



## Testing the Role of Verbal Narration in Implicit Theory of Mind Tasks

Louisa Kulke & Hannes Rakoczy

To cite this article: Louisa Kulke & Hannes Rakoczy (2018): Testing the Role of Verbal Narration in Implicit Theory of Mind Tasks, Journal of Cognition and Development, DOI: [10.1080/15248372.2018.1544140](https://doi.org/10.1080/15248372.2018.1544140)

To link to this article: <https://doi.org/10.1080/15248372.2018.1544140>



Published online: 17 Nov 2018.



Submit your article to this journal [↗](#)



Article views: 14



View Crossmark data [↗](#)



# Testing the Role of Verbal Narration in Implicit Theory of Mind Tasks

Louisa Kulke and Hannes Rakoczy

Göttingen University, Germany


## ABSTRACT

Theory of Mind (ToM), the ability to attribute mental states to agents, has usually been measured with explicit verbal tasks and found to develop slowly during the preschool years. New implicit ToM measures have lately revolutionized the field by suggesting that ToM may be present much earlier in development. However, recent replication studies of implicit ToM present a complex pattern of failed, partial and successful attempts. The big challenge is to identify an underlying system to this pattern that can explain why some tasks replicate while others do not. The rationale of the present study was to address this challenge by investigating one potential factor that may explain patterns of (non-)replications of implicit measures, namely elements of verbal narration in anticipatory looking tasks. Sixty-seven 4- to 5-year-old children completed modified versions of two different anticipatory looking implicit false belief tasks which recently proved difficult to replicate. The main modification was that verbal narration was added to the original stimulus videos. Results revealed that original looking patterns could still not be replicated. There was no improvement in one task, while a slight improvement was observed in the other task. In conclusion, adding verbal narration does not necessarily improve the replicability of anticipatory ToM tasks, suggesting either that these measures might not be sufficiently sensitive to tap implicit ToM, or that other factors are crucial for successful replications.

Theory of Mind (ToM), the ability to ascribe mental states such as beliefs and desires to others and ourselves, is fundamental to our social life. Traditionally, it has been assumed that ToM develops in protracted ways over the preschool years, building on linguistic experience and slowly developing cognitive capacities such as executive function (Perner, 1991). This assumption has been based on the findings of hundreds of studies with explicit false belief tasks. In these tasks, children hear vignettes about an agent whose action they have to predict on the basis of her mistaken belief (Wimmer & Perner, 1983). Children pass these tasks from the age of 4 years (Perner, 1991; Wellman, Cross, & Watson, 2001; Wimmer & Perner, 1983). However, more recently, novel implicit ToM tasks that tap looking times and related non-explicit measure rather than explicit judgments have

**CONTACT** Louisa Kulke  [lkulke@uni-goettingen.de](mailto:lkulke@uni-goettingen.de)  Department of Affective Neuroscience and Psychophysiology, Göttingen University, Goßlerstraße 14, Goettingen 37073, Germany

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/hjcd](http://www.tandfonline.com/hjcd).

 Both experiments in this article earned open materials and open data for transparent practices. Materials and data are available at DOI [10.17605/OSF.IO/R7CXZ](https://doi.org/10.17605/OSF.IO/R7CXZ). The Preregistered+Analysis plan is available at <https://osf.io/fb8fj/>

© 2018 Taylor & Francis

revolutionized the field. These alternative measures have suggested that implicit and spontaneous ToM is present much earlier, sometimes as early as in the first year of life, and remains in constant, unconscious operation throughout the lifespan. Varieties of implicit ToM task use violation of expectation looking-time (Kovács, Téglás, & Endress, 2010; Onishi & Baillargeon, 2005), interactive behavioral measures (Buttelmann, Carpenter, & Tomasello, 2009; Knudsen & Liszkowski, 2012; Southgate, Chevallier, & Csibra, 2010) and anticipatory looking (e.g., Clements & Perner, 1994; Low & Watts, 2013; Schneider, Bayliss, Becker, & Dux, 2012; Senju, Southgate, White, & Frith, 2009; Southgate, Senju, & Csibra, 2007; Surian & Geraci, 2012). From a theoretical point of view, these findings have stirred debates between competing interpretations such as nativism (assuming that these tasks tap full-blown, potentially innate ToM; Baillargeon et al., 2015; Carruthers, 2013; Leslie, 2005; Scott, 2017; Wang & Leslie, 2016), skepticism (these tasks do not necessarily tap any form of ToM; Heyes (2014)) and two-systems theories (these tasks tap simple forms of ToM qualitatively different from later-developing full-blown forms; Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Low, Apperly, Butterfill, & Rakoczy, 2016).

