibra, 2011; Senju, Southgate, White, & Frith, 2009). In contrast to Southgate et al. (2007) who used a between-subject design, all infants in our study saw both the FB1 condition and the FB2 condition in counterbalanced order. A between-subject analysis was still possible by using the first trial performance.

2.3.1.2.2. Anticipation + Outcome task. For our new eye-tracking-based VoE task, we recorded video clips based on the stimuli of Southgate et al. (2007), but included a belief-congruent or a belief-incongruent outcome and a FB and a TB version, as in other VoE tasks (Onishi & Baillargeon, 2005; Surian et al., 2007; Träuble et al., 2010). The videos thus contained familiarization trials, a belief-induction phase, and belief-incongruent outcomes. The agent disappeared during the belief-induction, and the desired object remained in the changed location. For a detailed description of this task see Appendix A. After infants had watched the object placements, the agent duck behind the screen, as in other anticipatory looking studies where the agent disappears before the anticipation phase (Clements & Perner, 1994; Surian & Geraci, 2012; Thoermer et al., 2012), and infants saw an anticipation phase comparable to Southgate et al. (2007) followed by a fixation cross that appeared between the two doors before the outcome to reorient infants' gaze before the outcome phase started. All infants saw a belief-congruent reaching trial and a belief-incongruent

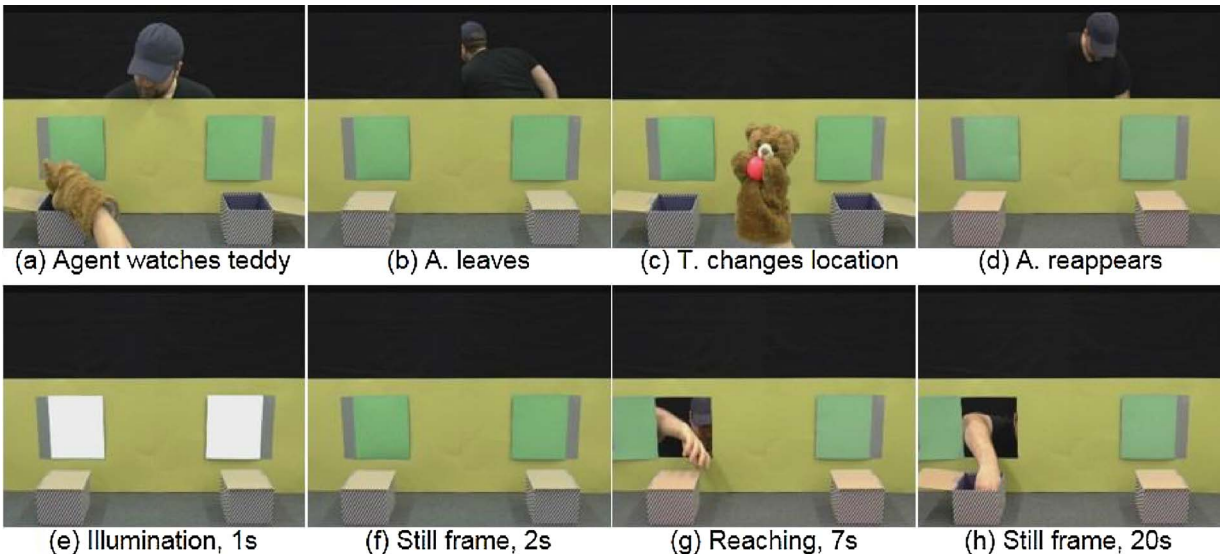

Fig 1. Selected scenes showing the main events in a false belief trial of the Anticipation + Outcome task in consecutive order (a)–(h). (a) Agent watches teddy during the first change of ball location; (c) Teddy changes ball location; (d) Agent reappears before (TB condition) or after location change (FB condition; depicted is the FB condition); (e) Agent ducks behind the screen before illumination; (g, h) Outcome phase, either belief-congruent or belief-incongruent.

reaching trial in counterbalanced order (thus yielding two anticipation trials per infant). We used an eye-tracker to measure looking times, instead of coding live by hand. However, the general procedure was similar: Fig. 1 shows selected scenes from the video clips describing the main events. Note, all infants saw the same film clips during the outcome phase (target side counterbalanced left or right) to ensure that there were no spatial or luminance differences between outcomes or conditions that could affect pupil size.

2.3.2. Interaction task

2.3.2.1. Set-up. The testing room was 3.7 m × 3.5 m in size, had white walls, a door in one corner and three cameras in the other corners recording the experimental procedure. Infants were seated on the floor in front of their parents, who leaned against a wall. A blue and a green box (L: 27 cm, H: 34 cm, W: 20 cm) were placed 120 cm from the infant and 100 cm apart. The front of the boxes facing the infants could be opened so that they remained in an upward position. For the two warm-up trials, we used a yellow bath duck and a small yellow shovel as objects. Fig. 2 shows the three different object pairs we used in the three test trials: (1) a purple lemon squeezer and a red funnel, (2) a black watering can spout with colorful glue strips and a yellow plastic toy ring with colorful glue strips, and (3) a purple pastry scraper and a grey piece of tube with colorful glue strips. For each pair we used a different novel label for the target object: (1) Sefo, (2) Toma, and (3) Nari.

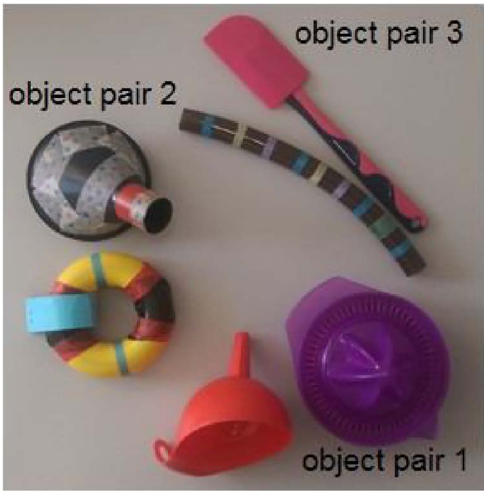

Fig. 2. Object pairs used in the three test trials of the Interaction task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3.2.2. Procedure. We adopted the experimental procedure of the Sefo-task in a FB and a TB version (Southgate et al., 2010, Experiment 1). An experimenter requested infants to retrieve one of two objects he either correctly or falsely believed to be in one of two boxes. In the TB condition, the agent witnessed another person swapping the objects, thus infants were required to retrieve the toy from the box that was indicated; in the false belief condition, on the contrary, the agent was outside during the object changes, infants were thus required to retrieve the toy from the opposite box which the experimenter indicated. For a detailed description of this task see Appendix A. Instead of one test trial, we administered three test trials in the same condition (either TB or FB) to gather continuous data for our correlational analyses. A first trial between-subject analysis was still possible.

2.4. Coding and analyses

For the video-based tasks, the two doors served as areas of interest (AOI) for the analyses during the anticipation phase. We measured first fixations starting onset of the illumination until 1.75 s after offset of the illumination (2.75 s in total) using Tobii Studio software (I-VT fixation filter). Infants could score 1 (first fixation in AOI of correct door; AL task: FB1 = last object location, FB2 = first object location; A + O task: FB = empty box, TB = full box) or 0 (first fixation in AOI of incorrect door; AL task: FB1 = first object location, FB2 = last object location; A + O task: FB = full box, TB = empty box). For the looking times between the two doors in the anticipation phase, we analyzed raw data using customized R scripts, and measured from onset of the illumination until 1.75 s after offset of the illumination (2.75 s in total), as authors did in follow-up studies (Senju et al., 2010, 2011, 2009). We calculated differential looking scores (DLS) for the anticipation phase by subtracting the looking time to the incorrect door from the looking time to the correct door and dividing it by the sum of both (a value of 0 would indicate no preference for one door; a value above 0 would indicate longer looking to the correct door; a value below 0 would indicate longer looking to the incorrect door). To compare results to the original analysis of Southgate et al. (2007), we also analyzed the total looking time instead of the DLS, and used the second familiarization trial as an inclusion criterion for the analyses of the test trials. Due to some ambiguity in the original Southgate et al. (2007) description of the analyses and stimuli, we report additional analyses of our measures for the time period of 1.75 s starting after offset of the illumination for the AL task in Appendix B.

In the A + O task, we measured looking times and mean pupil size during the outcome phase (reaching phase plus still frame phase, 27 s) and pupil size additionally at a time point at the beginning of the outcome, in the fourth second of the reaching phase. Note that the pattern of looking time results remained the same when analyzing only the still frame phase of the outcome phase (without the reaching phase). We used the total screen (size of video) as AOI and analyzed raw data using customized R scripts. Mean pupil size of left and right eye was computed at each sample. To analyze changes in pupil size, we calculated the relative change in pupil dilation (RCD) by subtracting a baseline from the focal phase and dividing it by the baseline (baselines are described in the result section). Tobii Studio calculates pupil size by taking distance to the stimuli into account (Tobii AB, 2016). To make sure, minor changes in infants' posture did not affect pupil size calculation, we ran subsidiary analyses (see Appendix B). First, we used the distance scores from Tobii as the dependent variable in the same manner as we did for the pupil dilation, to see if distance alone would yield similar results as pupil size change. Further, we correlated pupil size and distance to the eye-tracker. Both analyses confirmed that pupil size was not influenced by posture changes. Posture changes were minor anyhow, because infants were fastened in a car seat with a headrest.

In the Interaction task, we coded the box that was first approached or pointed to by the infant. Infants could score 1 (referred box; correct in TB, but incorrect in FB) or 0 (non-referred box; incorrect in TB, but correct in FB). We calculated a mean performance over the repeated trials, ranging from 0 (no trial correct) to 1 (all trials correct). A subsample of 12 FB and 12 TB participants (a total of 72 trials) was additionally analyzed by a second coder. Inter-rater reliability was excellent (Cohen's $k = 0.944$, $p < .001$).

To analyze relations between the measures, we used Pearson correlations for metric variables, and phi correlations for dichotomous variables. All statistical tests were performed in IBM SPSS Statistics Version 23. Alpha was set at 0.05. All presented *p*-values are two-tailed if not mentioned otherwise. We report lower and upper limits of 95% confidence intervals (CI).

