Cognitive Development xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Cognitive Development

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

How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span

Louisa Kulke[a,*], Mirjam Reiß[b], Horst Krist[b], Hannes Rakoczy[a]

[a] Department of Cognitive Developmental Psychology, Institute of Psychology, University of Göttingen, Leibniz-Science Campus Primate Cognition, Germany

[b] Department of Developmental Psychology and Educational Psychology, Institute of Psychology, University of Greifswald, Germany

ARTICLE INFO

Keywords:
Implicit theory of mind
Replication
Children
Life span
Validity

ABSTRACT

Recent findings from new implicit looking time tasks indicate that children show anticipatory looking patterns suggesting false belief processing from very early on; however, systematic and independent tests of their replicability and their convergent validity are still outstanding. The current paper reports three studies from two independent research labs that attempted to test the replicability and convergent validity (using correlation analyses) of the Southgate et al. (2007) and the Surian and Geraci (2012) paradigms. Results showed that the original findings can neither be replicated in children nor in elderly adults, and can only partially be replicated in adults. Furthermore, the two different paradigms did not correlate, which puts into question the convergent validity of these tasks as tapping the same capacity of an implicit Theory of Mind. In conclusion, the present studies suggest that the results from implicit Theory of Mind tasks should be treated with caution.

1. Introduction

Theory of Mind (ToM), the ability to attribute subjective mental states such as beliefs, desires and intentions is a capacity of fundamental importance to many aspects of our social lives. Explicit false belief (FB) tasks have been used as litmus tests for such an understanding. In particular, in so-called change-of-location false belief (FB) tasks, participants hear a vignette in which an object changes its location which is either witnessed by the protagonist (true belief [TB] control condition) or not (FB condition), and participants are then asked where the agent will look for the object (Wimmer & Perner, 1983). Decades of studies with such tasks have shown that children come to solve such FB tasks around age 4 (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983). Furthermore, superficially very different tasks systematically converge and correlate (Perner & Roessler, 2012; Wellman, Cross, & Watson, 2001), suggesting that they tap a common cognitive capacity, namely meta-representation (the capacity to represent others' representational states). This body of evidence was the basis for the traditional consensus that the development between 3 and 5 years marks a fundamental conceptual transition or even revolution.

However, this consensus has been challenged recently by a growing body of evidence from implicit ToM tasks. Since the pioneering work by Clements and Perner (1994) more and more studies have demonstrated implicit sensitivity to another agent's beliefs in younger children who do not yet master explicit tasks (e.g., Clements & Perner, 1994; Kovács, Téglás, & Endress, 2010; Low & Watts, 2013; Southgate, Chevallier, & Csibra, 2010; Southgate, Senju, & Csibra, 2007; Surian, Caldi, & Sperber, 2007;

* Corresponding author at: University of Göttingen, Institute of Psychology, Department of Cognitive Developmental Psychology, Waldweg 26, 37073 Göttingen, Germany.
E-mail address: lkulke@uni-goettingen.de (L. Kulke).

http://dx.doi.org/10.1016/j.cogdev.2017.09.001
Received 14 February 2017; Received in revised form 14 August 2017; Accepted 9 September 2017

Surian & Geraci, 2012). And recent adult work also suggests that these implicit and automatic capacities may remain intact and stable across the life span (Schneider, Bayliss, Becker, & Dux, 2012; Schneider, Lam, Bayliss, & Dux, 2012; Schneider, Slaughter, Bayliss, & Dux, 2013). Different kinds of implicit ToM tasks have been used, including violation of expectation (VoE) paradigms, interaction behavior (e.g., Buttelmann, Carpenter, & Tomasello, 2009; Buttelmann, Suhrke, & Buttelmann, 2015; Fizke, Butterfill, Van de Loo, Reindl, & Rakoczy, 2014; Knudsen & Liszkowski, 2012; Southgate et al., 2010), and anticipatory looking (AL) tasks (e.g., Clements & Perner, 1994; Low & Watts, 2013; Schneider, Bayliss et al., 2012; Senju, Southgate, White, & Frith, 2009; Southgate et al., 2007; Surian & Geraci, 2012). Of these, AL tasks are particularly interesting: In contrast to measures that tap differential retrodictive responses such as VoE, AL requires prediction; and in contrast to VoE and interaction measures, AL tasks can be (and have been) used in exactly the same kinds of ways across the lifespan (see e.g., Senju et al., 2009; Southgate et al., 2007).

Findings from AL measures have been part of the basis for ambitious and far-reaching theoretical conclusions to the effect that standard explicit tasks mask the true and early ToM competence that may even be innate (Baillargeon et al., 2015; Carruthers, 2013; Leslie, 2005). It is a very interesting question whether such positive findings, if they turned out to be robust, license such strong conclusions. It is another and much more fundamental question whether these findings actually are robust and replicable. In light of the general replication crisis (see e.g., Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013; Simmons, Nelson, & Simonsohn, 2011), questions of the reliability of these findings arise and need to be taken seriously. In particular, relatively few AL studies have been published to date, most of which have not been replicated outside of the lab in question or could not be replicated (Grosse Wiesmann, Steinbeis, Friederici, & Singer, 2017) or have used small sample sizes (Senju et al., 2009, 2010; Southgate et al., 2007).

Implicit ToM tasks involving AL measures were first used by Clements and Perner (1994) who studied anticipatory looking in response to a verbal prompt ("I wonder where she will …"). Subsequent studies then used spontaneous AL measures without any verbal prompting (e.g., Low & Watts, 2013; Schneider, Bayliss et al., 2012; Senju et al., 2009; Southgate et al., 2007; Surian & Geraci, 2012). Two AL tasks, in particular, have been used with children from very young ages on. Firstly, the Southgate/Senju paradigm has been used most widely in different populations, including infants (Southgate et al., 2007), children (Senju et al., 2010), adults (Senju et al., 2009) and participants with autism spectrum disorder (Senju et al., 2010, 2009). It is structured like standard change-of-location FB tasks (Clements & Perner, 1994), but the object is removed rather than transferred and there is no TB condition. Secondly, Surian and Geraci (2012) developed a standard change-of-location task with animated figures in which two figures chase each other and the protagonist forms a true or false belief about the other agent's location.

Both of these implicit ToM tasks are, in a broad sense, change-of-location FB tasks measuring anticipatory looking, but they differ in two crucial aspects. Strictly speaking, only the Surian & Geraci task is a proper change of location task. Firstly, the object is transferred between locations in the Surian & Geraci task and stays in the new location, like in standard change-of-location FB tasks, while it is relocated and then removed from the scene in the Southgate/Senju paradigm. Secondly, the Surian & Geraci task includes a TB control condition, as standard change-of-location tasks do, which is missing in the Southgate/Senju task.

Additionally, very little is known about whether different AL measures actually tap the same underling construct, which should result in high correlations of different tasks (convergent validity) (Heyes, 2014). Decades of studies have shown systematic, strongly converging and correlated performance in various superficially diverse explicit FB tasks and have thus supplied ample evidence for their converging validity. In contrast, so far hardly any tests for convergent validity of implicit tasks have been published. Rather, most studies only tested one local task (e.g., Clements & Perner, 1994; Low & Watts, 2013; Schneider, Bayliss et al., 2012; Senju et al., 2009; Southgate et al., 2007; Surian & Geraci, 2012). And those few studies that have used several tasks have failed to find any evidence for correlations (Yott & Poulin-Dubois, 2016).

Therefore, the current paper reports two sets of studies that systematically investigated the reliability and convergent validity of AL tasks independently in two labs. The first study describes findings from the first lab, attempting replication of one AL ToM task (Southgate et al., 2007) with a large sample of 2- to 6-year-old children. The second set of studies describes two replication experiments that stem from a second, independent lab. Study 2a tests an opportunity sample of children in the Southgate et al. task, including a comparably large age range. Based on previous research using the original paradigm, age should not affect our findings, as belief- congruent AL has previously been demonstrated in different age ranges for infants, (Southgate et al., 2007), children, (Senju et al., 2010), and adults (Senju et al., 2009). However, to ensure that the broad age range does not affect the findings, Study 2b tests narrow age ranges of children, adults and elderly adults to investigate developmental changes across the life span. It furthermore combines two implicit ToM paradigms (Senju et al., 2009; Southgate et al., 2007; Surian & Geraci, 2012) to investigate convergent validity.