These debates rest on the premise that the finding from implicit ToM tasks are robust, reliable and replicable. This empirical premise has recently been come into question. A growing body of systematic and independent attempts at replication have recently produced a complex and puzzling pattern of results. For implicit measures, there have now been several published non-replications, often with much larger samples sizes than in the original studies (for example, VoE: Dörrenberg, Liszkowski, and Rakoczy (2018); Powell, Hobbs, Bardis, Carey, and Saxe (2018); Yott and Poulin-Dubois (2016); interaction: Crivello and Poulin-Dubois (2018); Dörrenberg et al. (2018); Grosse Wiesmann, Friederici, Singer, and Steinbeis (2017); AL: Kulke, Reiß, Krist, and Rakoczy (2018); Kulke, von Duhn, Schneider, and Rakoczy (2018); Dörrenberg et al. (2018); Burnside, Ruel, Azar, and Poulin-Dubois (2018)). In addition, a survey to get at potential file-drawer problems has revealed more unpublished non-replications (Kulke & Rakoczy, 2018). The largest body of evidence comes from attempts to replicate AL ToM tasks. Several published studies report non-replications with several hundred participants across the lifespan, with much bigger sample sizes than the original studies, including direct replications with the exact same original stimulus material and procedures as well as conceptual replications.

The big challenge is now how to make sense of this complex pattern of findings, how to identify a system that underlies the pattern and explains why some tasks replicate and others do not. A closer look at the AL tasks that have and those that have not been successfully replicated (Kulke & Rakoczy, 2018) suggests the following. The AL measure that has been most often successfully replicated is the very first implicit ToM task ever used, the one by Clements and Perner (1994). This task is implicit in the sense that spontaneous and uninstructed anticipatory looking is measured. However, this measure is embedded in a verbally narrated standard FB task format (in which the experimenter tells and acts out a story: Max puts his chocolate in one location, the chocolate is then transferred to another location, then Max returns in search of his chocolate). In contrast, the AL tasks that have repeatedly been subject to non-replications by Southgate et al. (2007) and by Surian and Geraci (2012) are completely non-verbal. Participants see video events unfolding in which an agent acquires a false (or true) belief and participants can

predict the agent's action on this basis (revealed in their anticipatory looking to the location where the agent will go). May this difference between absence and presence of verbal narration explain the differences in replicability of the two types of tasks? Such a possibility seems not implausible in several respects: First of all, it is generally well known that language plays a crucial role in the development and execution of ToM (e.g., Milligan, Astington, & Dack, 2007; Newton & de Villiers, 2007). Second, we know from explicit ToM tasks that subtle differences in the verbal details of the dependent measures can make crucial differences (Wellman et al., 2001). Furthermore, some studies have shown slightly better implicit ToM performance in adults compared to children and elderly adults (Burnside et al., 2018; Kulke et al., 2018) and improvements due to learning in adults with autism but not children with autism (Schuwerk, Jarvers, Vuori, & Sodian, 2016; Schuwerk, Vuori, & Sodian, 2015), suggesting developmental changes, possibly related to executive functions. Thus, adding verbal narration may help younger children to overcome executive task demands that have masked their competence (evidence for such executive tasks demands in AL tests in both children and adults was recently found by Wang and Leslie (2016)).