3. Results

For each of our four measures we first report the first trial between-subject performance, to compare it to the original studies. To be most lenient in achieving replication results, we report one-tailed analyses for those comparisons that have revealed significant results in previous studies. We then report analyses on the full set of our data, to seek for confirmatory support with the larger sample. Where appropriate we use order as between-subject factor and report results for specific orders. Because previous findings predict effects specifically for false belief conditions, we analyze conditions also separately.

3.1. Anticipatory Looking task

3.1.1. Familiarization trials

48 participants provided data for the first familiarization trial, 51 for the second familiarization trial, and 43 infants for both trials. Across the two familiarization trials, the first look was significantly more often directed to the correct door than expected by chance (M = 0.605, SD = 0.279; $t(42) = 2.46$, $p = .018$, $d_z = 0.38$, CI: 0.019, 0.191). In the first trial, 58% of the infants directed their first look to the correct door (binomial test, n = 48, $p = .312$, odds ratio (OR) = 1.38). In the second trial, infants directed their first look significantly more often to the correct door than expected by chance (65% correct; binomial test, n = 51, $p = .049$, OR = 1.86). There was a negative correlation between infants' first look in the first familiarization trial and in the second

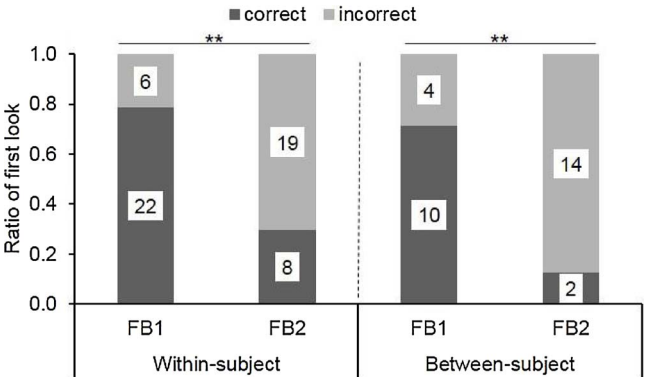

Fig. 3. Anticipatory looking task. Percentage of infants who passed the second familiarization trial and first looked to the correct or the incorrect door in the FB1 and the FB2 conditions. The left panel shows a within-subject analysis and the right panel shows the between-subjects replication analysis of Southgate et al. (2007) for the first trial. Numbers in bars show number of infants. $^{**}p < .01$.

familiarization trial ($phi(43) = -0.361$, $p = .018$). Analyses with the DLS as dependent measure revealed a similar pattern. Infants tended to look longer to target than distractor across the two trials (M = 0.12, SD = 0.44; $t(42) = 1.84$, $p = .073$, $d_z = 0.28$, CI: −0.012, 0.258); and in the second trial (M = 0.19, SD = 0.71; $t(50) = 1.87$, $p = .068$, $d_z = 0.26$, CI: −0.014, 0.387); but not in the first trial (M = 0.07, SD = 0.79; $t(47) = 0.64$, $p = .527$, $d_z = 0.09$, CI: −0.157, 0.302). Also for the DLS, the first and the second familiarization trial tended to correlate negatively ($r(43) = -0.296$, $p = .054$).

3.1.2. Anticipation phase

3.1.2.1. First trial analyses. To exclude any effects of repeated trial exposure, we analyzed the first trial separately for each of the two conditions, and compared these to each other. When analyzing the first trial of those infants who had looked at the correct door in the second familiarization trial, 14 participants provided data for the FB1 condition and 16 for the FB2 condition.

3.1.2.1.1. First look. The right panel of Fig. 3 shows the first look results for the replication analysis of the finding by Southgate et al. (2007). In the FB1 condition, 71% of the infants directed their first look to the correct door, which was not different from chance (binomial test, n = 14, $p = .090$, one-tailed, OR = 2.45). Contrary to the original study, in the FB2 condition, infants' first look went significantly more often to the incorrect door than expected by chance (13% correct; binomial test, n = 16, $p = .004$, OR = 6.69). Further, contrary to the original study, the two conditions differed significantly from each other (Fisher's exact test, n = 30, $p = .002$, $\phi = -0.600$). The pattern of results for our analyses with measurements starting offset of the illumination was similar, see Appendix B.

3.1.2.1.2. DLS. In the FB1 condition, the DLS did not differ from chance (M = 0.316, SD = 0.819; $t(13) = 1.44$, $p = .088$, one-tailed, $d_z = 0.40$, CI: −0.157, 0.789). In the FB2 condition, the DLS revealed that infants looked significantly longer to the incorrect door than expected by chance (M = −0.450, SD = 0.665; $t(15) = 2.71$, $p = .016$, $d_z = 0.70$, CI: −0.805, −0.095). Condition differed significantly from each other ($t(28) = 2.83$, $p = .009$, $d_s = 1.07$, CI: 0.211, 1.32).

3.1.2.1.3. Looking time. To be as close as possible to the original analyses, we additionally analyzed the total looking time to each door for the first trial for those who passed the second familiarization trial. A 2 × 2 ANOVA with window (correct, incorrect) as within-subject factor and condition (FB1, FB2) as a between-subject factor yielded no significant main effects of window ($F(1, 28) = 0.03$, $p = .428$, one-tailed, $\eta_p^2 = 0.001$, CI: −353, 423) or condition ($F(1, 28) = 0.39$, $p = .535$, $\eta_p^2 = 0.014$, CI: −329, 175), but a significant interaction between condition and window ($F(1, 28) = 9.78$, $p = .004$, $\eta_p^2 = 0.259$). Infants looked significantly longer at the correct door (M = 960 ms, SD = 819) compared to the incorrect door (M = 332 ms, SD = 339) in the FB1 condition ($F(1, 28) = 5.15$, $p = .031$, $\eta_p^2 = 0.155$, CI: 61, 1194), but significantly longer at the incorrect door (M = 1002 ms, SD = 661) compared to the correct door (M = 444 ms, SD = 551) in the FB2 condition ($F(1, 28) = 4.64$, $p = .040$, $\eta_p^2 = 0.142$, CI: −1087, −27).

3.1.2.2. Analyses of full set. In the test trials, 51 participants provided data for the FB1 condition (50 provided a valid first fixation), and 51 for the FB2 condition (49 provided a valid first fixation). Regarding those infants who passed the second familiarization trial, 29 participants of each condition provided data for the test trials (28 provided a valid first fixation in the FB1, and 27 in the FB2).

3.1.2.2.1. First look. When analyzing all infants, in the FB1 condition, 62% directed their first look to the correct door, which is not different from chance (binomial test, n = 50, $p = .119$, OR = 1.63). In the FB2 condition, infants' first look went significantly more often to the incorrect door than expected by chance (only 24% correct; binomial test, n = 49, $p < .001$, OR = 3.17). Comparing the two conditions, infants were significantly more often correct in the FB1 condition compared to the FB2 condition (McNemar, n = 44, $p = .002$, OR = 4.82). Order of condition presentation (FB1 first, FB2 first) had no influence on the FB1 condition (Fisher's exact test, n = 50, $p = .383$, $\phi = 0.155$). In the FB2 condition, performances differed between orders (Fisher's exact test, n = 49, $p = .022$, $\phi = -0.345$). When FB2 was administered first, infants performed significantly below chance (see first trial analysis). When FB2 was second, infants' first looks were not different from chance (41% correct; binomial test, $p = .523$, OR = 1.44).

The left panel of Fig. 3 shows the ratio of first looks in each condition for the infants who had correctly anticipated in the second familiarization trial. When analyzing those infants who passed the second familiarization trial, in the FB1 condition, infants' first look went significantly more often to the correct door than expected by chance (79% correct; binomial test, n = 28, $p = .004$, OR = 3.76);

in the FB2 condition, infants' first look went almost significantly more often to the incorrect door than expected by chance (30% correct; binomial test, n = 27, p = .052, OR = 2.33). Comparing the two conditions, infants were significantly more often correct in the FB1 condition compared to the FB2 condition (McNemar, n = 24, p = .004, OR = 9.27).

3.1.2.2.2. DLS. When analyzing the DLS of all infants, in the FB1 condition, infants looked significantly longer to the correct door than expected by chance (M = 0.245, SD = 0.747; $t(50) = 2.35$, p = .023, d_z = 0.34, CI: 0.035, 0.455); in the FB2 condition, infants looked significantly longer to the incorrect door than expected by chance (M = −0.373, SD = 0.669; $t(50) = 3.98$, $p < .001$, d_z = 0.56, CI: −0.561, −0.185). Conditions differed significantly ($t(45) = 3.74$, p = .001, d_z = 0.55, CI: 0.275, 0.917). Order of condition presentation had no influence in the FB1 condition ($t(49) = 1.58$, p = .121, d_s = 0.45, CI: −0.741, 0.089). In the FB2 condition, performances tended to differ between orders ($t(49) = 1.74$, p = .088, d_s = 0.49, CI: −0.049, 0.690). When FB2 was administered first, infants performed significantly below chance (see first trial analysis). When FB2 was second, the DLS was not different from chance (M = −0.203, SD = 0.707; $t(23) = 1.41$, p = .173, d_z = 0.28, CI: −0.502, 0.095).

When analyzing the DLS of those infants who had passed the second familiarization trial, in the FB1 condition, infants looked significantly longer to the correct door than expected by chance (M = 0.486, SD = 0.704; $t(28) = 3.72$, p = .001, d_z = 0.70, CI: 0.219, 0.754). In the FB2 condition, infants looked significantly longer to the incorrect door than expected by chance (M = −0.331, SD = 0.679; $t(28) = 2.63$, p = .014, d_z = 0.50, CI: −0.590, −0.073). Conditions differed significantly ($t(25) = 3.66$, p = .001, d_z = 0.72, CI: 0.354, 1.262).