2. Study 1

In this first study, the anticipatory-looking paradigm from Southgate et al. (2007) and Senju et al. (2010) was employed in an attempt to replicate the original findings and to trace the developmental course of (implicit) false-belief understanding in children aged 2–6 years. In close correspondence to Southgate et al. (2007), we presented children with a change-of-location task in which an actor reached through one of two windows to retrieve an object hidden in one of two containers. In a first condition of this task (FB1), the object was first moved between containers and then simply removed while the protagonist was not looking; whereas in a second condition (FB2), the protagonist did not observe how the object was first moved to the other container and then removed from the scene. As a potential measure of implicit false-belief attribution, children's anticipatory looking was analyzed via an eye tracker. Additionally, their explicit false-belief understanding was assessed by asking them where the actor would search for the object.

2.1. Method

2.1.1. Participants

For this study, 495 participants were recruited. Of these, 35 children had to be excluded because of missing looking data (n = 21) or other technical issues (n = 2), discontinuation of the experimental session (n = 6), experimenter error (n = 1), noncompliance (n = 2), or known mental retardation (n = 3). The remaining 460 children were divided into five age groups: There were 116 two-year-olds (44 female, *M* = 29.5 months, *SD* = 3.5, range = 24–35 months); 88 three-year-olds (45 female, *M* = 42.0 months, *SD* = 3.5, range = 36–47 months); 90 four-year-olds (38 female, *M* = 54.1 months, *SD* = 3.1, range = 48–59 months); 83 five-year-olds (43 female, *M* = 65.2 months, *SD* = 3.3, range = 60–71 months); and 83 six-year-olds (44 female, *M* = 76.6 months, *SD* = 3.3, range = 72–83 months). All children lived in or close to the city of Greifswald, located in the Northeast of Germany. They participated on a voluntary basis and with the consent of their parents. Children were rewarded with small presents and parents were compensated for travel costs.

2.1.2. Stimuli

Like in the original studies by Southgate and her collaborators (Senju et al., 2010; Southgate et al., 2007), familiarization and test events were presented as video clips and children's looking behavior was registered via an eye tracker. All videos were recorded following the scripts for the original videos employed by Southgate et al. (2007). A female actor stood behind a panel containing two windows, and in front of each window there was an opaque box with a lid. At the start of each trial, a puppet appeared at the center of the stage and placed a colored ball in one of the boxes (see Fig. 1). The actor wore a visor covering her eyes to eliminate the possibility of using eye gaze as a cue to where she would search. She did, however, always follow the puppet's movements with her head.

In four familiarization trials, children observed how the puppet opened the lid of the left- or right-hand box (Fig. 1a, b), put the ball into it, closed the lid, and then disappeared. After this, the actor retrieved the ball by reaching through the corresponding window on the left- or right-hand side. Before the window began to open, both windows were illuminated (for 1 s) and a chime sounded simultaneously to signal the forthcoming reaching action (2.75 s after the onset of illumination). Familiarization trials were presented in two different orders with respect to the placement of the ball: (a) *left − right − right − left*, or (b) *right − left − left − right*. After each familiarization trial an attention getter (i.e., a jiggling and sounding toy animation) appeared in the center of the computer screen.

Following familiarization, one of two false-belief test trials (FB1 or FB2) was presented. In the FB1 condition (Fig. 1c), the puppet opened the lid of the left-hand box, put the ball into it, closed the lid, returned to the center, retrieved the ball again, placed it in the center of the stage, opened the right-hand box, put the ball into it, closed the right- and the left-hand-box (in this order), and then disappeared from the scene. Then the sound of a phone ringing was heard, whereupon the actor turned away from the stage as if she was attending to the sound. As soon as the actor turned around, the puppet returned, retrieved the ball from the box, closed the lid, and disappeared from the scene. Once the phone stopped ringing, the actor turned back, the windows were illuminated, and the chime sounded.

In the FB2 condition (Fig. 1d), the puppet also put the ball into the left-hand box but disappeared from the scene right after this and before the phone was ringing and the actor turned around. Then the puppet returned, retrieved the ball from the left-hand box, placed it in the center of the stage, opened the right-hand box, put the ball into it, closed the right- and then the left-hand box, paused shortly, opened the right-hand-box, retrieved the ball, closed the box, and disappeared from the scene. After this the phone stopped ringing, the actor turned back, the windows were illuminated, and the chime sounded. Each test video continued for 5 s after the onset of illumination of the windows without any more action.

In the original studies (Senju et al., 2010; Southgate et al., 2007), the actor's turning direction was confounded with the belief condition: In the FB1 condition, the actor turned around in a clockwise direction and back in a counter-clockwise direction, whereas, in the FB2 condition, she moved in the opposite direction, respectively. We eliminated this confound by using both types of turning movements in both belief conditions.[1]

2.1.3. Procedure

Children were tested individually by a female experimenter. They watched the video clips on a computer monitor (17" monitor, 1280 × 1024 pixels) while their eye movements where recorded using a Tobii T120 eye tracker (Stockholm, Sweden). The videos were presented using Tobii Studio 2.2.7 software. Eye gaze was registered at a sampling frequency of 60 Hz and analyzed using the Tobii Fixation Filter with default parameters (velocity and distance threshold: 35 pixels/samples). Children sat at a distance of approximately 60 cm from the monitor, either on their parent's lap or on a child's seat (Tripp Trapp, Stokke AS, Norway). The eye tracker was calibrated individually using the regular 5-point calibration procedure of Tobii Studio with default settings (red dots, gray background, medium speed, full screen). If necessary, the calibration procedure was repeated before the experimental session started with the familiarization trials. After the test trial, while the last frame of the scene was still visible, children were asked "Where do you think is she going to look for the ball first?".

2.1.4. Design and distinctive features

Participants were randomly assigned (between participants) to one of the eight conditions resulting from the combination of the experimental variable *condition* (FB1 vs. FB2) with the control variables *order of presentation* (of the familiarization trials) and *turning*

[1] Examples of the videos employed in the present study can be viewed at https://youtu.be/21dII2jlpxw, https://youtu.be/f7YtiTCjRWE, and https://youtu.be/1pkNA2zzIrQ, for a familiarization, FB1, and FB2 trial, respectively.

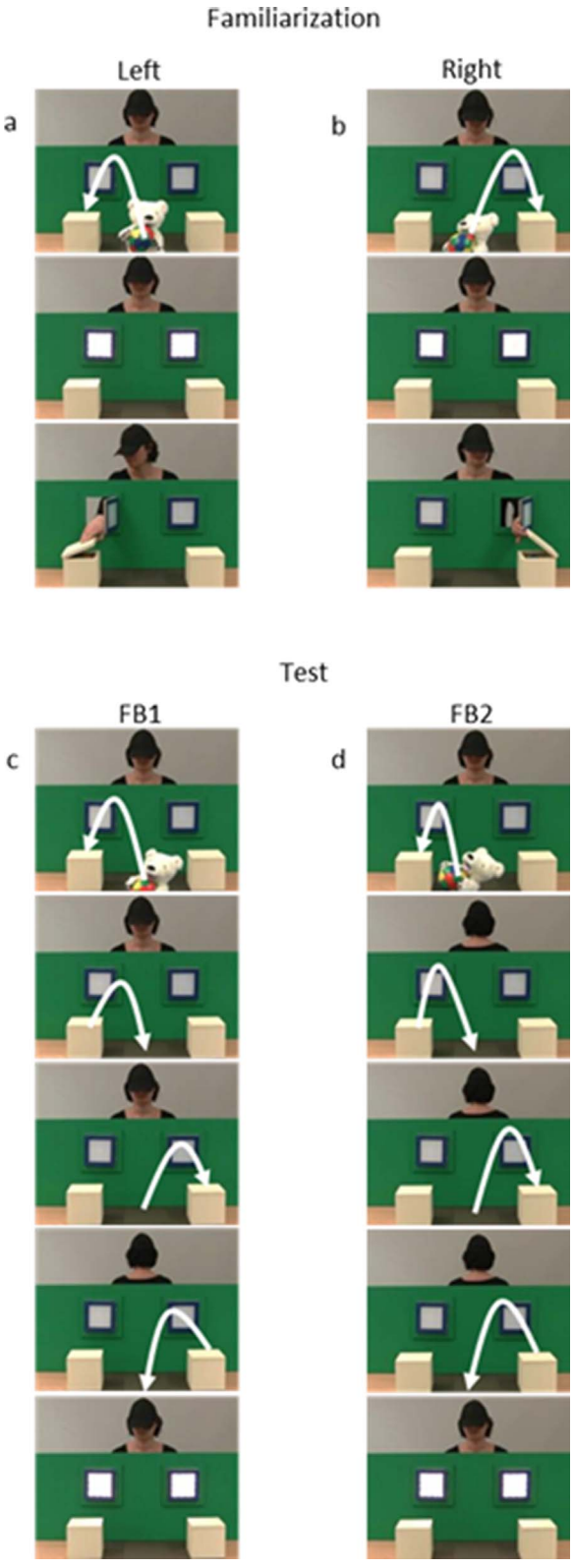

Fig. 1. Selected frames from familiarization and test trials in Study 1.

direction. Age group was included as an additional variable in a full-factorial between-subjects design.