The rationale of the present study was, therefore, to explore whether existing successful vs. unsuccessful replications of AL FB tasks may be due to differences in verbal narration. To this end, we followed up on our previous large-scale replication attempts of the two AL tasks by Southgate et al. (2007) and by Surian and Geraci (2012) (Kulke, Reiß, Krist, & Rakoczy, 2017; Kulke et al., 2018, 2018). An identical age range as in the study by Kulke et al. (2018) was tested (children between 4 and 5 years) to ensure comparability. At this age, children already show explicit Theory of Mind; therefore, any failure to pass the task should not be related to a lack of Theory of Mind, but rather to the task itself. Again, we used the original procedure and stimuli, yet the latter modified in such ways that they now contained verbal narration like the Clements and Perner (1994) task previously replicated successfully. If verbal narration is indeed crucial for implementing sensitive and reliable implicit ToM tasks, participants should now show belief-based anticipatory looking in each of the two modified tasks. Furthermore, performance in the two tasks should be correlated since both tap the same kind conceptual capacity.

## Method

### Participants

The current study was preregistered with the Open Science Framework (<https://osf.io/fb8fj/>). Sixty-seven neurotypical children between 4.5 and 5.5 years ( $M = 59.6$  months,  $SD = 3.34$ , 32 female) volunteered to participate. Age did not differ between the current study and the study by Kulke et al. (2018),  $t(129) = 0.73$ ,  $p = .465$ , and the gender ratio was comparable in both studies (current study: 48%, previous study: 50% female). They were recruited from local kindergartens and from the departmental child volunteer database. All participants spoke German as their primary language. For separate analyses of the Southgate/Senju and the Surian and Geraci task, the maximum number of participants who passed the inclusion criteria of the study was included ( $n = 48$  in the Southgate/Senju task; FB1:  $n = 26$ , FB2:  $n = 22$ , and  $n = 53$  in the Surian and Geraci task), while only those participants who passed the inclusion criteria based on both original studies were included in correlation analyses, leading

to a total participant number of 40 children. Therefore, the overall attrition rate was 40%, being in line with previous research (Southgate et al., 2007). The study was approved by the ethics committee of Göttingen University (Ethics code: 143a), carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) and consent was obtained from the children's parents.

### ***Materials and stimuli***

The original stimuli from Senju et al. (2009) and Surian and Geraci (2012), provided by the authors,<sup>1</sup> were used as a basis for the current study. In the Senju et al. (2009) task, a hand puppet and an actress wearing a visor interact in a setting with two boxes and an occluder with windows at the position of the boxes. In the first two familiarization trials, a chime sounds and the actress reaches through a window to retrieve a ball lying on either of the boxes. In the subsequent two familiarization trials, the hand puppet moves a ball into a box, the chime sounds and the actress reaches for the ball through the window close to the respective box. In the test trial, the puppet moves the ball into a box while the actress is watching and then moves the ball to the other box while the actress is watching (FB1) or turned away (FB2) and finally removes the ball from the scene without the actress watching. The actress turns back and the chime sounds again. Belief-based anticipatory looking would reveal itself in increased looking to the empty box in FB2 and the box containing the ball in FB1. 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, the ball enters the Y-shaped tunnel from the bottom end, reappears at one top end of the Y-shape (left or right) and hides in a box at that end. The triangle enters 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. In the true belief condition the ball moves from the box at one end of the tunnel to the box at the other end, witnessed by the triangle; then the triangle disappears. In the false belief condition the triangle leaves the scene before the ball moves between boxes. The triangle reappears at the scene, enters the tunnel and exits the tunnel at the belief-congruent location (i.e., the true location of the ball in the TB and the opposite location in the FB condition).

The main difference of the current study to the original tasks was that a German narration was added to the original videos. A male, native German narrator described in a calm voice what was happening in the scene. Although he described the events in the scene, no information regarding the false belief of the actress was provided to avoid an explicit statement in the previously implicit task. Before the cue for gaze recording (chime sound in the Southgate/Senju task or triangle entering the tunnel in the Surian and Geraci task), the narrator stated “I wonder where she will reach for the ball” (Southgate/Senju) or “I wonder where the triangle will exit the tunnel” (Surian & Geraci, 2012). The full audio recordings used for the current study are provided at the OSF: DOI 10.17605/OSF.IO/R7CXZ.