3.2. Anticipation + Outcome task

3.2.1. Looking times in outcome phase

3.2.1.1. First trial analyses. In a first trial analysis, 26 participants provided data in the FB condition (16 congruent, 10 incongruent) and 26 in the TB condition (14 congruent, 12 incongruent). The right panel of Fig. 4 shows the results of the between-subject analysis on the looking time during the outcome phase for the first trial. A univariate ANOVA with condition (FB, TB) and congruency (congruent outcome, incongruent outcome) as between-subject factors provided confirmatory support to previous studies that infants in the group with the incongruent outcome tended to look longer compared to infants in the group with the congruent outcome (main effect of congruency: $F(1, 48) = 2.97$, p = .046, one-tailed, η_p^2 = 0.058, CI: −6485, 501), with no significant difference between conditions and no interaction. To test directly the prediction that the effects were present in each condition, simple comparisons based on the variance of the overall ANOVA revealed that infants in the TB condition looked significantly longer when the outcome in the first trial was incongruent compared to when it was congruent ($F(1, 48) = 3.97$, p = .026, one-tailed, η_p^2 = 0.076, CI: −9712, 46), but not in the FB condition ($F(1, 48) = 0.21$, p = .323, one-tailed, η_p^2 = 0.004, CI: −6150, 3849).

3.2.1.2. Analyses of full set. 52 infants provided data for both congruent and incongruent trials, 26 per condition. For a within-subject analyses on the outcome phase, a 2 × 2 × 2 ANOVA on the mean looking time with congruency as within-subject factor (congruent outcome, incongruent outcome) and condition (FB, TB) and order (congruent first, incongruent first) as between-subject factors revealed a significant main effect of congruency ($F(1, 48) = 5.52$, p = .023, η_p^2 = 0.103, CI: −3484, −270) and an interaction with order ($F(1, 48) = 8.37$, p = .006, η_p^2 = 0.149). Only infants who first saw the incongruent outcome looked significantly longer at the incongruent than the congruent outcome ($F(1, 48) = 11.89$, p = .001, η_p^2 = 0.199, CI: −6633, −1746); infants who first saw the congruent outcome did not look significantly longer at the incongruent than the congruent outcome ($F(1, 48) = 1.76$, p = .677, η_p^2 = 0.004, CI: −1653, 2523). The left panel of Fig. 4 shows the mean looking times of infants who saw the incongruent trial first for the congruent and incongruent outcome for both conditions.

To ensure that this effect was present in both conditions, we re-ran the 2 × 2 ANOVA for each condition. We obtained again significant interactions between congruency and order for both conditions, and simple comparisons based on the overall variance of the ANOVA revealed that in both conditions infants looked longer to the incongruent than congruent outcome when the incongruent

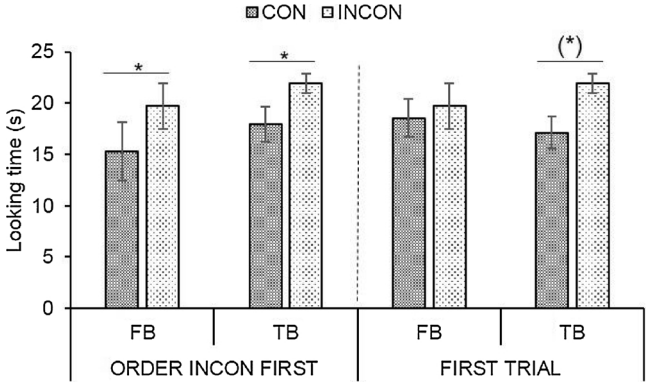

Fig. 4. Anticipation + Outcome task. Participants' mean looking time ± s.e.m. for congruent (CON) and incongruent trials (INCON) in the false belief (FB) and true belief (TB) conditions. Left panel: Within-subject analysis for the order incongruent outcome first (n = 22). Right panel: Between-subject analysis for the first trial (n = 26 per condition). *$p < .05$; (*)p = .026, one-tailed.

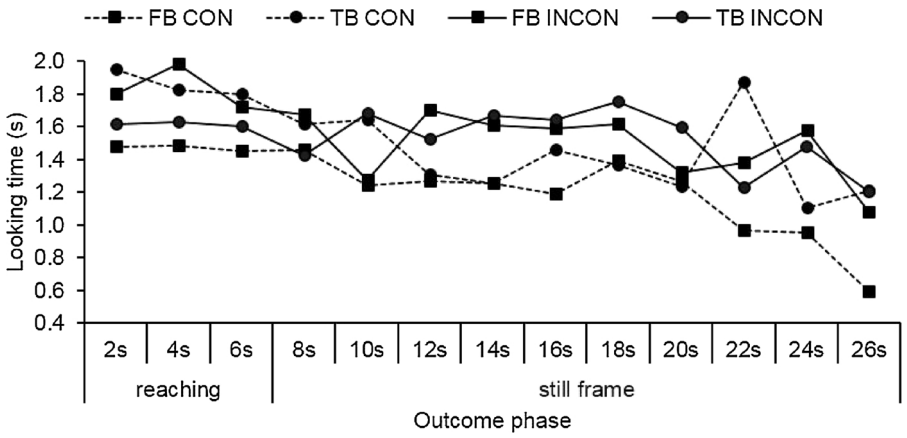

Fig. 5. Anticipation + Outcome task. Participants' mean looking time in two-second-increments in congruent and incongruent trials across the outcome phase for each condition in the incongruent outcome first order. First second of reaching phase is not depicted for better illustration.

trial had been first (FB: $F(1, 24) = 5.22$, $p = .031$, $\eta_p^2 = 0.179$, CI: −8387, −427; TB: $F(1, 24) = 6.95$, $p = .014$, $\eta_p^2 = 0.225$, CI: −7084, −864), but not when the congruent trial had been first (FB: $F(1, 24) = 0.16$, $p = .690$, $\eta_p^2 = 0.007$, CI: −2532, 3761; TB: $F(1, 24) = 0.03$, $p = .856$, $\eta_p^2 = 0.001$, CI: −2623, 3135). Fig. 5 displays the looking times across the time of the test period for the incongruent first order. The effect appears to emerge most pronounced in the still frame period after 4 s and lasts for around 8–10 s.

A similar pattern was obtained when analyzing the number of infants who showed a differential looking pattern across the two trials: 73% of infants who saw the incongruent outcome first looked longer at the incongruent outcome compared to the congruent outcome, which is marginally significant (binomial test, n = 22, $p = .052$, OR = 2.70). In contrast, only 63% of infants who saw the congruent outcome first looked longer at the incongruent outcome compared to the congruent outcome (binomial test, n = 30, $p = .20$, OR = 1.70).

3.2.2. Pupil dilation

3.2.2.1. Outcome phase

3.2.2.1.1. First trial analyses. We analyzed the RCD from the baseline defined as the mean of the last second of the anticipation phase to the mean of the outcome phase. In a first trial analysis, 25 participants provided data in the FB condition (15 congruent, 10 incongruent) and 26 in the TB condition (14 congruent, 12 incongruent). A between-subject univariate ANOVA with congruency (congruent outcome, incongruent outcome) and condition (FB, TB) as factors revealed that infants in the group with the incongruent outcome had a significantly larger increase in pupil size compared to infants in the group with the congruent outcome (main effect of congruency: $F(1, 47) = 10.54$, $p = .002$, $\eta_p^2 = 0.183$, CI: −0.083, −0.020), with no significant difference between conditions and no interaction. To test directly whether the effects were present in each condition, simple comparisons based on the variance of the overall ANOVA confirmed that in the FB condition infants' relative pupil size increase was larger when the outcome in the first trial was incongruent (M = 0.112, SD = 0.065) compared to when it was congruent (M = 0.043, SD = 0.052; $F(1, 47) = 9.10$, $p = .004$, $\eta_p^2 = 0.162$, CI: −0.115, −0.023); in the TB condition means were in the same direction but did not reach significance (incongruent: M = 0.091, SD = 0.069; congruent: M = 0.057, SD = 0.038; $F(1, 47) = 2.40$, $p = .128$, $\eta_p^2 = 0.049$, CI: −0.078, 0.010).

3.2.2.1.2. Analyses of full set. 50 participants provided data for both congruent and incongruent trials, 25 per condition. A 2 × 2 × 2 ANOVA for the RCD (from last second of anticipation phase to mean of outcome phase) with congruency (congruent outcome, incongruent outcome) as within-subject factor, and condition (FB, TB) and order (congruent first, incongruent first) as between-subject factors revealed a main effect for congruency ($F(1, 46) = 8.20$, $p = .006$, $\eta_p^2 = 0.151$, CI: −0.044, −0.008) which interacted with order ($F(1, 46) = 4.87$, $p = .032$, $\eta_p^2 = 0.096$), with no difference between the conditions. Simple comparisons based on the variance of the overall ANOVA revealed that infants' relative pupil increase was larger in the incongruent (M = 0.101, SD = 0.067) compared to the congruent outcome (M = 0.056, SD = 0.057) when the incongruent trial was presented first ($F(1, 46) = 11.46$, $p = .001$, $\eta_p^2 = 0.199$, CI: −0.072, −0.018), but not when the congruent trial was presented first (incongruent: M = 0.056, SD = 0.062; congruent: M = 0.050, SD = 0.046; $F(1, 46) = 0.246$, $p = .622$, $\eta_p^2 = 0.005$, CI: −0.030, 0.018). To ensure this effect was present in both conditions, we analyzed each condition separately. For the FB condition, when the incongruent trial was presented first, the difference in the RCD between the incongruent (M = 0.112, SD = 0.065) and the congruent (M = 0.051, SD = 0.062) outcome remained significant ($F(1, 23) = 7.75$, $p = .011$, $\eta_p^2 = 0.252$, CI: −0.105, −0.016). Also for the TB condition, when the incongruent trial was presented first, we found a similar pattern (incongruent: M = 0.091, SD = 0.069; congruent: M = 0.061, SD = 0.055; $F(1, 23) = 3.53$, $p = .074$, $\eta_p^2 = 0.133$, CI: −0.064, 0.003). On the level of individual infants the measure did not differentiate between infants who showed a larger RCD in incongruent compared to congruent trials. Fig. 6 shows the temporal unfolding of the pupil sizes during the outcome period for congruent and incongruent trials in the TB and FB conditions when the incongruent trial was administered first. Differences seem to arise as early as 3 s into the reaching phase and last almost until the end of the testing period.