The major deviations of the present study from those of Southgate et al. (2007) and Senju et al. (2010) can be summarized as follows: (1) While Southgate et al. (2007) tested 24- and 25-month-old toddlers and Senju et al. (2010) 6- to 8-year-old normally developing children (besides children with autism spectrum disorder), we recruited a large sample of toddlers and preschoolers (ranging from 2;0 to 5;11). (2) Like Senju et al. (2010), but unlike Southgate et al. (2007), we presented participants with four rather than two familiarization trials. (3) The order of the familiarization trials as well as the actor's turning direction were varied across participants.

2.2. Results

2.2.1. Replication analyses of original Senju measures

The full dataset of this study is provided in Kulke, Reiß, Krist, and Rakoczy (submitted). Children's anticipatory looking was analyzed in two ways: (1) by measuring their looking times to each window (during an interval of 5 s starting with the onset of illumination of the windows), and (2) by registering which of the two windows they fixated first (within an interval of 2 s after the onset of illumination of the windows). For this purpose, two rectangular areas of interest (AOIs) were defined covering each of the two windows. For the first measure, a differential looking score (DLS) was calculated as the difference between the looking time to the correct window and the incorrect window divided by the total looking time to both windows (no preference = 0, minimum = −1, maximum = 1):

$$DLS = \frac{\text{t(corr)-t(incorr)}}{\text{t(corr)} + \text{t(incorr)}}$$

There were 8 children (3 two-year-olds, 2 five-year-olds, and 3 six-year-olds) whose anticipatory looking could not be analyzed because of missing data.

Originally, it was planned to use the inclusion criterion applied by Senju et al. (2010), according to which only children who looked longer to the correct than the incorrect window in the last familiarization trial (after the illumination of the windows and before the respective window was opened) should be included for analysis (henceforth referred to as the *Senju criterion*). It turned out, however, that according to this criterion 220 children (47.8%) would have had to be excluded from analysis. In preliminary analyses, we therefore tested for both anticipatory-looking measures whether there were any significant differences between those children who passed the Senju criterion and those who did not. An ANOVA performed on the DLS including age group, condition (FB1 vs. FB2), and the Senju criterion as between-subjects variables, did not reveal any significant effects involving the Senju criterion, all Fs < 1 (all ps > 0.35, all $\eta_p^2 < 0.002$). Similarly, for the first-fixation measure ($n = 431$), there were no significant differences between children who passed the Senju criterion and those who did not, $p = 0.70$ (Fisher's exact test). Thus, the Senju criterion was not applied in the present study. Similar results were obtained with respect to the criterion used by Southgate et al. (2007, henceforth called *Southgate criterion*), according to which only those children were included for analysis who correctly anticipated the opening of the window in the last familiarization trial (DLS: all ps > 0.38, all $\eta_p^2 < 0.01$; first fixations: $p = 0.70$).

In the same manner, we tested for potential effects involving order of presentation and turning direction. As there were no such effects (order of presentation: all ps > 0.17; turning direction: all ps > 0.15), only age group and condition (FB1 vs. FB2) were included as between-subjects variables in the main ANOVA performed on the DLS.

2.2.1.1. Looking times. Overall, a mean DLS of 0.12 was obtained which indicated a small but reliable preference for the correct window, $F(1, 442) = 18.42$, $p < 0.001$, $\eta_p^2 = 0.04$. This overall result was qualified by a highly significant effect of condition, $F(1, 442) = 24.36$, $p < 0.001$, $\eta_p^2 = 0.05$. Children performed reliably above chance (i.e., > 0) in the FB1 condition ($M = 0.26$, 95% CI = 0.19 − 0.34), but, unlike in the original study, not in the FB2 condition ($M = -0.02$, 95% CI = −0.10 − 0.06). There was also a marginally significant effect of age group, $F(4, 442) = 2.22$, $p = 0.07$, $\eta_p^2 = 0.02$. As indicated by the confidence intervals included in Fig. 2, 2-year-olds did not perform reliably above chance in either condition, whereas 3-, 4-, 5-, and 6-year-olds exhibited a significant preference for the correct window in the FB1 but not the FB2 condition. The interaction between age group and condition did not reach statistical significance, however, $F(4, 442) = 1.70$, $p = 0.15$, $\eta_p^2 = 0.02$.

2.2.1.2. First looks. For the second measure of anticipatory looking, we analyzed children's first fixations within an interval of 2 s after the onset of illumination of the windows. From 29 children (6.3%) no first-fixation data were available because they did not fixate either of the windows during this interval. Among the remaining 431 children, 232 (53.8%) fixated the correct window first, $p = 0.12$ (binomial test). Preliminary analyses yielded no significant effects of order of presentation, $p = 0.29$, or turning direction, $p = 1.0$ (Fisher's exact test). Children performed reliably above chance in the FB1 condition (58.6% correct), $p = 0.01$, but, unlike in the original study, not in the FB2 condition (48.8% correct), $p = 0.78$. This difference was statistically reliable, $p = 0.04$ (Fisher's exact test). There was no significant age effect, $\chi^2(4, N = 431) = 4.46$, $p = 0.35$.[2] Fig. 3 illustrates the first-fixation results as a function of age group and condition.

2.2.1.3. Answers. Children's verbal responses or pointing gestures were coded into three categories: (1) correct, (2) incorrect, and (3)

[2] There were 113 children (25% of the total sample) who already fixated one of the AOIs when the signal appeared, implying that their first saccade was directed away from the respective AOI. For analyses of the first saccades (which require a movement of the eye rather than a fixation, as first looks do), these participants had therefore to be excluded. Results were essentially the same for first saccades as for first fixations, with the exception that, for the former, neither the preference for the correct window in the FB1 condition, p = 0.11 (binomial test), nor the condition effect (FB1: 56.6% vs. FB2: 49.7%) reached statistical significance, p = 0.26 (Fischer's exact test).

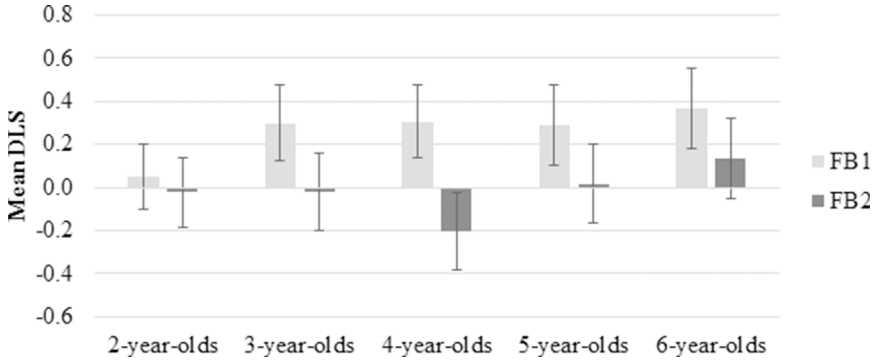

Fig. 2. Mean differential looking scores (DLS) in the Southgate/Senju task for the different age groups in FB1 and FB2 conditions. Error bars depict 95% CIs.

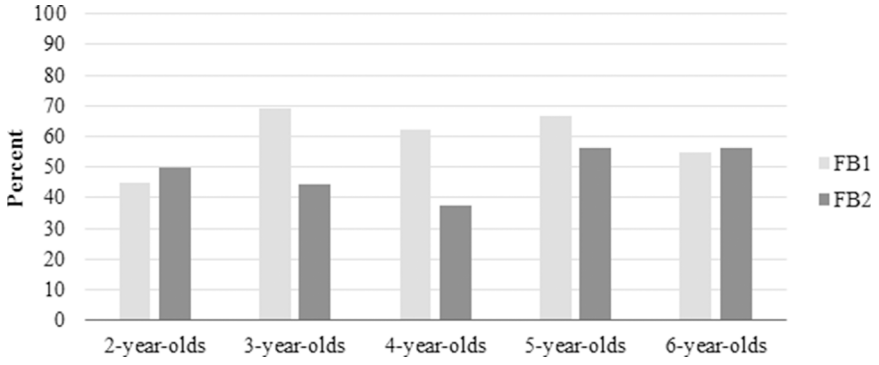

Fig. 3. Percent of correct first fixations in the FB1 and FB2 conditions.