An SMI REDm 250 mobile remote eye tracker recorded participants' gaze at a rate of 60 Hz. Participants completed a standard 5-point calibration and validation routine before

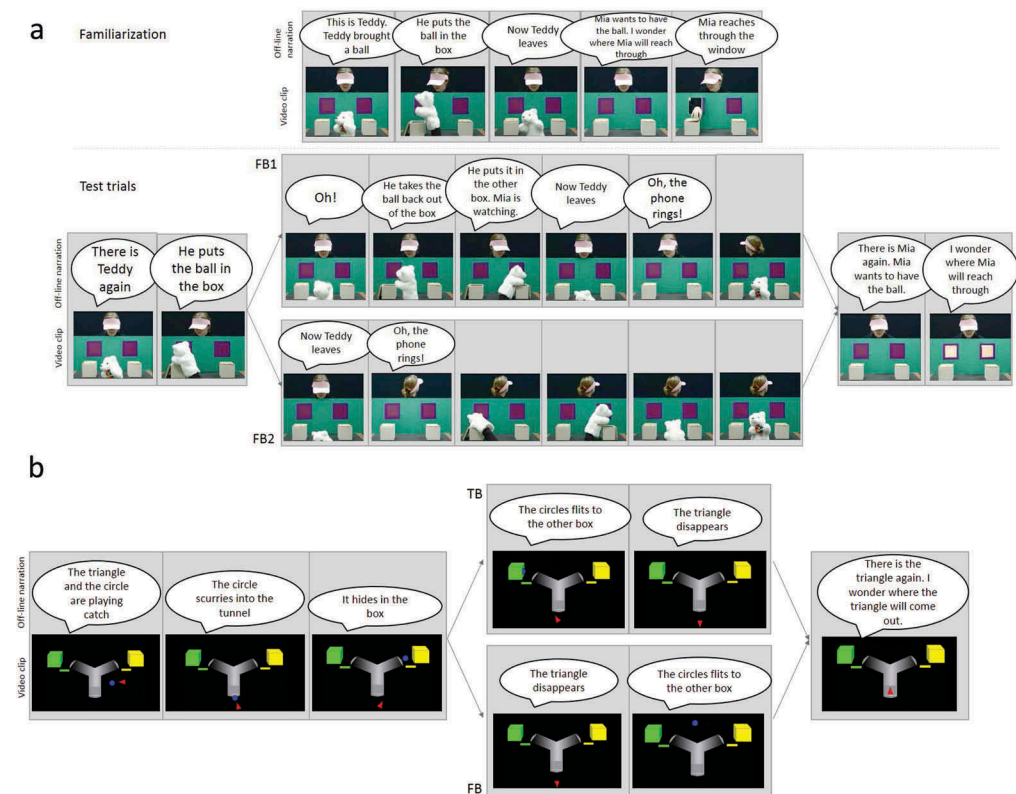
---

<sup>1</sup>As of the Surian and Geraci task only the original FB and TB videos for one object location were available from the authors, the mirrored videos from Kulke et al. (2018) were used to counterbalance sides as in the original study.

each task. Movies were presented on a laptop computer (HP ZBook with a Windows 10 operating system) controlled via the SMI Experiment Center (version 3.6.53) and presented on a 15.6" LCD display (1920 × 1080 pixel). Offline analysis of gaze information was conducted using BeGaze Software (version 3.6.52) for visualization and exported to IBM SPSS Statistics (version 25) for further analysis.

### Design

All participants completed the verbal versions of the Southgate/Senju and the Surian and Geraci task (see Figure 1a and b for schematic displays of the tasks and narrations). In a mixed design, conditions were manipulated as in the original studies, with the Southgate/Senju belief condition (FB1/FB2) being counterbalanced between participants and the Surian and Geraci conditions (TB/FB) being compared within participants. The original outcome measures were used in the current study. For the Southgate/Senju task, we report first saccades and a differential looking score (DLS), describing relative looking to the belief-congruent location, computed as the time spent looking at the correct area of interest (AOI) minus the time spent looking at the incorrect AOI, divided by the time spent looking at both AOIs. For the Surian and Geraci task, we report first saccades and



**Figure 1.** (a) Schematic display of one familiarization and the two test trials (FB1 and FB2) of the Southgate/Senju task with verbal narration. (b) Schematic display of the FB and the TB trial in the Surian and Geraci task with verbal narration.

proportion of looking times. In order to analyze correlations of the two tasks, a DLS was computed for both based on Kulke et al. (2018) to account for differences due to task specifics (AOI size and duration). Furthermore, findings in the current verbal task were compared between participants to the study by Kulke et al. (2018), in which identical stimuli and procedures were used but without verbal instructions, the data of which has been published by Kulke et al. (2017). Only children of the same age range (4.5 to 5.5 years) were included in the comparison. There were approximately 6 months between data recording for the current study and the study by Kulke et al. (2017) and no children participated in both studies.