We also tested whether the effect was immediately present following the outcome. If the effect was present early, this would render interpretations that invoke longer looking as one cause of the effect less likely. We focused on the reaching phase when the

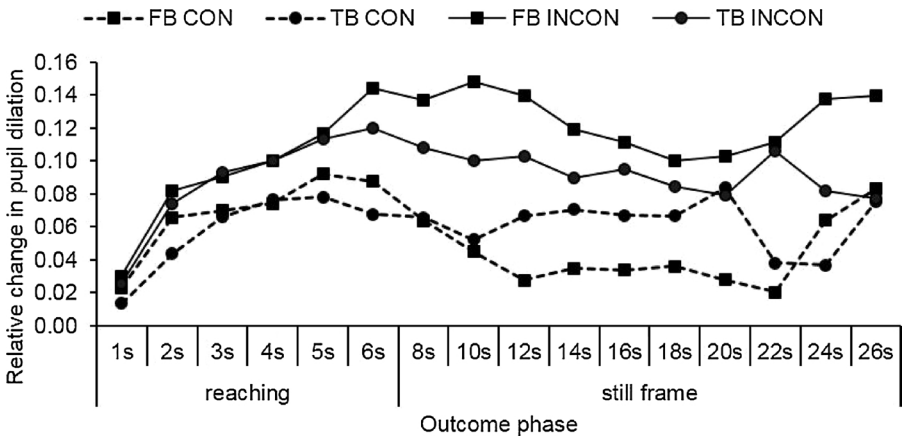

Fig. 6. Mean relative change in pupil dilation for participants in the order incongruent outcome first in congruent and incongruent trials across the outcome phase for each condition. First five time points of outcome phase in one-second-increments, all subsequent time points in two-second-increments.

hand started approaching the container (fourth second of outcome phase, see Fig. 6) and re-ran the analyses. The pattern of results with order interactions and selective congruency effects remained the same. Full results are reported in Appendix B.

3.2.2.2. Belief induction phase. We also analyzed the mean RCD over the two trials from a pre-induction baseline (the mean of the five seconds before the teddy disappeared and the phone rang) to after the belief induction (the first second after the agent reappeared and sat behind the screen again). 25 participants per condition provided data. A univariate ANOVA with condition (FB, TB) and order (congruent first, incongruent first) as between subject factors revealed a main effect of condition ($F(1, 46) = 18.81$, $p < .001$, $\eta_p^2 = 0.290$, CI: 0.033, 0.091) with no effect of order and no interaction. Infants' RCD was significantly larger in the false belief condition (M = 0.055, SD = 0.044) compared to the TB condition (M = −0.006, SD = 0.055). To ensure that there were no learning or other internal effect involved, we analyzed the first trial. The ANOVA yielded again a main effect of condition ($F(1, 41) = 15.59$, $p < .001$, $\eta_p^2 = 0.267$, CI: 0.036, 0.111). In the first trial, infants had a larger pupil size increase in the FB condition (M = 0.059, SD = 0.066) compared to the TB condition (M = −0.015, SD = 0.053).

To investigate whether this effect was due to the sudden appearance of the agent in the FB condition who had been present in the TB condition already for a longer time, we defined as a new time window for the TB condition the moment when the agent had just appeared and compared it to the moment in the FB condition when the agent had just appeared. A univariate ANOVA with condition and order as between subject factors for the RCD from the pre-induction baseline to the appearance of the agent (early in TB, late in FB) yielded no significant difference between the FB condition (M = 0.052, SD = 0.043) and the TB condition (M = 0.045, SD = 0.068; $F(1, 46) = 0.20$, $p = .656$, $\eta_p^2 = 0.004$, CI: −0.025, 0.039), suggesting that the effect was based on the appearance of the agent rather than the induction of a false belief.

3.2.3. Anticipatory looking

To provide additional evidence for our analyses in the AL task, we also analyzed infants' anticipatory looking pattern in the current A + O task.

3.2.3.1. Familiarization trials. 51 participants provided data for the first familiarization trial (48 provided a valid first fixation), 47 for the second familiarization trial, and 46 provided data for both trials (44 provided two valid first fixations). Across the two familiarization trials of both conditions, the first look tended to be less often to the correct door than expected by chance (M = 0.421, SD = 0.263; $t(43) = 2.01$, $p = .051$, $d_z = 0.30$, CI: −0.160, 0.000). In the first trial, 52% of the infants directed their first look to the correct door (binomial test, n = 51, $p = .885$, OR = 1.08). In the second trial, infants' first look was significantly more often directed to the incorrect door than expected by chance (32% correct; binomial test, n = 47, $p = .019$, OR = 2.13). There was a negative correlation between infants' first look in the first familiarization trial and in the second familiarization trial ($phi(44) = -0.409$, $p = .007$). A similar pattern was obtained with the DLS measure. Infants performed at chance level across the two trials (M = −0.074, SD = 0.552; $t(45) = 0.90$, $p = .371$, $d_z = 0.13$, CI: −0.237, 0.090), as well as in the first trial (M = 0.033, SD = 0.810; $t(50) = 0.29$, $p = .773$, $d_z = 0.04$, CI: −0.195, 0.261). In the second trial, infants looked significantly longer at the incorrect door than expected by chance (M = −0.224, SD = 0.744; $t(46) = 2.06$, $p = .045$, $d_z = 0.30$, CI: −0.442, −0.005). There was no correlation between the first and second familiarization trial for the DLS.

3.2.3.2. Anticipation phase

3.2.3.2.1. First trial analyses. 24 participants provided data in the FB condition, and 25 in the TB condition (24 provided a valid first fixation).

3.2.3.2.1.1. First look

In the first test trial, in the FB condition, infants' first looks were not different from chance (63% correct; binomial test, n = 24,

p = .307, OR = 1.70); also in the TB condition, infants' first looks were not different from chance (54% correct; binomial test, n = 24, p = .839, OR = 1.17). Conditions were not different from each other (Fisher's exact test, n = 48, p = .385, ϕ = −0.167). Because there were only few infants who passed the second familiarization trial, we could not use this as an inclusion criterion for further analyses.

3.2.3.2.1.2. DLS

The DLS was not different from chance in the FB condition (M = −0.188, SD = 0.602; t(23) = 1.53, p = .140, d_z = 0.31, CI: −0.442, 0.066), as well as in the TB condition (M = −0.037, SD = 0.625; t(24) = 0.30, p = .767, d_z = 0.06, CI: −0.300, 0.221). Conditions were not different from each other (t(47) = 0.86, p = .395, d_s = 0.12, CI: −0.504, 0.202)

3.2.3.2.2. Analyses on repeated trials. In contrast to the AL task which had one trial per condition, in the current A + O task, infants had two test trials per condition. In the second trial, 25 participants provided data in each condition (24 provided a valid first fixation in the TB). 23 participants provided data for both trials in the FB condition, 25 in the TB condition (24 a valid first fixation).

3.2.3.2.2.1. First look

Across the two test trials, first looks were not significantly different from chance in the FB condition (M = 0.587, SD = 0.389; t(22) = 1.07, p = .295, d_z = 0.22, CI: −0.081, 0.255), or in the TB condition (M = 0.500, SD = 0.417; t(23) = 0.00, p = 1.00, d_z = 0.00, CI: −0.176, 0.176). Conditions were not different from each other (t(45) = 0.74, p = .464, d_s = 0.22, CI: −0.150, 0.324). This was also true for the second anticipation trial (FB: 56% correct; binomial test, n = 25, p = .690, OR = 1.27; TB: 54% correct; binomial test, n = 24, p = .839, OR = 1.17; Fisher's exact test, n = 49, p = 1.00, ϕ = −0.018). First looks in the second trial were not different between infants that had first seen a congruent or an incongruent trial (Fisher's exact test, n = 49, p = 1.0, ϕ = −0.036).

3.2.3.2.2.2. DLS

Across the two test trials, in both conditions the DLS was not significantly different from chance level (FB: M = −0.013, SD = 0.484; t(22) = 0.13, p = .896, d_z = 0.03, CI: −0.223, 0.196; TB: M = 0.090, SD = 0.454; t(24) = 0.99, p = .334, d_z = 0.20, CI: −0.098, 0.277), with no difference between conditions (t(46) = 0.76, p = .451, d_s = 0.22, CI: −0.375, 0.170). In the second test trial, in the FB condition, the DLS was not different from chance (M = 0.223, SD = 0.734; t(24) = 1.52, p = .141, d_z = 0.31, CI: −0.080, 0.526). However, in the TB condition, in the second test trial, infants tended to look longer to the correct door (M = 0.217, SD = 0.624; t(24) = 1.74, p = .095, d_z = 0.36, CI: −0.041, 0.474). Conditions were not different from each other (t(48) = 0.04, p = .972, d_s = 0.01, CI: −0.381, 0.394). There was no difference in the performance of infants in the second trial who had first seen a congruent or an incongruent trial (t(48) = 0.79, p = .938, d_s = 0.23, CI: −0.377, 0.407). Infants performed significantly better in the second trial compared to the first trial (t(47) = 2.79, p = .008, d_z = 0.40, CI: −0.601, −0.098).

3.3. Interaction task

3.3.1. First trial analyses

To replicate the result by Southgate et al. (2010) we conducted a between-subject analyses and analyzed the first trial. Fig. 7 shows the ratio of choice in the first trial for both conditions. In the TB condition, infants chose the referred box significantly more often than expected by chance (71% correct; binomial test, n = 28, p = .036, OR = 2.45). In the FB condition, infants' choice did not differ from chance, 63% incorrectly chose the referred box (binomial test, n = 32, p = .215, OR = 1.70). In contrast to the original study, there was no significant difference in the number of infants who chose the referred box between the TB and the FB condition (Fisher's exact test, n = 60, p = .293, one-tailed, ϕ = 0.094).