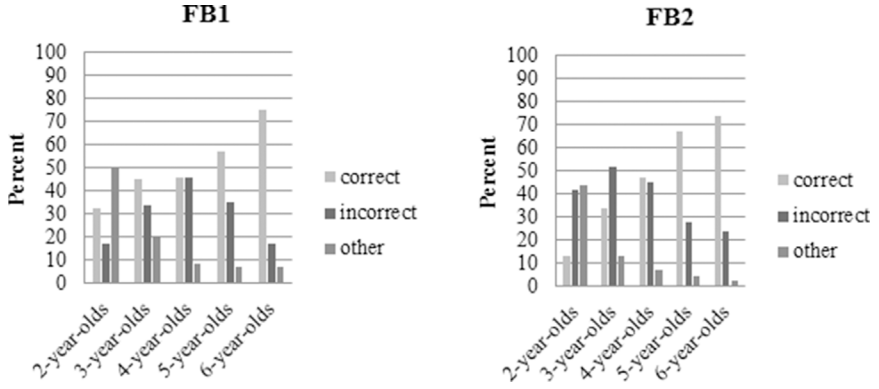

Fig. 4. Percent of correct, incorrect, and other responses in the verbal response task.

other (no answer, “don’t know”, etc.). An answer was coded as correct if the child referred to the correct window and as incorrect if he or she referred to the incorrect window. The overall percentage of correct, incorrect, and other answers amounted to 47.6%, 33.7%, and 18.7%, respectively. Performance increased significantly with age, $\chi^2(8, N = 460) = 115.5, p < 0.001$, both in the FB1 condition, $\chi^2(8, N = 237) = 56.2, p < 0.001$, and in the FB2 condition, $\chi^2(8, N = 233) = 68.4, p < 0.001$ (Fig. 4). There was also a marginally significant effect of condition, $\chi^2(1, N = 460) = 5.45, p = 0.07$. This effect tended to favor the FB1 over the FB2 condition; it reached statistical significance with the 2-year-olds, $\chi^2(2, N = 116) = 11.0, p = 0.004$, but not with the other age groups, $ps > 0.20$.

To shed light on the relation between children’s verbal responses as a direct measure of their false-belief understanding and the indirect measures obtained from their looking behavior, we compared the DLS as well as the first fixation (correct vs. incorrect window) of the children who gave correct answers to the AL measures from those who did not give a correct answer. DLS and first fixations were therefore analyzed as in the main analyses including answer (correct vs. incorrect/other) as an additional between-subjects variable. It turned out that children who answered correctly, indeed, exhibited a higher DLS ($M = 0.20$; 95% CI = 0.11 − 0.29) than those who did not ($M = 0.06$; 95% CI = −0.03 − 0.15). This difference was statistically reliable, $F(1, 432) = 4.91$, $p = 0.03$, $\eta_p^2 = 0.01$, and did not depend on age group or condition (FB1 vs. FB 2), $ps > 0.41$. First fixations were, however, not

associated with children's answers, neither overall, $p = 0.70$, nor within any age group, $ps > 0.38$ (Fisher's exact test).

2.2.2. Replication analyses of original Southgate measures: 24- and 25-month-olds' anticipatory looking

For a comparison of our results to those obtained by Southgate et al. (2007) with a sample of 24- and 25-month-olds, we followed their analytic method to assess the anticipatory-looking behavior within a corresponding subsample of our participants. Instead of using the (statistically more appropriate) DLS measure,[3] Southgate et al. compared the absolute looking times to each window. In our subsample of 24- and 25-month-olds, there were 20 children who looked at one of the windows during the first two seconds after the onset of illumination of the windows[4] (FB1 condition: six 24- and five 25-month-olds; FB2 condition: five 24- and four 25-month-olds). These children looked about equally to the correct (457 ms) and the incorrect window (438 ms). A repeated measures ANOVAs with window (correct vs. incorrect) as a within-subjects variable and condition (FB1 vs. FB2), did neither yield a significant window effect, $F < 1$, nor a significant interaction between condition and window, $F(1, 18) = 1.58$, $p = 0.23$.

As Southgate et al. (2007) only included children in their analysis who anticipated the correct window in the last familiarization trial, we also examined those 24- and 25-month-olds separately who satisfied the Southgate criterion ($n = 9$). The respective children did not look significantly longer to the correct (477 ms) than to the incorrect window (364 ms), $F < 1$. Compared to the looking times reported by Southgate et al. (2007; approx. 1000 and 500 ms for the correct and incorrect window, respectively), the corresponding looking times were much shorter in the present study. However, the difference of the looking times to the correct and the incorrect window (i.e., the window effect) and the sum of these looking times were neither significantly correlated in the subsample of the 24- and 25-month-olds ($N = 20$), $r = 0.07$, $p = 0.76$, nor in our entire sample (N = 431), r = 0.03, $p = 0.56$. Thus, the window effect did not vary with children's total looking time to either of the two windows.

With regard to 24- and 25-month olds' first fixations, the findings reported by Southgate et al. (2007) could not be replicated either: There were only 8 children (40%) who fixated the correct window first, while 12 (60%) fixated the incorrect window first, $p = 0.50$ (binomial test). The percentage of correct first fixations neither depended on the false-belief condition (FB1: 27.3%, FB2: 55.6%), $p = 0.36$ (Fisher's exact test), nor did it differ between children who met the Southgate criterion and those who did not (33.3% vs. 45.5%, respectively), $p = 0.67$.

2.3. Discussion

Using the Southgate/Senju anticipatory-looking paradigm, we investigated false-belief understanding in a sample of 460 children aged 2–6 years. Only in the FB1 condition in which, unnoticed by the protagonist, the object was removed from the scene, did we find clear evidence for correct anticipation: both with respect to the relative duration children looked at the correct window (differential looking score, DLS) and their first fixation after the illumination of the windows. No such evidence was found, however, for the FB2 condition, where, unnoticed by the protagonist, the object was relocated and then removed from the scene. Southgate et al. (2007) designed the two false-belief conditions to exclude certain low-level accounts of children's performance: Success in the FB1 condition would rule out the possibility that children's looking was due to the last position of the protagonist's attention, and success in the FB2 condition would exclude that children's looking was due the last (hiding) position of the ball. As, in the present study, children did not perform significantly above chance in the FB2 condition, the positive results in the FB1 condition do not provide unequivocal evidence for false-belief understanding. Children's looking might indeed have been influenced by low-level cues such as the ball's last hiding location.

As expected, the percentage of correct answers to the verbal question at the end of the test trial improved markedly with age. Children tended to give more correct answers in the FB1 than in the FB2 condition but the condition effect was only moderate and reached significance in the youngest age group only. The reason for this particular, somewhat puzzling pattern of results concerning the condition effect is not yet clear. One confound, potentially affecting the side children choose, was that the screen side containing the correct AOI was confounded with condition (FB1/FB2), like in the original study. In future research using similar paradigms, the side of the correct AOI should be controlled for.

Interestingly, children's answers were not independent of their looking behavior. Children who tended to look to the correct window, presumably in anticipation of the protagonist's reaching action, also gave more correct answers. This was true independently of age and condition. However, children's answers were significantly associated with their DLS but not with their first fixations. Therefore, it cannot be ruled out that the former association stemmed from children's tendency to point to the window to which their last gaze was directed rather than from the fact that their looking behavior reflected their explicit reasoning.

3. Study 2

Studies 2a and 2b describe replication attempts from a second, independent lab, that test the Southgate/Senju paradigm (Studies 2a and 2b) and the Surian & Geraci paradigm (Study 2b) in children and adults.

[3] The DLS accounts for differences in absolute looking time to the areas of interest by dividing looking times to each AOI by the total time spent looking at both AOIs. It is thus less susceptible to outliers.

[4] Southgate et al. (2007) used an interval of 1.75 s in their analyses. We chose the slightly larger interval of 2 s because children were actually familiarized to a delay of 1.75 s between the *end* of illumination and the opening of a window, both in the original videos and in the present ones, but not to a delay of 1.75 s after the *onset* of illumination (as accidently stated by Southgate et al., 2007, p. 590).

3.1. Study 2a

Study 2a aimed to replicate the Southgate/Senju paradigm described in Study 1 in an independent lab setting by testing a fairly wide age range of children.

3.1.1. Method

3.1.1.1. Participants. An opportunity sample of $N = 52$ neurotypical children ($M = 60.7$ mos, $SD = 25.75$ mos, *range* = 24-127 mos, 34 female) participated at a family fair after parental consent was obtained. Children received a sticker in return for their participation.