### **Procedure**

Participants were tested in a child-friendly environment. A laptop with the remote eye tracker attached to it was positioned in front of the participant and the distance and angle of the eye-tracker were adjusted to the height of the participant to improve the eye-tracking signal. Participants completed a 5-point calibration and validation routine before each task. They saw both tasks in a counterbalanced order. The Southgate/Senju condition (FB1 or FB2) was selected based on a predetermined randomization list. Participants completed a TB and an FB condition in the Surian and Geraci task, the order of which were randomized. The final screen side (left, right) of the ball was counterbalanced. The testing procedure took approximately 10 minutes.

### **Analysis**

Based on the original studies, participants were excluded from the analysis if they failed the inclusion criteria of the task to be analyzed. To be included in the Senju et al. (2009) task, participants needed to look longer towards the correct than the incorrect box in the last familiarization trial. For the Surian and Geraci (2012) task, participants needed to look longer towards the correct than towards the incorrect box in at least one of the two familiarization trials. To be included in the correlation analysis of the two tasks, participants needed to fulfill both criteria. If participants needed to be excluded due to failing the original inclusion criteria, additional participants were tested until the predetermined sample size was reached.

The same outcome measures as in the original studies were used and the analyses were identical to the paper by Kulke et al. (2018). For the videos based on Senju et al. (2009), a differential looking score (DLS) was calculated. AOIs were defined as the area covering the left and right window and gaze was recorded in a 5 s time interval starting with the illumination of the windows. For the videos based on Surian and Geraci (2012), looking times (depicted as proportion of looking) in the time interval between the triangle entering the tunnel and its reappearance were measured in two AOIs including the outer edge of the left and right box up to the respective tunnel exit. Furthermore, first saccades to the correct and incorrect location were analyzed. Note that the number of saccades is smaller than the sample size, as several participants did not make any first saccade within the defined time window or were already looking at either of the AOIs at the beginning of the predefined time interval and could therefore not make a saccade to this direction. For the latter case we report an additional measure, “First look”, in the open



data file (DOI 10.17605/OSF.IO/R7CXZ) which included both those trials in which participants initially already fixated on the AOI and those in which participants made a saccade in the direction. Correlation analyses of the two tasks used the DLS, as it accounts for the different time windows and AOIs used in the original tasks. Full datasets are provided in Supplement B. The results from the current study using explicit narration were compared to results from the previous study using the original implicit tasks (Kulke et al., 2018). In order to utilize the maximum power, each analysis and figure included the maximum number of participants that could be included in the analysis. Therefore, all 48 participants who passed the Southgate/Senju familiarization (FB1:  $n = 26$ , FB2:  $n = 22$ ) were included in the Southgate/Senju analysis and all 53 who passed the Surian and Geraci familiarization were included in that analysis, respectively. For the correlation of both paradigms, participants needed to pass both familiarizations.