3.3.2. Analyses on repeated trials

48 participants took all three trials, eight took two trials, and four took one trial. When analyzing the mean performance across the repeated trials (including all participants), in the TB condition, infants chose the referred box significantly more often than

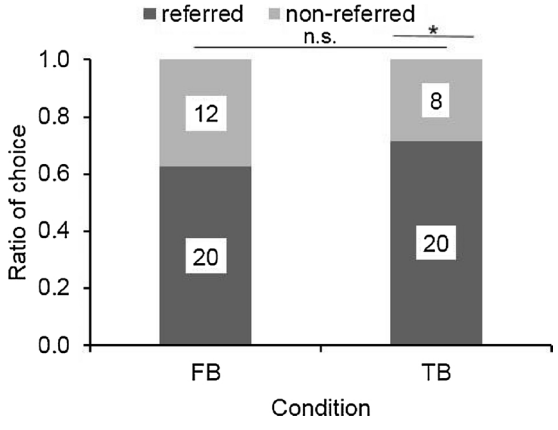

Fig. 7. Interaction task. Percentage of infants who chose each box in the first trial in both conditions. Numbers in bars show number of infants. *p < .05.

Table 1
Overview of correlations between the different measures: DLS for anticipation (1) in the FB1 condition and (2) in the second trial of the A + O task, (3) looking time difference of incongruent minus congruent trials in the outcome phase, (4) the relative change in pupil dilation in incongruent trials, and (5) the mean performance over the repeated interaction trials.

Measures	1	2	3	4	5
1. Anticipation FB1	–				
2. Anticipation A + O	0.036 n = 21	–			
3. Looking time	−0.251 n = 21	**TB: 0.413**[*] n = 25 FB: 0.009 n = 25	–		
4. Pupil dilation	0.105 n = 21	0.027 n = 50	**0.711**[**,a] n = 22	–	
5. Interaction	0.175 n = 24	−0.272 n = 44	0.004 n = 46	**TB: 0.677**[**] n = 22 FB: −0.039 n = 23	–

[*] $p < .05$.
[**] $p < .001$.
[a] For the mean pupil size, not the RCD.

expected by chance (M = 0.71, SD = 0.33; $t(27) = 3.41$, $p = .002$, $d_z = 0.66$, CI: 0.085, 0.343). However, in the FB condition, infants also chose the referred box significantly more often than expected by chance (M = 0.64, SD = 0.35; $t(31) = 2.19$, $p = .036$, $d_z = 0.39$, CI: 0.009, 0.262). Across the test trials, there was no significant difference between the conditions in infants' choice of the referred box ($t(58) = 0.89$, $p = .377$, $d_s = 0.23$, CI: −0.256, 0.098).

3.4. Correlations between measures

We correlated performance in our key measures of anticipatory looking, looking time, pupil dilation, and interaction with each other. As evident from the group results, there were various ways of composing the variables for each measure (e.g. DLS, first look, number of infants, performance in a given trial or order). We report the most meaningful relations between variables for this type of exploratory analyses. No other correlations were significant. Table 1 provides a summary of the main results. Sample sizes of tests vary because only infants who provided data for both measures that were compared could be considered.

3.4.1. Anticipatory looking

In the Anticipatory Looking task, the first look and the DLS of the FB1 condition did not correlate with the anticipatory looking measures of the A + O task in the FB condition; did not correlate with any of the FB looking time measures of the A + O task during the outcome phase; did not correlate with any of the FB pupil size measures of the outcome phase; and did not correlate with any of the interaction measures in the FB condition. Also, none of the measures correlated with the FB2 condition. This was also true when only considering those infants that passed the second familiarization trial of the AL task.

For the anticipatory looking in the A + O task, the first look measure revealed no correlations. The mean DLS across both trials correlated with the looking time difference between incongruent and congruent outcomes. This correlation was only present in the TB condition ($r(25) = 0.506$, $p = .010$), and absent in the FB condition. In the TB condition, it was present only in the second trial (in which the group performed better; $r(25) = 0.413$, $p = .040$), but not in the first trial. In the second trial of the TB condition, this correlation emerged also on the level of individual infants ($phi(25) = 0.345$, $p = .085$). There were no relations to the Interaction task.

3.4.2. Looking time

In addition to the reported correlation with the anticipatory looking in the A + O task, the looking time difference between incongruent and congruent outcomes correlated with the mean pupil size during the incongruent outcome ($r(52) = 0.382$, $p = .005$). The correlation was also found when correlating the looking time difference with the pupil size in the fourth second ($r(50) = 0.319$, $p = .024$; see our additional analyses, Appendix B). Because infants who saw the incongruent trial first seemed to perform better than the other infants, we ran the correlation for only that group of infants. This yielded an even stronger correlation for the looking time difference and the mean pupil size in incongruent trials ($r(22) = 0.711$, $p < .001$). This correlation was also significant for each condition separately (FB: $r(10) = 0.749$, $p = .013$; TB: $r(12) = 0.700$, $p = .011$) and with the pupil size in the fourth second of the reaching phase ($r(22) = 0.504$, $p = .017$). None of the other looking time measures yielded any correlations.

3.4.3. Pupil dilation

In addition to the reported correlation with looking times, there was a correlation between the RCD in the incongruent trial and the mean number of correct interaction trials, but only in the TB condition ($r(22) = 0.677$, $p = .001$). This correlation held when only analyzing the first interaction trial of the TB condition ($r(22) = 0.578$, $p = .005$).

4. Discussion

The rationale of the present study was to test for the robustness and replicability of implicit ToM tasks used in infants and toddlers, as well as for their convergent validity. To this end, we collected four implicit measures across three different infant ToM

tasks, testing visual anticipation of belief-congruent actions, visual responses to violations of expected belief-congruent actions, and interactive reactions to belief-congruent requests.

4.1. Reliability of implicit ToM measures

Regarding replicability, the study revealed a mixed pattern of findings. The fact that we could replicate the pattern of findings in the true belief conditions and the FB1 condition of the AL task suggests that the measures were in principle sensitive to infants' processing of the situation and that there were no general problems with the task setups. However, none of the measures provided strong and conclusive evidence for false belief processing and, in a strict sense, failed to replicate previous findings. At the same time, our additional analyses on repeated trials, order effects and our new measure of pupil dilation indicate some degree of false belief processing under certain conditions. These indications pertained to the looking time and pupil size measures, but were absent for the anticipatory looking and interaction measures.

Regarding the looking time measure, the first trial analysis of our new VoE task which we conducted analogously to single trial analyses of previous studies (e.g., Onishi & Baillargeon, 2005), yielded only a weak effect in the TB condition, and no effect in the FB condition, thus failing to provide robust evidence for belief-based processing. This might call into question the robustness and size of single trial effects, and should be taken into consideration when designing future studies. Because we did a conceptual replication of the VoE paradigm, we do not interpret our findings as a failure of a direct replication. Note also that the original studies coded looking times live from video while we relied on automated eye-tracking recordings. Our analyses on tracking ratios, however, revealed very good tracking and no firm grounds to question its reliability (see participants section and Appendix B). In light of the findings of the diverse and procedurally very dissimilar VoE studies in this field one could have expected better performance (e.g., Scott et al., 2012; Surian et al., 2007; Träuble et al., 2010). However, when looking at our full-factorial analyses, the looking time measure did reveal sensitivity to belief-congruent action outcomes for true and false beliefs – but only if infants watched the incongruent outcome first. This kind of order effect is common in visual habituation research when congruent and incongruent outcomes alternate in a test phase following habituation (e.g., Baillargeon, 1987; Baillargeon, Spelke, & Wasserman, 1985). Given that our participants were already 2-years-old, it is plausible that they kept track of what they had seen in the first trial. The effect was strong enough to remain when measured at the level of individual children. The underlying mechanisms of this measure are unclear, but our descriptive findings on the temporal unfolding suggest that in the case of belief-congruent outcomes, infants begin to divert their attention fairly early, already after the first few seconds of a still frame, while in the case of belief-incongruent outcomes they keep looking for a long time. One possible interpretation is that infants wait for a congruent closure and expect a further step in the sequence and hence keep watching. If this was the case, this effect should reduce across repeated trials with incongruent outcomes.

Regarding the pupil size measure, the prediction was that a belief-incongruent outcome would yield heightened attention compared to an expected outcome, resulting in a larger increase in pupil size. Indeed, our measure of relative pupil size increase revealed sensitivity to belief-congruent action outcomes for true and false beliefs, and, like the looking duration measure, this was only the case for infants who had first watched an incongruent outcome. While the measure was not sensitive enough to reveal the effect on the level of individual children, the effect remained in a first trial analysis. Given that this is the first report of a belief-congruency processing effect in pupil size change we need to be cautious with its interpretation. The temporal dynamics of the effect show that it emerged early (after 2–3 s into the unfolding of the reaching event) and lasted astonishingly long. The similarity to the looking time measure suggests similar attentional processes underlying 2-year-olds looking time and pupil size change to violations of expected action outcomes. This is in line with recent findings showing that both pupil dilation and looking times increased in response to stimuli violations, for example in object permanence tasks (Jackson & Sirois, 2009) and in face processing tasks (Falck-Ytter, 2008; Gredebäck, Eriksson, Schmitow, Laeng, & Stenberg, 2012).

Because the A + O task was always administered last, it could be that fatigue or other carry-over effects influenced performance. Previous VoE studies, however, have used multiple trial procedures, long sessions, within-designs or even included several FB tasks, and found no effects of trial or task order (Poulin-Dubois & Yott, 2017; Träuble et al., 2010; Yott & Poulin-Dubois, 2016). At any rate, fatigue should not be selective to a specific condition but affect processing on a general level. However, this is not what we found. Our study revealed selective significant processing differences in the TB condition and under specific circumstances (order) in the FB condition. Note also that the rate of tracking loss did not differ drastically between first and last task, suggesting that infants' visual attention was not influenced by potential fatigue effects. It remains possible, of course, that for currently unknown reasons false belief processing is demonstrable only in single task studies using single trials.

Regarding infants' visual anticipations, one interpretation of the pattern of results across tasks is that anticipatory looking reveals neither belief tracking nor goal anticipation but rather experience-dependent anticipations which may change within a task (Paulus et al., 2011). Findings from the familiarization trials question the suitability of visual saccades in revealing action anticipation: Although 2-year-olds have a robust understanding of reaching as goal-directed act, they anticipated in the AL task correctly only after three trials (two warm-up and one familiarization), and were still far from ceiling. The negative correlations between the first and second trials of the familiarization phases indicate some form of perseveration such that infants tended to look where they had last seen the hand appear, suggesting extraneous task demands.