3.1.1.2. Materials and stimuli. Children watched one of two original videos from Senju et al. (2009) (between-subjects variable condition FB1/FB2), which is comparable in structure to the paradigm described in Study 1 (see Fig. 1 for a schematic display of the conditions). A remote SMI REDn eye tracker was used to monitor participants' gaze. Before recording started, participants completed a standard 5-point calibration and validation routine. Movies were controlled in the SMI Experiment Center (version 3.5.169) and presented on a 15.6 inch LCD screen (1920 × 1080 pixel). Gaze information was saved for offline analysis, which was conducted using BeGaze Software (version 3.5.101) for visualization and AOI processing exported to IBM SPSS Statistics (version 22) for analysis.

3.1.1.3. Procedure. Participants were seated in front of a laptop with the eye tracker attached to it and a 5-point calibration and validation was obtained. The whole testing procedure took approximately 3 min. The condition was randomly assigned to each participant.

3.1.1.4. Analysis. Based on the original analysis procedure (Senju et al., 2009), the first saccade after the chime sound and fixations to both windows were measured during the freeze period and a DLS was calculated as described in Study 1. As there was no significant difference in first saccade direction or DLS between participants passing or failing the inclusion criteria (see analyses in Supplement A), all participants were included in the following analyses to increase the sample size. The full datasets are provided in .

3.1.2. Results

3.1.2.1. Looking times. Mean looking times, measured through DLS are depicted in Fig. 5 as a function of condition (FB1/FB2). A *t*-test showed that the looking score ($M = 0.27$, $SD = 0.656$) was significantly higher than zero, $t(42) = 2.70$, $p = 0.010$, $d = 0.833$, suggesting that participants looked significantly longer to the correct location. Separate *t*-tests for both conditions showed that DLS only significantly differed from zero in FB1 ($M = 0.51$, $SD = 0.495$), $t(23) = 5.08$, $p < 0.001$, $d = 2.119$, but, unlike in the original study, not in FB2 ($M = -0.04$, $SD = 0.716$), $t(18) = -0.23$, $p = 0.821$, $d = -0.108$.

3.1.2.2. First saccades. Eight participants showed missing data at the chime tone, meaning that their first saccade direction could not be investigated. A binomial test for single samples was used to investigate whether the first saccade was significantly more often directed towards the correct than towards the incorrect location. Participants with missing data were excluded from the analysis. Unlike in the original study, the difference between saccades to the correct (27) and incorrect (17) location was not significant in a binomial test, $p = 0.174$. Note that findings are comparable if first looks are used as an outcome measure (see Supplement B).

3.1.3. Discussion of study 2a

Study 2a aimed to replicate findings from Southgate et al. (2007) and Senju et al. (2009) in an independent setting on a sample of children between two and eleven years. Results show that participants spent significantly more time looking at the correct than the

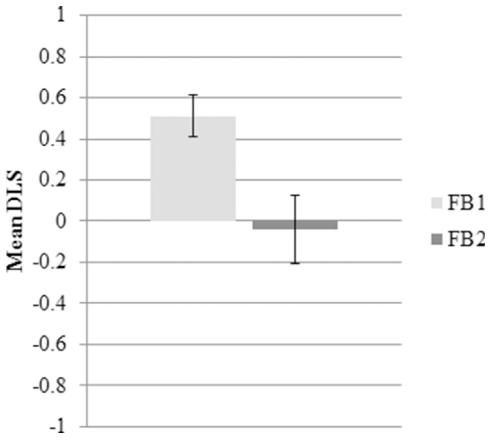

Fig. 5. Differential looking scores (DLS) as a function of condition in Study 2a. Error bars display the Standard Error.

incorrect location, replicating findings from Senju et al. (2009). However, a more detailed analysis showed that this effect was only significant in the condition that involved immediate disappearance of the ball after the agent turns away from the scene (FB1), not in the condition in which the ball was moved between boxes before being removed from the scene (FB2). These findings are in line with Study 1. Higher scores in this FB1 condition than in the FB2 condition have also been described by Senju et al. (2010), however the current study suggests that the effect is not significant in the FB2 condition. The direction of the first saccade was not significantly more often directed towards the correct location, in contrast to previous studies using this paradigm (Senju et al., 2009; Southgate et al., 2007). This is in line with Senju et al. (2010), who also did not replicate previous findings on first saccade direction. However, Study 2a tested an opportunity sample with a fairly wide age range, which might have affected the findings.

3.2. Study 2b

To address the possibility that the age range might affect the findings, we conducted a second systematic replication study (Study 2b), testing a narrow age range of children (4;6 to 5;6 years), adults, and elderly adults. In this study, we combined the Southgate/Senju task with the Surian & Geraci paradigm. If both tasks are valid and measure the same construct, they should both be replicable in this controlled study with narrow age groups and they should correlate. Furthermore, if implicit ToM is a robust phenomenon that remains largely unchanged over the lifespan, similar AL behavior should be found in all age groups.

3.2.1. Method

3.2.1.1. Participants. Sixty-four neurotypical children (M = 5.01 years, SD = 0.392, range = 4.5-5.5 years), 80 neurotypical adults (M = 24.07 years, SD = 2.952) and 83 elderly adults (M = 71.43 years, SD = 6.953) volunteered to participate. All participants had normal or corrected to normal vision and reports of known visual symptoms (e.g., cataracts in elderly participants) were noted. Participants who did not pass inclusion criteria based on original studies were excluded from further analyses, leading to a total participant number of 40 children, 40 adults and 40 elderly adults. Consent was obtained from adult participants and from the children's parents. The study was approved by the ethics committee of Göttingen University (Ethics code: 143a).

3.2.1.2. Materials and stimuli. The original stimuli from Senju et al. (2009) and Surian and Geraci (2012) were provided by the authors.[5] A description of the Senju et al. (2009) can be found in the methods section of Study 2a. The Surian and Geraci (2012) videos show a triangle chasing a ball through a Y-shaped tunnel. In two familiarization trials the triangle chases the ball until it enters the Y-shaped tunnel from the bottom end. It then waits until the ball reappears (either at the left or right top end) and hides in a box at that end. The triangle then moves into the tunnel and reappears at the same end as the ball. The two test trials are identical to the familiarization trials up to the point where the ball enters the box. At this point the triangle either leaves the scene, the ball moves from the box at one end of the tunnel to the box at the other end and the triangle then reappears to the scene (false belief condition) or the triangle witnesses that the ball moved to the other box, then disappears and reappears (true belief condition). The triangle then enters the tunnel and reappears at the belief-congruent location.

An SMI REDn Scientific remote eye tracker was used to monitor participants' gaze. Before recording started, participants completed a standard 5-point calibration and validation routine. Movies were presented on a laptop computer (DELL Precision M4800 with a Windows 8.1 pro operating system) controlled via the SMI Experiment Center (version 3.5.169) and presented on a 15.6" LCD display (1920 × 1080 pixel). Gaze information was recorded at 60 Hz and saved for offline analysis, which was conducted using BeGaze Software (version 3.5.101) for visualization and exported to IBM SPSS Statistics (version 22) for further analysis.

3.2.1.3. Design. In a mixed design, all participants completed the Southgate/Senju and the Surian & Geraci task. The belief condition (FB1/FB2) in the Southgate/Senju paradigm was manipulated between participants and the Surian & Geraci conditions (TB/FB) were compared within participants, as in the original studies respectively. The original outcome measures were adopted in the current study; therefore, first saccades and DLS are reported for the Southgate/Senju paradigm and first saccades and proportion of looking times for Surian & Geraci. However, DLS is used to analyze correlations of the different paradigms as it is independent of the time window chosen for looking time analysis.

3.2.1.4. Procedure. Participants were seated in front of a laptop with the remote eye tracker attached to it. The participants' distance and angle to the screen were adjusted to the height of the participant to ensure the best possible eye-tracking signal. Participants were presented with both paradigms in a randomized order. Each paradigm was preceded by a 5-point calibration and validation routine. The Southgate/Senju condition (FB1 or FB2) was randomly selected for each participant. Participants received both a TB and an FB condition in the Surian & Geraci paradigm, however, the order of the test trials was randomized and the screen side (left, right) was counterbalanced. The whole testing procedure took approximately 10 min.

3.2.1.5. Analysis. Participants were only included in the analysis if they passed the inclusion criteria defined by Senju et al. (2009) (longer looking towards the correct than the incorrect box in the last familiarization trial, applied to analyses of the Southgate/Senju

[5] Note that of the Surian & Geraci paradigm only the original FB and TB videos for one object location were available from the authors. Therefore, in order to randomize the object location (left/right), as described in the original study, we mirrored the original videos at the central vertical axis to create identical videos with the object location on the respective other side.

data) or by Surian and Geraci (2012) (longer looking towards the correct than towards the incorrect box in at least one of the two familiarization trials, applied to analyses of the Surian & Geraci data) or both (applied to correlation analyses of both paradigms). However, Supplement C shows that similar results are achieved when all participants are included in the analysis.