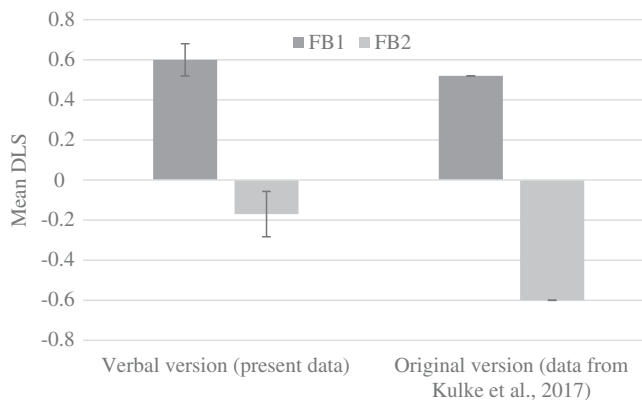
## Results

### Southgate/Senju task

#### DLS

Participants who passed the inclusion criteria based on Senju et al. (2009) were included in the analysis. Mean DLS as a function of condition is displayed in Figure 2. DLS was significantly positive in the FB1 condition, indicating significantly more looking to the belief-congruent than the belief-incongruent location ( $M = 0.60$ ,  $SD = 0.41$ , 95% CI [0.44, 0.77]),  $t(25) = 7.583$ ,  $p < .001$ ,  $d = 3.03$ , but not in the FB2 condition ( $M = -0.17$ ,  $SD = 0.53$ , 95% CI [-0.40, 0.07]),  $t(21) = -1.466$ ,  $p = .157$ ,  $d = -0.64$ ,  $BF = 0.567$ .<sup>2</sup>

DLS did not differ between the explicit ( $M = 0.60$ ,  $SD = 0.41$ , 95% CI = [0.44, 0.77]) and the implicit version ( $M = 0.52$ ,  $SD = 0.44$ , 95% CI = [0.32, 0.73]) of the study in the FB1



**Figure 2.** Mean DLS in the Senju/Southgate task in the verbal (left) and original (right) FB1 and FB2 conditions.

<sup>2</sup>Note that in the paper by Kulke et al. (2018) DLS in the FB1 condition was significantly positive,  $M = 0.526$ ,  $SD = 0.441$ ,  $t(19) = 5.341$ ,  $p < 0.001$ ,  $d = 2.451$ , and in the FB2 condition significantly negative,  $M = -0.598$ ,  $SD = 0.405$ ,  $t(19) = -6.597$ ,  $p < 0.001$ ,  $d = -3.027$ , with both conditions differing significantly from another,  $t(38) = 8.397$ ,  $p < 0.001$ ,  $d = 2.724$ .



condition,  $t(44) = 0.622$ ,  $p = .537$ ,  $d = 0.19$ ,  $BF = 0.344$ , but was significantly larger in the explicit ( $M = -0.17$ ,  $SD = 0.53$ , 95% CI =  $[-0.40, 0.07]$ ) than in the implicit ( $M = -0.60$ ,  $SD = 0.41$ , 95% CI =  $[-0.79, -0.41]$ ) version in the FB2 condition,  $t(40) = 2.93$ ,  $p = .006$ ,  $d = 0.93$ .

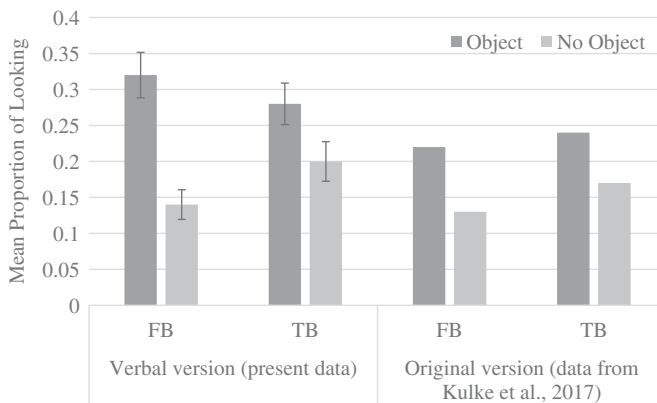
### First saccade

A binomial test showed no significant difference between first saccades to the correct ( $n = 8$ ) and incorrect ( $n = 4$ ) location,  $p = .388$ ,  $BF = 0.902$ . There was no significant difference between saccades to the correct ( $n = 4$ ) and incorrect ( $n = 0$ ) location in the FB1 condition,  $p = .125$ , but note  $BF = 1.900$ , or in the FB2 condition (correct:  $n = 4$ , incorrect:  $n = 4$ ),  $p = 1$ ,  $BF = 0.650$ . Fisher's exact test showed a significant difference between the current explicit study (correct:  $n = 8$ , incorrect:  $n = 4$ ) and the previous implicit study by Kulke et al. (2018) (correct:  $n = 3$ , incorrect:  $n = 11$ ),  $p = .045$ ,  $d = 1.10$ ,  $BF = 5.938$ .