Findings from the false belief conditions strongly speak against an understanding of false beliefs. The replication of the Southgate et al. (2007) results failed for the FB2 condition, which is the crucial condition for crediting participants with false belief processing. To succeed in FB1 infants just had to look at where the ball had been last. This same strategy led to failure in FB2 where the last location of the ball was the belief-incongruent location. Importantly, we did not just fail to replicate the conventional *p*-value, but the patterns in the FB2 condition were in the opposite direction of the original study, and they were significantly different from the FB1

condition, rejecting the hypothesis that common FB processing underlies these two FB conditions. Our A + O task provided converging negative evidence. Although the task may perhaps exert higher demands because the agent disappears (Rubio-Fernández & Geurts, 2013) and the object remains present (Wang & Leslie, 2016), it is important to note that the task matched the verbal standard false belief task conceptually most closely.

The interaction-based measure of belief-congruent reacting, finally, did not reveal clear-cut false-belief understanding either. The replication of results by Southgate et al. (2010) failed because infants performed at chance in the FB condition (and across trials even in the opposite direction). Although infants were above chance in the TB condition, one could have expected better performance given the findings by Southgate et al. and the generally fairly easy structure of the task for 2-year-olds. Because we kept the element of deception in the TB condition (up to the point when E1 came back and watched) for better comparison with the FB condition, it is theoretically possible that this cue detracted from E1's actual epistemic state. However, even if the cue perhaps underestimated infants' above chance performance in the TB condition, the problem is that infants clearly failed in the FB condition. Why infants performed poor in the FB condition remains to be explained. Informal observations revealed that infants often made a second offer (not reported by Southgate et al.), which perhaps indicates that infants did not fully understand the referential specificity of the request.

It is at least theoretically also possible that our participants failed because they were too old for this task, since the original study tested 17-month-olds. However, this seems very unlikely, because other recent replication studies have also failed to replicate interaction-based tasks with 18-month-old infants (Poulin-Dubois & Yott, 2017), and a study using a modified version of the Southgate et al. (2010) task has recently reported positive evidence at 36 months (Király et al., 2016). It could also be that the preceding AL task influenced performance in the Interaction task. However, this too appears unlikely, because all infants engaged in the warm-up trials and enjoyed interacting with the experimenter across all trials. Further, any task order effect should not be selective but influence performance in true and false belief conditions equally. Again, this is not what we found.

4.2. Convergent validity of implicit ToM measures

Children's verbal ToM at 4 years of age is characterized by a broad and systematic competence that emerges in synchronized and correlated fashion across various superficially dissimilar tasks and measures (Astington & Gopnik, 1988; Perner & Roessler, 2012). Our correlation analyses did not reveal a comparable level of broad ToM competence for infants. Correlations were mostly absent and scattered. This is in line with other recent studies that have failed to find unified performances between and within different FB paradigms (Poulin-Dubois & Yott, 2017; Yott & Poulin-Dubois, 2016). Bearing in mind our exploratory approach, two main findings emerged. First, the few correlations were mostly selective to the true belief condition. This indicates that the tasks did not measure totally different aspects, or suffered from very different extraneous task demands. It also underscores the interpretation that infants do not have a unified concept of belief. The TB correlation between the looking time difference and the DLS of the second A + O anticipation trial indicates convergent validity and supports the interpretation that by the second test trial infants had learnt to correctly anticipate (but only in the TB condition). The TB correlation between the pupil size increase in incongruent trials and the Interaction task is interesting as it may reflect as common denominator between the two variables a concern for others (Hepach, Vaish, & Tomasello, 2012). Again, however, the correlations were selective to the TB condition and hence do not indicate an understanding of false beliefs.

The second finding was the relation between the looking time and the pupil size. It is the only correlation that holds for both true and false beliefs, which provides convergent validity to the positive group level results of false belief understanding in these two measures. Confirmatory results are required, especially because the correlation concerned the absolute tonic size of the pupil, not the relative increase of the pupil. However, it is unlikely that the finding reflects only a peripheral physiological correlate of the looking time pattern (longer looking leads to larger pupil), because the correlation was already evident when calculating with the fourth second of the reaching phase. Our additional control analyses (see Appendix B) further speak against measurement artifacts.

The current study revealed mixed findings depending on measures and analyses. The two attentional measures, looking time and pupil change, were most promising in revealing false belief processing. This could be because they were least demanding and rather unconscious, at least in the case of pupil change (which cannot be produced consciously). Anticipatory looking and interactive behaviors are more goal-driven then looking time and pupil change. They entail anticipating specific consequences of the behavior (e.g., to see an event in a specific location; to satisfy a requester's need), while looking time and pupil change rather retrodict than project events. Anticipatory looking and interactive behaviors entail elements of choice (a decision where to look, or what to offer), which perhaps requires more reasoning processes than for the other two measures. Therefore, and because of their anticipatory direction, anticipatory looking and interactive behaviors depend on pragmatically very clear situations. AL, and especially interaction tasks, however, often suffer from weak pragmatic soundness. For example, in the video-based tasks, the agent was only weakly introduced to reach reliably correctly – a necessity to anticipate a correct reach; and in the Interaction task, the requester ultimately may have wanted both toys. Further, it was not conveyed why he did not retrieve the toys himself. Because the pragmatics of these non-verbal tasks are often difficult to convey, these measures may be more difficult to replicate. It remains debatable, however, how to best account for the level of ToM understanding exhibited in the looking time and pupil change measures given that there were no comparable indications of such an understanding in the other tasks.

5. Conclusion

From a theoretical point of view, the pattern of findings in the present study challenges strong claims about infant ToM, in particular nativist and two-systems views. Despite fundamental disagreements in some respects, these two kinds of accounts converge

on the claim that different implicit tasks all measure the same basic ToM capacity (ToM proper, according to nativist accounts; a basic, efficient and automatic ToM system, according to the two-systems-view). These tasks should thus each by itself be robust and there should be convergence and correlation across them.

The present study fails to find robust evidence for either replicability or convergent validity of the implicit ToM tasks. Now, what does such absence of evidence amount to? On the one hand, the limited support from the VoE task, given that this was a conceptual rather than direct replication, with many differences between original and replication task, by itself cannot be considered evidence for the absence of robust replicability. On the other hand, however, two of the three tasks used here (AL and interaction) involved direct replication attempts of previous studies, and the clearly negative present findings can thus constitute at least prima facie evidence of absence of robust replicability of these tasks. But clearly, the present findings taken by themselves cannot settle the broader question of the robustness and replicability of implicit ToM tasks. In order to understand whether implicit ToM in infants and toddlers is a real and robust phenomenon, future research will need to design and administer systematic, large-scale, pre-registered multi-lab replication and meta-analytical studies.

In the meantime, the present findings (together with other convergent evidence reported in this Special Issue) suggest that the empirical foundation for positing early and implicit ToM competence in infants is much less robust and conclusive than previously assumed.

Acknowledgements

This study was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft), research unit “Crossing the borders: The interplay of language, cognition, and the brain in early human development” (Project: FOR 2253, Grants: LI 1989/3-1 and RA 2155/4-1). We want to thank Victoria Southgate for sharing her video material and giving advice on her tasks. We also want to thank the colleagues and students who helped in data collection and recruiting participants, especially Marianna Jartó, Wiebke Pätzold, Nicola Ballhausen, Betty Timmermann, Maria Häberlein, Maximilian Kaliski, Jula Brüning-Wessels and Jana Klose.

Appendix A. Stimuli and procedure

Anticipatory Looking task

We showed infants two warm-up trials, two familiarization trials and two false belief test trials. One test trial was in the false belief 1 (FB1) condition, the other test trial in the false belief 2 (FB2) condition. In all trials, an agent stood behind a screen that contained two doors, each in front of one box. The agent wore a visor cap hiding her eyes, while she was following the displayed actions with her head movements. In the final phase of each trial, the two doors were illuminated for 1 s which was accompanied by a chime sound. This was followed by a 1.75 s still frame (without illumination), after which the agent would reach through one of the doors. The warm-up and familiarization trials served to increase infants’ understanding of the actor’s goal and the predictability that the agent will reach through the door after the illumination and still frame phase.

In the warm-up trials (Senju et al., 2009), infants saw a red whale toy sitting on one box (one trial on each box, order counterbalanced between subjects). After the illumination and the still frame, the agent reached through the door for the toy. In the two familiarization trials, the agent watched a teddy putting a toy in a box (one trial on each side, order counterbalanced between subjects) and afterwards leaving the scene. After the illumination and the still frame, the agent reached through the door into the box that contained the toy (only in the first trial she also took it out).

In a FB1 test trial, the teddy put the toy in a box but suddenly decided to change the toy to the other box and to leave the scene (note: this was all seen by the agent). Afterwards, the agent was distracted by a phone call and turned to the back. Thus, she did not witness how the teddy reappeared, took the toy out of the box and left the scene giggling with the toy. In a FB2 test trial, the teddy would leave the scene after he put the toy in the first box. While he was gone, the agent got a phone call and turned to the back. So, contrary to the FB1 condition, in the FB2 condition the agent was distracted and did not witness how the teddy reappeared and changed the ball to the other box. Also in the FB2 condition, the teddy decided to take the toy out of the second box again and to leave the scene with the toy (unseen by the agent). After the illumination and the still frame, infants had to predict through which door the agent would reach in order to search for the toy. No outcome was shown in the test trials. As in the original study, the last object location was always the same in both conditions but we counterbalanced the side between subjects.

Anticipation + Outcome task

We showed each infant two familiarization trials and two test trials. We counterbalanced the order for target side and congruency. An agent (same agent as E1 in Interaction task) sat behind a screen that contained two doors with two boxes in front. He followed the actions of a teddy bear with his head movements. All trials started similarly, the agent noticed a ball that was already positioned centrally between the boxes and he said, “Ah.” A teddy appeared centrally from the bottom of the scene and waved to the infant and the agent, while the agent said, “Hello.” The teddy opened the lid of the first box, put the ball inside and closed the lid. In a familiarization trial, the teddy waved goodbye and left the scene centrally to the bottom. Then, the agent said, “Okay,” and ducked down behind the screen, so that he was not visible anymore. Subsequently, the two doors were illuminated for 1 s which was accompanied by a chime sound. After a delay of another 2s, a fixation cross appeared centrally between the doors for 550 ms, which served to center infants’ gaze before the outcome. Afterwards, the agent reached through the door inside the baited box (being visible

through the door). He grabbed the ball and reappeared above the screen. He held the ball in his hand, smiled and said, “Ah,” alternating his gaze between the ball and the infant.