Measurements of looking times towards the AOIs were based on the original studies. In the Senju et al. (2009) paradigm, looking times were measured using the original measures described above. As the main Senju/Southgate measure, a differential looking score (DLS) was calculated as described in the previous studies. In the original paper introducing the Surian and Geraci (2012) paradigm, looking times towards an area including the boxes in the time window between the triangle entering the tunnel and its reappearance were measured. As in the original paper, looking times were the main measure, here depicted as the proportion of looking time on the AOIs during the defined time window. In addition to looking time measures, first saccades were analyzed, in line with the original analyses. Additionally, the DLS was used for correlation analyses of the different paradigms, as it is comparable and accounts for the different time windows defined for the AOIs. Full datasets of Study 2b are provided in .

3.2.2. Results

3.2.2.1. Replication of original Southgate/Senju analyses

3.2.2.1.1. Looking times. The patterns of looking times, indicated by mean DLS, are depicted in Fig. 6 as a function of age and condition. A mixed linear model including age group (children, adults, elderly adults) and belief condition (FB1 or FB2) as fixed factors showed a significant effect on DLS of age group, $F(2, 119) = 4.567$, $p = 0.012$, belief, $F(1, 119) = 39.442$, $p < 0.001$, and a significant interaction, $F(2, 119) = 7.288$, $p = 0.001$. Follow-up independent samples *t*-tests showed that there was a significant difference in DLS between children ($M = -0.036$, $SD = 0.706$) and adults ($M = 0.323$, $SD = 0.629$), $t(80) = -2.433$, $p = 0.017$; there was no significant difference in DLS between adults and elderly adults ($M = 0.156$, $SD = 0.628$), $t(83) = 1.230$, $p = 0.222$ or between children and elderly adults, $t(81) = -1.306$, $p = 0.195$. There was a significant difference in DLS between belief condition FB1 ($M = 0.458$, $SD = 0.517$) and FB2 ($M = -0.142$, $SD = 0.662$), $t(123) = 5.631$, $p < 0.001$.

One sample *t*-tests were conducted to test whether the DLS significantly differed from zero, and differences between FB1 and FB2 conditions were investigated with independent samples *t*-tests for each age group.

In children, DLS significantly differed from zero in the positive direction in the FB1 condition, $M = 0.526$, $SD = 0.441$, $t(19) = 5.341$, $p < 0.001$, $d = 2.451$, and, unlike in the original study, significantly differed from zero in the negative direction in the FB2 condition, $M = -0.598$, $SD = 0.405$, $t(19) = -6.597$, $p < 0.001$, $d = -3.027$. The difference between conditions was significant, $t(38) = 8.397$, $p < 0.001$, $d = 2.724$.

In adults, DLS significantly differed from zero in the positive direction in the FB1 condition, $M = 0.435$, $SD = 0.590$, $t(18) = 3.211$, $p = 0.005$, $d = 1.514$, but, unlike in the original study, not in the FB2 condition, $M = 0.231$, $SD = 0.658$, $t(22) = 1.686$, $p = 0.106$, $d = 0.719$. However, there was no significant difference between FB1 and FB2 conditions, $t(40) = 1.044$, $p = 0.303$, $d = 0.330$.

In elderly adults, DLS significantly differed from zero in the positive direction in the FB1 condition, $M = 0.416$, $SD = 0.531$, $t(21) = 3.678$, $p = 0.001$, $d = 1.605$, but, unlike in the original study, not in the FB2 condition, $M = -0.117$, $SD = 0.616$, $t(20) = -0.872$, $p = 0.394$, $d = -0.390$. The difference between conditions was significant, $t(41) = 3.046$, $p = 0.004$, $d = 0.951$.

3.2.2.1.2. First saccades. The frequency of correct and incorrect saccades in FB1 and FB2 conditions is depicted in Table 1. Sixteen participants showed missing data at the chime tone, meaning that their first saccade direction could not be investigated and another 44 participants already looked at one AOI during coding onset. A binomial test for single samples was used to investigate whether the first saccade was significantly more often directed towards the correct than towards the incorrect location. Participants with missing

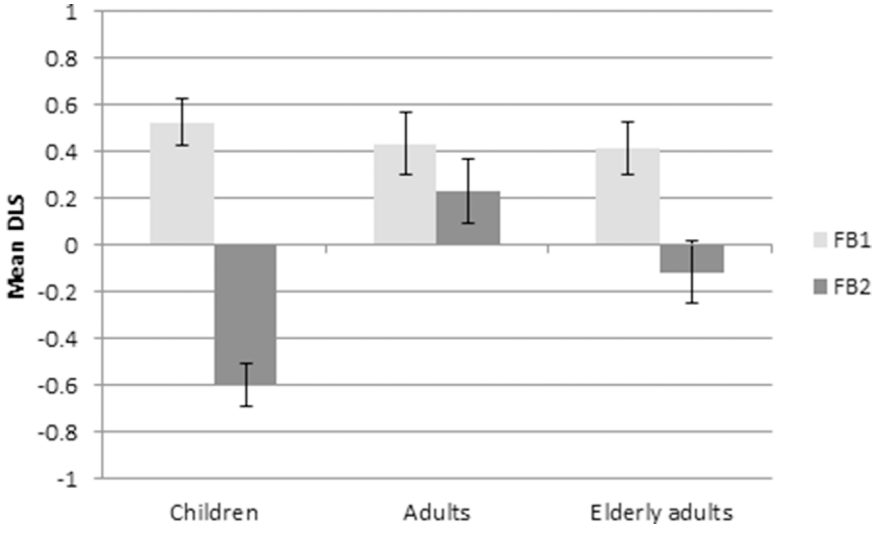

Fig. 6. Mean differential looking scores (DLS) in the Southgate/Senju task as a function of age group and condition in Study 2b. Error bars depict the Standard Error.

Table 1
Frequency of looks to the correct and incorrect location and significance level of the differences for different age groups and conditions.

	Children			Adults			Elderly		
	correct	incorrect	p-value	correct	incorrect	p-value	correct	incorrect	p-value
FB1	1	3	0.625	8	3	0.227	4	10	0.180
FB2	2	8	0.109	9	5	0.424	8	5	0.581

data were excluded from the analysis. Unlike in the original study, the difference between saccades to the correct (32) and incorrect (34) location was not significant in a binomial test, $p = 0.902$. Follow up tests showed that there was no significant difference in any age group or condition (Table 1). Note that findings are comparable if first looks are used as an outcome measure (see Supplement B).

3.2.2.2. Replication of original Surian & Geraci analyses

3.2.2.2.1. Looking times. Looking times, measured through mean proportions of looking in the different age groups and conditions are depicted in Fig. 7. Mixed linear models including participant ID as random factor and belief (true or false), location (containing object or not) and age group (child, adult, or elderly adults) as fixed factors computed the effect on the proportion of looking time on the AOIs during the predefined time window.

Overall, there was a significant effect of object location, $F(1, 613) = 53.442$, $p < 0.001$, with higher proportions of looking towards the object location, $M = 0.243$, $SD = 0.257$, than towards the empty location, $M = 0.130$, $SD = 0.184$. No other effects were significant.

When conducting the analyses for each age group separately, in children, there was a significant effect of location, $F(1, 232) = 12.252$, $p < 0.001$, with higher proportions of looking towards the object location, $M = 0.233$, $SD = 0.187$ than towards the empty location, $M = 0.152$, $SD = 0.166$ (but, unlike in the original study, no interaction of belief and location, $F(1, 232) = 0.189$, $p = 0.664$).

In adults, there was a significant effect of location, $F(1, 280) = 25.695$, $p < 0.001$ with higher proportions of looking towards the object location, $M = 0.305$, $SD = 0.312$, than towards the empty location, $M = 0.143$, $SD = 0.220$, and a significant interaction of belief and object location, $F(1, 280) = 4.296$, $p = 0.039$. Follow-up t-tests showed that participants looked significantly longer at the object location ($M = 0.354$, $SD = 0.323$) than the empty location ($M = 0.126$, $SD = 0.205$) in the TB condition, $t(119) = -5.017$, $p < 0.001$, $d = -0.920$ and also in the FB condition, $t(133) = -2.133$, $p = 0.035$, $d = -0.370$, (object location: $M = 0.256$, $SD = 0.295$, empty location: $M = 0.160$, $SD = 0.235$). There was no significant difference between proportions of looking time in the FB and TB conditions at the empty location, $t(140) = -0.922$, $p = 0.358$, $d = -0.156$ or at the ball location, $t(140) = -1.888$, $p = 0.061$, $d = -0.319$.

In elderly adults, there was a significant effect of location, $F(1, 219) = 18.601$, $p < 0.001$ with higher proportions of looking towards the object location, $M = 0.192$, $SD = 0.235$ than towards the empty location, $M = 0.098$, $SD = 0.155$ (but, unlike in the original study, no interaction of belief and location, $F(1, 219) = 0.043$, $p = 0.836$).

Note that in the original paper Surian & Geraci conducted a second analysis in which they computed the effects on looking time described above after preselecting hypothesis-conform participants based on first saccades. Supplement D reports our results for this preselection of participants. Note, however, that preselecting participants based on their hypothesis-conform behavior in one outcome measure (first saccades) that is related to the second measure (looking time) distorts the sample towards a hypothesis-conform outcome. Overall, the findings are comparable showing no interaction of location and belief in children and elderly adults but a significant interaction in adults.

3.2.2.2.2. First saccades. The frequency of correct and incorrect saccades in TB and FB conditions is depicted in Table 2. A

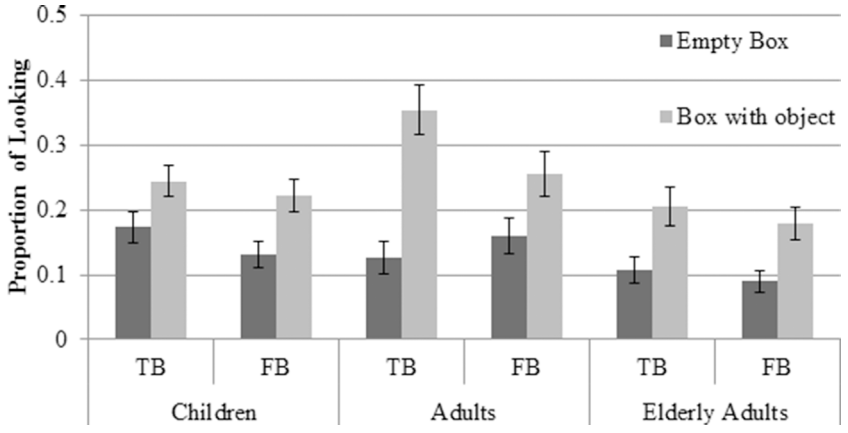

Fig. 7. Proportion of looking as a function of condition and age group. Error bars depict the standard error.

Table 2
Frequency of looks to the correct (corr.) and incorrect (incorr.) location and significance level of the differences for different age groups and conditions in the Surian & Geraci paradigm.

	Children			Adults			Elderly		
	Corr.	Incorr.	p-value	Corr.	Incorr.	p-value	Corr.	Incorr.	p-value
TB	27	23	0.672	34	23	0.185	31	26	0.597
FB	24	29	0.583	26	29	0.788	21	38	0.036

binomial test for single samples was used to investigate whether the first saccade was significantly more often directed towards the correct than towards the incorrect location, separately for the TB and FB condition. Participants with missing data were excluded from the analysis. Unlike in the original study, the difference between saccades to the correct and incorrect location was not significant in a binomial test, in the TB condition (correct: 92, incorrect: 72), $p = 0.138$ or in the FB condition (correct: 71, incorrect: 96), $p = 0.063$. Follow up tests showed that there was no significant difference in any age group or condition, except that elderly adults showed more first looks to the incorrect ($n = 38$) than the correct location ($n = 21$) in the FB condition, $p = 0.036$ (Table 2).

In contrast to the other paragraphs, this paragraph is not formatted in the "justified" but in the "left-aligned" format.3.2.2.3 Correlation of Southgate/Senju and Surian & Geraci paradigms

Pearson correlations were computed to investigate the relation between, firstly, the TB and the FB condition within the Surian & Geraci paradigm and secondly, the Southgate/Senju paradigm and the TB and FB condition of the Surian & Geraci paradigm (Table 3). To investigate overall belief processing in the Surian & Geraci paradigm, an additional composite DLS was calculated that accounts for belief processing in both TB and FB conditions (see Supplement E for further details). Within the Surian & Geraci paradigm, the TB and FB conditions negatively correlated in elderly adults only, $r(31) = -0.450$, $p = 0.011$. There were no significant correlations between the Southgate/Senju and the Surian & Geraci paradigm for the full sample or for the separate age groups for any of the three measures (TB, FB, composite TB/FB score), but only marginal correlation of TB and FB conditions within the same (Surian & Geraci) paradigm (see Table 3 for details).

4. General discussion

The current paper reports two sets of studies that attempted to replicate implicit ToM tasks and correlated the outcomes to test their convergent validity. Study 1 tested 460 children between two and six years of age in the Southgate/Senju paradigm and could only reliably replicate the findings from the FB1 but not the FB2 condition of the paradigm. Study 2a tested children between two and eleven years on the Southgate/Senju task and also showed that earlier findings from the FB1 condition could be replicated, but not from the FB2 condition. Study 2b tested children, adults, and elderly adults on two different paradigms − the Southgate/Senju and the Surian & Geraci task. In the Southgate/Senju task, the original findings from the FB1 condition could reliably be replicated in all age groups whereas the FB2 condition could only partially be replicated in adults and only if including all participants irrespective of whether they passed the familiarization trials or not (see Supplement C). In the Surian & Geraci paradigm, the original belief-congruent looking patterns (looking longer to the location with the target object in TB, but to the empty location in FB) could only be replicated in the TB condition. In the FB condition, in contrast to the original findings, participants mostly looked at the belief-incongruent location where the object really was. The only indication of belief sensitivity was a small interaction effect between location and condition in adults, such that participants' looking to the location with the target object was less pronounced in FB than in TB. Concerning convergent validity, no correlations between the two types of tasks were found. In summary, neither the Southgate/Senju, nor the Surian & Geraci paradigm could be fully and robustly replicated in the current paper. Only the more ambiguous and inconclusive (FB1) or the control conditions (TB) which are subject to alternative explanations could be replicated.

The familiarization trials show that almost half of the children tested in the present studies did not look longer to the correct than the incorrect window in the fourth and last familiarization trial (i.e., they did not meet the Senju criterion for inclusion) and that there was no significant age effect in this respect. Furthermore, it turned out that children who met the Senju or the Southgate criterion for inclusion exhibited a similar looking behavior in the test trial as those who did not. This questions the appropriateness

Table 3
Correlations of the Southgate/Senju and Surian & Geraci paradigms in different age groups as well as in the overall sample.

Correlated conditions	**Children**	**Adults**	**Elderly adults**	**Overall**
Surian & Geraci TB & FB	$r(38) = 0.316$, $p = 0.053$	$r(32) = 0.346$, $p = 0.052$	$r(31) = -0.450$, $p = 0.011$	$r(101) = 0.101$, $p = 0.315$
Surian & Geraci TB and *Southgate/Senju*	$r(37) = -0.171$, $p = 0.311$	$r(35) = 0.210$, $p = 0.226$	$r(34) = 0.256$, $p = 0.144$	$r(106) = 0.119$, $p = 0.224$
Surian & Geraci FB and *Southgate/Senju*	$r(38) = -0.127$ $p = 0.446$	$r(34) = 0.044$, $p = 0.804$	$r(33) = 0.064$, $p = 0.725$	$r(105) = 0.002$, $p = 0.982$
Surian & Geraci Comp. and *Southgate/Senju*	$r(37) = -0.186$, $p = 0.269$	$r(31) = 0.239$, $p = 0.196$	$r(35) = 0.276$, $p = 0.109$	$r(103) = 0.115$, $p = 0.247$

for assessing participants' (correct or incorrect) expectations concerning the protagonist's actions. Furthermore, gaze patterns reflecting ToM processing were also not found in children tested at an age at which it is established that they pass explicit ToM tasks. Our results thus cast doubts on the validity of the Southgate/Senju paradigm with regard to assessing false-belief understanding and question its reliability with respect to measuring children's expectations.