In the test trials, after the teddy put the ball in the first box, he opened that box again, placed the ball between the boxes, opened the second box and put the ball inside. He closed the lid of the second box and afterwards the lid of the first box (the teddy always closed the lid of the new toy position first, so that the agent’s last view went to the empty box). Subsequently, the teddy waved goodbye and left the scene centrally to the bottom. A phone call sound was played, which was commented by the agent with, “Oh, telephone,” while he faced the infant. The agent stood up, turned around and disappeared centrally through a gap in between the two black walls in the background. In the true belief condition, the agent would reappear after a delay of 3 s and, thus, he would witness all subsequent events. In the false belief condition, on the contrary, the agents was still gone while the teddy reappeared, opened the baited box, placed the ball in between the boxes, opened the other box, placed the ball inside and closed the lids in the former manner. The teddy disappeared right before the agent in the false belief condition reappeared from behind the background (note: the ball was still present). The agent sat down behind the screen and from now on all events happened parallel between the two test conditions. As in a familiarization trial, the agent said, “Okay,” and ducked down behind the screen, so that he was not visible anymore. The two doors were illuminated for 1 s which was accompanied by a chime sound. After a delay of another 2 s, a fixation cross appeared centrally between the doors for 550 ms. Afterwards, infants were shown an outcome phase that consisted of a reaching phase (7 s) in which the agent reached through the door inside the baited box (belief-congruent in TB, but incongruent in FB) or the empty box (belief-congruent in FB, but incongruent in TB), followed by a still frame phase (20 s) in which infants were shown a still frame of the agent with his hand inside of the box.

Interaction task

Before the test trials, infants were presented with two warm-up trials designed to familiarize them with searching for objects in the boxes. The experimenter (E1) gave the infant two familiar objects and allowed it to explore them for about 10s. Afterwards, E1 put the objects into the two boxes and closed the lids. Then he asked the infant to bring him one of the objects by naming it. If the infant succeeded by bringing the correct toy first, the child was asked for the other object. This was repeated until the infant brought the requested object in two consecutive trials.

In the three test trials, E1 showed the infant two novel toys and allowed it to explore them for about 10 s, placed them in the two boxes and closed the lids (objects were not labelled yet). E1 told the infants that he had to go out because he forgot something but he would be back soon. E1 left the room through the door and another experimenter (E2) that was unknown to the infants entered the room from behind the curtains. E2 emphasized her deceptive plan by giggling and gesturing, “Shush.” E2 sat down between the boxes and interchanged the objects. E2 opened both boxes, placed one object in front of its box, took the other object, showed it to the infant, placed in the other box, picked up the first object, showed it to the infant, placed it in the other box, and closed both boxes simultaneously. In the TB condition, E1 would reappear as ‘early bird’ in the moment when E2 had opened the boxes and placed the first object in front of the box, but right before she would start interchanging the objects and he would conspicuously watch E2. From the moment when E1 entered the room in the TB condition, E2 stopped acting deceptive. In the FB condition, however, E1 would re-enter the room shortly after E2 hid behind the curtains again. E1 sat down between the boxes in a position from where he could not look inside and asked the infant, “Do you remember what I put in here? There is a Sefo in here. Shall we play with the Sefo? Can you give the Sefo to me?” whilst pointing to one of the two boxes. E1 opened both boxes simultaneously and faced the infant. E1 asked repeatedly for the objects until the infant began to point to or to approach one of the boxes. We counterbalanced target side, object pair and target object between the trials for this task.

Appendix B. Results

Analyses on looking time in A + O task including participants with more weighted gaze samples

We repeated our main analysis on the looking time measure but only included participants with higher amounts of gaze samples (> 50%). The interaction between congruency and order remained significant up to the point when only participants with gaze samples > 80% were included (see Table A1). For all configurations, simple comparisons revealed that infants looked significantly longer in the incongruent trial compared to the congruent trial, only if the incongruent trial was administered first (all *ps* < .05); and there was no difference between both trials when the congruent trial was first. The effect vanished at > 90% gaze samples and only 22 participants left.

Table A1
Main results of the 2 × 2 × 2 ANOVA with congruency as within subject factor and condition and order as between subject factors on the looking time during the outcome phase of the A + O task by only including participants with more weighted gaze samples (%WGS) and the resulting sample sizes (N).

%WGS/Factors	congruency	congruency*order	condition	order	*N*
> 50	$F(1, 44) = 6.1$, $p = .017$, $\eta_p^2 = 0.122$	$F(1, 44) = 7.2$, $p = .010$, $\eta_p^2 = 0.141$	n.s.	n.s.	48
> 60	$F(1, 41) = 5.1$, $p = .030$, $\eta_p^2 = 0.110$	$F(1, 41) = 7.5$, $p = .009$, $\eta_p^2 = 0.154$	n.s.	n.s.	45
> 70	n.s.	$F(1, 36) = 5.6$, $p = .023$, $\eta_p^2 = 0.135$	n.s.	n.s.	40
> 80	n.s.	$F(1, 28) = 4.6$, $p = .040$, $\eta_p^2 = 0.142$	n.s.	n.s.	32
> 90	n.s.	n.s.	n.s.	n.s.	22

Mirrored vs. un-mirrored videos in the AL task

To exclude that the mirroring of the test videos had an effect on infants' performances in the AL task, we compared performances of those infants who passed the second familiarization between mirrored and un-mirrored videos. There was no difference between performances for the first look measure in both conditions (Fisher's exact test; FB1: n = 28, p = .375; FB2: n = 27, p = 1.0) and no difference for the DLS measure (FB1: $t(27) = 1.63$, p = .114; FB2: $t(27) = 0.410$, p = .685).

Analyses on distance to the eye-tracker

To rule out that a variable distance between the eye and the eye-tracker could explain findings of pupil dilation during the outcome phase of the A + O task, we conducted a 2 × 2 × 2 ANOVA for a relative change in distance (baseline was the last second of the anticipation phase; the focal phases were the mean distance over the whole outcome phase and the distance in the fourth second of the outcome phase) with congruency as within subject factor and condition and order as between subject factors. There were no main effects for congruency (whole outcome: $F(1, 45) = 0.10$, p = .752, $\eta_p^2 = 0.002$; fourth second: $F(1, 44) = 0.02$, p = .888, $\eta_p^2 = 0.000$) or condition (whole outcome: $F(1, 45) = 1.83$, p = .182, $\eta_p^2 = 0.039$; fourth second: $F(1, 44) = 0.83$, p = .369, $\eta_p^2 = 0.018$) and no interaction between congruency and order (whole outcome: $F(1, 45) = 0.00$, p = .960, $\eta_p^2 = 0.000$; fourth second: $F(1, 44) = 0.06$, p = .812, $\eta_p^2 = 0.001$). Further, pupil size and distance to the eye-tracker were not correlated in congruent (whole outcome: $r(52) = -0.099$, p = .484; fourth second: $r(51) = -0.190$, p = .183) or incongruent trials (whole outcome: $r(52) = -0.060$, p = .673; fourth second: $r(51) = -0.034$, p = .811). According to this, a variable distance to the eye-tracker cannot explain our findings on pupil dilation.

Offset illumination analyses for the first trial of the AL task

When analyzing or measurements offset of the illumination, in the second familiarization trial, 29 of 48 (60%) of the infants directed their first look to the correct door (binomial test, p = .193, OR = 1.50). Of those, 26 provided data for the first test trial, 11 in the FB1 condition, and 15 in the FB2 condition. When analyzing the first test trial of those infants who had looked at the correct door in the second familiarization trial, in the FB1 condition 55% of the infants directed their first look to the correct door (binomial test, n = 11, p = .500, one-tailed, OR = 1.22); and their looking time to the correct door did not differ from chance (M = 0.394, SD = 0.809; $t(10) = 1.62$, p = .137, $d_z = 0.49$, CI: −0.149, 0.937). In the FB2 condition, infants first look went significantly more often to the incorrect door than expected by chance (13% correct; binomial test, n = 15, p = .007, OR = 6.69); and infants tended to look longer to the incorrect door (M = −0.368, SD = 0.736; $t(14) = 1.94$, p = .073, $d_z = 0.50$, CI: −0.776, 0.040). The two conditions differed significantly from each other on both measures (first look: Fisher's exact test, n = 26, p = .038, $\phi = -0.441$; DLS: $t(24) = 2.50$, p = .020, $d_s = 0.98$, CI: 0.133, 1.391).

We additionally analyzed the total looking time to each door during the 1.75 s after offset of the illumination for the first trial for those who passed the second familiarization trial. A 2 × 2 ANOVA with window (correct, incorrect) as within-subject factor and condition (FB1, FB2) as a between-subject factor yielded no significant main effects for window ($F(1, 24) = 0.20$, p = .331, one-tailed, $\eta_p^2 = 0.008$, CI: −251, 388) or condition ($F(1, 24) = 1.60$, p = .219, $\eta_p^2 = 0.062$, CI: −373, 90), but a significant interaction between condition and window ($F(1, 24) = 5.52$, p = .027, $\eta_p^2 = 0.187$). Infants tended to look longer at the correct door (M = 632 ms, SD = 601) compared to the incorrect door (M = 199 ms, SD = 255) in the FB1 condition ($F(1, 24) = 3.38$, p = .078, $\eta_p^2 = 0.124$, CI: −53, 918), but look longer at the incorrect door (M = 704 ms, SD = 529) compared to the correct door (M = 409 ms, SD = 458) in the FB2 condition ($F(1, 24) = 2.15$, p = .156, $\eta_p^2 = 0.082$, CI: −710, 121), though not significantly.

RCD in the fourth second of the reaching phase

To test whether the effect was immediately present at the beginning of the outcome phase, we focused on the reaching phase when the hand started approaching the container and calculated the RCD for the fourth second of the reaching phase (see Fig. 7) with the last second of the anticipation phase as baseline. Findings were the same. A 2 × 2 × 2 ANOVA with congruency as within subject factor and condition and order as between subject factors revealed a significant main effect of order ($F(1, 44) = 4.50$, p = .040, $\eta_p^2 = 0.093$, CI: −0.052, −0.001) and an interaction between congruency and order ($F(1, 44) = 3.99$, p = .052, $\eta_p^2 = 0.083$). Infants' relative pupil increase tended to be larger in the incongruent (M = 0.101, SD = 0.064) compared to the congruent outcome (M = 0.076, SD = 0.050) when the incongruent outcome was presented first ($F(1, 44) = 3.80$, p = .058, $\eta_p^2 = 0.079$, CI: −0.051, 001), but not when the congruent outcome was presented first ($F(1, 44) = 0.69$, p = .412, $\eta_p^2 = 0.015$, CI: −0.014, 034). For each condition separately, this comparison did not reach conventional significance level.

A first trial analysis provided confirmatory support: A between subject univariate ANOVA with order and condition as factors revealed that infants in the group with the incongruent outcome had a significantly larger increase in pupil size compared to infants in the group with the congruent outcome ($F(1, 46) = 4.49$, p = .039, $\eta_p^2 = 0.089$, CI: −0.067, −0.002), with no significant difference between conditions and no interaction. For each condition separately, this comparison did not reach conventional significance level.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970. http://dx.doi.org/10.1037/a0016923.

Astington, J. W., & Gopnik, A. (1988). *Knowing you've changed your mind: Children's understanding of representational change.* New York: Cambridge University Press.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition, 20*(3), 191–208. http://dx.doi.org/10.1016/0010-0277(85)90008-3.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118. http://dx.doi.org/10.1016/j.tics.2009.12.006.

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology, 67*(1), 159–186. http://dx.doi.org/10.1146/annurev-psych-010213-115033.

Baillargeon, R. (1987). Object permanence in 3½- and 4½-month-old infants. *Developmental Psychology, 23*(5), 655–664. http://dx.doi.org/10.1037/0012-1649.23.5.655.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. http://dx.doi.org/10.1177/1745691612459060.

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. http://dx.doi.org/10.1016/j.cognition.2009.05.006.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. http://dx.doi.org/10.1038/nrn3475.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language, 28*(2), 141–172. http://dx.doi.org/10.1111/mila.12014.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*(4), 377–395. http://dx.doi.org/10.1016/0885-2014(94)90012-4.

Elsner, B., Pauen, S., & Jeschonek, S. (2006). Physiological and behavioral parameters of infants' categorization: Changes in heart rate and duration of examining across trials. *Developmental Science, 9*(6), 551–556. http://dx.doi.org/10.1111/j.1467-7687.2006.00532.x.

Falck-Ytter, T. (2008). Face inversion effects in autism: A combined looking time and pupillometric study. *Autism Research, 1*(5), 297–306. http://dx.doi.org/10.1002/aur.45.

Gredebäck, G., & Melinder, A. (2011). Teleological reasoning in 4-month-old infants: Pupil dilations and contextual constraints. *Public Library of Science, 6*(10), e26487. http://dx.doi.org/10.1371/journal.pone.0026487.

Gredebäck, G., Eriksson, M., Schmitow, C., Laeng, B., & Stenberg, G. (2012). Individual differences in face processing: Infants' scanning patterns and pupil dilations are influenced by the distribution of parental leave. *Infancy, 17*(1), 79–101. http://dx.doi.org/10.1111/j.1532-7078.2011.00091.x.

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. http://dx.doi.org/10.1111/desc.12445.

Hamilton, A. F. D. C., Brindley, R., & Frith, U. (2009). Visual perspective taking impairment in children with autistic spectrum disorder. *Cognition, 113*(1), 37–44. http://dx.doi.org/10.1016/j.cognition.2009.07.007.

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. http://dx.doi.org/10.1111/j.1467-7687.2010.00980.x.

Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *Journal of Cognition and Development, 17*(3), 359–377. http://dx.doi.org/10.1080/15248372.2015.1135801.

Hepach, R., Vaish, A., & Tomasello, M. (2012). Young children are intrinsically motivated to see others helped. *Psychological Science, 23*(9), 967–972. http://dx.doi.org/10.1177/0956797612440571.

Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on Psychological Science, 9*(2), 131–143. http://dx.doi.org/10.1177/1745691613518076.

Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science, 12*(4), 670–679. http://dx.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. http://dx.doi.org/10.1177/0956797615627426.

Király, I., Oláh, K., Kovács, Á., & Csibra, G. (2016). Do 18- and 36-month-old infants update attributed beliefs by re-evaluating past events? *Poster session presented at the Budapest CEU conference on cognitive development*.

Knudsen, B., & Liszkowski, U. (2012a). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy, 17*(6), 672–691. http://dx.doi.org/10.1111/j.1532-7078.2011.00105.x.

Knudsen, B., & Liszkowski, U. (2012b). Eighteen- and 24-month-old infants correct others in anticipation of action mistakes. *Developmental Science, 15*(1), 113–122. http://dx.doi.org/10.1111/j.1467-7687.2011.01098.x.

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. http://dx.doi.org/10.1126/science.1190792.

Kovács, A. M., Téglás, E., & Endress, A. D. (2016). *Automatic belief tracking effects cannot be explained by attention check timing: Reply to Phillips et al.* [Unpublished Manuscript].

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences, 9*(10), 459–462. http://dx.doi.org/10.1016/j.tics.2005.08.006.

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives, 10*(3), 184–189. http://dx.doi.org/10.1111/cdep.12183.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542. http://dx.doi.org/10.1177/1745691612460688.

Moll, H., Kane, S., & Mcgowan, L. (2016). Three-year-olds express suspense when an agent approaches a scene with a false belief. *Developmental Science, 19*(2), 208–220. http://dx.doi.org/10.1111/desc.12310.

Moll, H., Khalulyan, A., & Moffett, L. (2017). 2.5-year-olds express suspense when others approach reality with false expectations. *Child Development, 88*(1), 114–122. http://dx.doi.org/10.1111/cdev.12581.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. http://dx.doi.org/10.1126/science.aac4716.

Paulus, M., Hunnius, S., van Wijngaarden, C., Vrins, S., van Rooij, I., & Bekkering, H. (2011). The role of frequency information and teleological reasoning in infants' and adults' action prediction. *Developmental Psychology, 47*(4), 976–983. http://dx.doi.org/10.1037/a0023785.

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences, 16*(10), 519–525. http://dx.doi.org/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: Reconsidering Kovács, Téglás, and Endress (2010). *Psychological Science, 26*(9), 1353–1367. http://dx.doi.org/10.1177/0956797614558717.

Poulin-Dubois, D., & Yott, J. (2017). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science,* e12600. http://dx.doi.org/10.1111/desc.12600.

Rakoczy, H., Bergfeld, D., Schwarz, I., & Fizke, E. (2015). Explicit Theory of Mind is even more unified than previously assumed: Belief ascription and understanding aspectuality emerge together in development. *Child Development, 86*(2), 486–502. http://dx.doi.org/10.1111/cdev.12311.

Rhodes, M., & Brandone, A. C. (2014). Three-year-olds' theories of mind in actions and words. *Frontiers in Psychology, 5*(263), 1–8. http://dx.doi.org/10.3389/fpsyg.2014.00263.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33. http://dx.doi.org/10.1177/0956797612447819.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Scott, S. E. B. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people Ssee. *Journal of Experimental Psychology, 36*(5), 1255–1266. http://dx.doi.org/10.1037/a0018729.

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, 141*(3), 433–438. http://dx.doi.org/10.1037/a0025458.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition, 162*, 27–31. http://dx.doi.org/10.1016/j.cognition.2017.01.018.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249. http://dx.doi.org/10.1016/j.tics.2017.01.012.

Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Developmental Science, 15*(2), 181–193. http://dx.doi.org/10.1111/j.1467-7687.2011.01103.x.

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

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous Theory of Mind in Asperger syndrome. *Science, 325*(5942), 883–885. http://dx.doi.org/10.1126/science.1176170.

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*(2), 353–360. http://dx.doi.org/10.1017/S0954579410000106.

Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science, 22*(7), 878–880. http://dx.doi.org/10.1177/0956797611411584.

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. http://dx.doi.org/10.1177/0956797611417632.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology, 143*(2), 534–547. http://dx.doi.org/10.1037/a0033242.

Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants. *Cognition, 130*, 1–10. http://dx.doi.org/10.1016/j.cognition.2013.08.008.

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

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science, 13*(6), 907–912. http://dx.doi.org/10.1111/j.1467-7687.2009.00946.x.

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. http://dx.doi.org/10.1111/j.2044-835X.2011.02046.x.

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

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *British Journal of Developmental Psychology, 30*(1), 172–187. http://dx.doi.org/10.1111/j.2044-835X.2011.02067.x.

Tobii AB (2016). *Tobii Studio user's manual v3.4.5.* Tobii AB.

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

van der Wel, R. P. R. D., Sebanz, N., & Knoblich, G. (2014). Do people automatically track others' beliefs? Evidence from a continuous measure. *Cognition, 130*(1), 128–133. http://dx.doi.org/10.1016/j.cognition.2013.10.004.

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. http://dx.doi.org/10.1111/mila.12099.

Wellman, H. (2014). *Making minds: How Theory of Mind develops.* Oxford: Oxford University Press.

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

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

Yott, J., & Poulin-Dubois, D. (2016). Are infants' Theory of Mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development, 17*(5), 683–698. http://dx.doi.org/10.1080/15248372.2015.1086771.

Zmyj, N., Prinz, W., & Daum, M. M. (2015). Eighteen-month-olds' memory interference and distraction in a modified A-not-B task is not associated with their anticipatory looking in a false-belief task. *Frontiers in Psychology, 6*, 857. http://dx.doi.org/10.3389/fpsyg.2015.00857.