It is interesting to note that–by contrast to the original study conducted with 25-month-olds (Southgate et al., 2007)–Senju et al. also found a significant condition effect favoring the FB1 condition with the DLS measure and that their first-fixation results were also negative (for both conditions combined; see Senju et al., 2010, p. 358, for details and for an additional analysis yielding positive results). In other words, Senju et al. (2010) could not fully replicate the original findings by Southgate et al. (2007) themselves. In Study 1, we attempted to replicate the original findings in the original age group by analyzing the looking behavior of 24- and 25-month-olds separately and in the same way as in the original study. Neither for this roughly age-matched group of participants as a whole, nor for the subgroup of children who met the Southgate criterion was there any evidence for correct anticipatory looking. It is unlikely that the sample size used in the current paper affected the findings, as sample size calculations based on Chow, Wang, and Shao (2007, Chap. 4) using the effect sizes from the original Southgate et al. (2007) study suggest that only a minimum sample size of 13 subjects is required to show the original effect for first looks, whereas 20 children fell within the original age group in Study 1. In summary, we conclude that the original findings reported by Southgate et al. (2007) are not fully replicable with older children and probably neither with 24- to 25-month-olds.

Taken together, the present findings thus sketch the following picture. Robust replicability is only given for the FB1 condition of the Southgate/Senju paradigm. However, this task by itself is difficult to interpret, most fundamentally because in the FB1 condition the last ball location is identical to the belief-congruent location. This leaves room for alternative and more parsimonious explanations, for example, that the participants simply keep their gaze at the last ball location. The clearer and more stringent tasks (FB2 in the Southgate/Senju, paradigm and the task by Surian & Geraci), in contrast, could not be replicated. In summary, none of the paradigms could be replicated as a whole. Both paradigms can only fully be interpreted as measuring false belief processing when both conditions are replicated, as only the FB2 condition of the Southgate/Senju paradigm excludes low-level explanations for success in the FB1 condition and only the FB condition of the Surian & Geraci task reflects false belief processing, while the TB condition is subject to low-level explanations like ball tracking.

Therefore, the current paper shows that two major implicit Theory of Mind paradigms cannot reliably be replicated in independent research labs. Original findings from these paradigms have served as crucial empirical basis for ambitious theories. Should they turn out not to be robustly replicable, this would have far-reaching theoretical implications, shaking the empirical foundations of the ambitious theories in question.

Only in adults, there was some potential evidence for belief-sensitive gaze patterns: Firstly, in the FB2 condition of the Southgate/Senju task, if participants who failed the familiarization were included in the analysis and secondly, in the Surian & Geraci task, indicated by the location by condition interaction (although adults generally looked more to the location with the object than to the empty location across conditions, they tended to do so somewhat less strongly in the FB than in the TB condition). While this is far from a replication of the original results (such that participants looked more at the empty location than the location with the object in FB and showed the reverse pattern in TB), it does reflect the same kind of pattern as recently found in a series of studies by Schneider and colleagues (Schneider, Bayliss et al., 2012; Schneider, Lam et al., 2012; Schneider et al., 2013). However, it remains largely unclear, currently, how to interpret this whole pattern in the light of the general non-replication of original findings.

The interpretation of earlier findings is particularly difficult given the lack of any convergence or correlation in any of the populations tested here between different kinds of tasks – despite the fact that these were superficially very similar variations of anticipatory looking change-of-location FB tasks. If the same cognitive capacity, implicit ToM, was underlying both paradigms, the measures should converge. However, the lack of this convergence suggests that alternative, possibly low-level explanations may account for the findings. There is therefore no evidence that the paradigms measure the same underlying concept (Theory of Mind). The lack of correlation found here is in line with other recent findings (Yott & Poulin-Dubois, 2016), questioning the convergent validity of implicit ToM tasks in general.

In summary, the present findings constitute absence of evidence for the robustness and convergent validity of AL FB tasks. But this obviously does not imply evidence for the absence of the implicit ToM capacities under discussion. In general, the present findings leave open two broad possibilities. Firstly, previous published findings may constitute false positives. Implicit ToM capacities are indeed less robust than previously assumed, and in light of the missing convergent validity subject to alternative, more parsimonious explanations (e.g., Heyes, 2014). An extreme version would be that there is no such thing as an implicit ToM apart from standard explicit processes. From this perspective, two pieces of evidence from the current study might be taken as indicators that the tasks used here indeed tap explicit processes. First, the relation between explicit answers and anticipatory looking in Study 1 may suggest that both tapped, in fact, the same (explicit) processes. Second, it might be speculatively argued that the apparent U-curve regarding FB2 over the lifespan in Study 2b could simply reflect the inverted U-curve typically found in explicit ToM development (e.g., Bernstein, Thornton, & Sommerville, 2011; Henry, Phillips, Ruffman, & Bailey, 2013). Clearly, however, more systematic research is needed to substantiate such speculations.

The second broad possibility the present findings leave open is that implicit ToM may be a real and unitary cognitive capacity – yet, one that is fragile and thus difficult to tap empirically. For example, competence may be argued to translate into performance only under very specific conditions (e.g., concerning the design, timing of the stimuli etc.). Yet, this argument is problematic in the present context since the studies here used the exact same original stimuli and procedures, in Studies 2a and 2b, or highly similar ones, in Study 1 (compare Senju et al., 2009; Southgate et al., 2007; Surian & Geraci, 2012). Alternatively, children and adults may reveal their implicit ToM capacities only when the stimuli and tasks are, for example, engaging and ecologically valid enough. A

recent study with apes suggests that decades of negative findings in ToM tasks with non-human primates may have been due to a combination of excessive task demands and lacking ecological validity (Krupenye, Kano, Hirata, Call, & Tomasello, 2016).

5. Conclusion and future directions

The present findings show consistent lack of replicability and convergent validity of AL measures of implicit ToM across studies, methods and labs. Whether this suggests that implicit ToM is not as robust a phenomenon as previously assumed, or robust yet difficult to detect, one cannot tell from the present findings alone. More systematic future research, ideally involving multiple labs, will be needed to settle this question. In the meantime, though, the robustness of implicit ToM should be treated with more caution than indicated in the previous literature.

Acknowledgements

We would like to thank Luca Surian, Victoria Southgate, and Atsushi Senju for sharing their original stimuli with us. Thanks to the students and research assistants involved in the testing and data processing for this project, particularly Virginie Bihari, Jacqueline Ewert, Josefine Grzesko, Julia Henke, Kathrin Heyn, Josefin Johannsen, Jonas Koch, Annika Nöhring, Julia Schmuggerow, Friederike Schreiber, Lisa Wenzel, and Marieke Wübker, and to Wolfgang Bartels for technical assistance. We also would like to thank the volunteers and families for taking part in our research. This work was supported by the German Science Foundation grant RA 2155/4-1 (research unit “Crossing the borders: The interplay of language, cognition, and the brain in early human development”) and by grant KR-1213/3-2.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogdev.2017.09.001.

References

Baillargeon, R., Scott, R., He, Z., Sloane, S., Setoh, P., Jin, K., ... Bian, L. (2015). *Psychological and sociomoral reasoning in infancy. APA handbook of personality and social psychology, vol. 1*, 79–150.

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.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition, 21*(1), 37–46.

Bernstein, D. M., Thornton, W. L., & Sommerville, J. A. (2011). Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. *Experimental Aging Research, 37*(5), 481–502.

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.

Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. *Journal of Experimental Child Psychology, 131*, 94–103.

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

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language, 28*(2), 141–172.

Chow, S.-C., Wang, H., & Shao, J. (2007). *Large sample tests for proportions sample size calculations in clinical research.* New York: CRC press.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*(4), 377–395.

Fizke, E., Butterfill, S., Van de Loo, L., Reindl, E., & Rakoczy, H. (2014). *Signature limits in early theory of mind: Toddlers spontaneously take into account false beliefs about an object's location but not about its identity*. Department of Psychology, University of Göttingen.

Grosse Wiesmann, C., Steinbeis, N., Friederici, A., & Singer, T. (2017). *The developmental trajectory of an anticipatory looking false belief task from infancy to preschool-age – a longitudinal study*.

Henry, J. D., Phillips, L. H., Ruffman, T., & Bailey, P. E. (2013). A meta-analytic review of age differences in theory of mind. *Psychology and Aging, 28*(3), 826.

Heyes, C. (2014). Submentalizing I am not really reading your mind. *Perspectives on Psychological Science, 9*(2), 131–143.

Knudsen, B., & Liszkowski, U. (2012). 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy, 17*(6), 672–691.

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

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science, 354*(6308), 110–114.

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (submitted). Implicit Theory of Mind across the life span – anticipatory looking data. Data in Brief.

Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences, 9*(10), 459–462.

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science, 24*(3), 305–311.

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences, 16*(10), 519–525.

Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition, 129*(2), 410–417.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General, 141*(3), 433–438.

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

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.

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*(02), 353–360.

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.

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

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

13(6), 907–912.
Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30*(1), 30–44.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586.
Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655–684.
Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103–128.
Yott, J., & Poulin-Dubois, D. (2016). Are infants' Theory of Mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development, 17*(5), 683–698.