### Surian and Geraci task

#### Proportion of looking

Mean proportions of looking as a function of condition are displayed in Figure 3. An ANOVA computing the effects of object location (empty or object) and belief (true or false) on proportion of looking including all participants who passed the Surian and Geraci inclusion criteria showed a significant effect of object location,  $F(1, 52) = 13.41$ ,  $p < .001$ ,  $\eta_p^2 = .205$ , but no significant interaction,  $F(1, 52) = 2.65$ ,  $p = .110$ ,  $\eta_p^2 = .048$ ,  $BF = 1.25$ , and no main effect of belief,  $F(1, 52) = 0.55$ ,  $p = .463$ ,  $\eta_p^2 = .010$ ,  $BF = 0.16$ . Pre-registered follow-up t-tests showed that participants looked longer at the object location ( $M = 0.32$ ,  $SD = 0.23$ , 95% CI =  $[0.26, 0.39]$ ) than the no object location ( $M = 0.14$ ,  $SD = 0.15$ , 95% CI =  $[0.10, 0.18]$ ) in the FB,  $t(52) = -4.04$ ,  $p < .001$ ,  $d = 1.11$ , but not significantly longer at the object location ( $M = 0.28$ ,  $SD = 0.21$ , 95% CI =  $[0.23, 0.34]$ ) than the no object location ( $M = 0.20$ ,  $SD = 0.20$ , 95% CI =  $[0.15, 0.26]$ ) in the TB condition,  $t(52) = 1.60$ ,  $p = .116$ ,  $d = 0.44$ ,  $BF = 0.491$ . The difference between TB and FB condition was not significant at the empty location,  $t(52) = -1.92$ ,



**Figure 3.** Proportion of looking to the object and no object location in the FB and TB condition of the Surian and Geraci task in the current and the original study.

$p = .061$ ,  $d = 0.53$ ,  $BF = 0.818$ , or the object location,  $t(52) = -1.04$ ,  $p = .304$ ,  $d = -0.40$ ,  $BF = 0.249$ .

Looking time in the false belief condition was compared with the corresponding original stimulus condition using an ANOVA including location (empty or object location) and stimulus type (original or narration) as well as the interaction of both variables. There was a significant main effect of object location,  $F(1, 110) = 22.67$ ,  $p < .001$ ,  $\eta_p^2 = 0.171$ , and of study,  $F(1, 110) = 7.61$ ,  $p = .007$ ,  $\eta_p^2 = 0.065$ , but no interaction,  $F(1, 110) = 2.55$ ,  $p = .113$ ,  $\eta_p^2 = 0.023$ ,  $BF = 0.99$ . Pre-registered follow up tests showed that participants looked longer at the ball location ( $M = 0.22$ ,  $SD = 0.19$ , 95% CI = [0.17, 0.27]) than the empty location ( $M = 0.13$ ,  $SD = 0.15$ , 95% CI = [0.09, 0.17]) in the previous study,  $t(58) = -2.51$ ,  $p = .015$ ,  $d = -0.65$ , and also in the current study,  $t(52) = -4.04$ ,  $p < .001$ ,  $d = 1.11$ . Separate analyses for the locations showed that there were no differences in looking time to the empty location between the previous implicit study by Kulke et al. (2018) ( $M = 0.13$ ,  $SD = 0.15$ , 95% CI = [0.09, 0.17]) and the current study ( $M = 0.14$ ,  $SD = 0.15$ , 95% CI = [0.10, 0.18]),  $t(110) = -0.30$ ,  $p = .764$ ,  $d = -0.06$ ,  $BF = 0.21$ , but a significant difference at the object location,  $t(110) = -2.52$ ,  $p = .013$ ,  $d = 0.48$ , with more looking to the object location in the current study ( $M = 0.32$ ,  $SD = 0.23$ , 95% CI = [0.26, 0.39]) than in the previous study ( $M = 0.22$ ,  $SD = 0.19$ , 95% CI = [0.17, 0.27]).

To exclude the possibility of a change in the overall looking pattern in all conditions of the verbal compared to the original version of the Surian and Geraci task, an exploratory (non-preregistered) analysis was performed that included both the TB and the FB condition of the original and the narrated version. A general linear model was used to compute the effects of belief (TB or FB), location (empty or object location) and stimulus type (original or narration) and their interactions on proportional looking time. There was no significant three-way interaction of belief, location and stimulus type,  $F(1, 110) = 0.96$ ,  $p = .329$ ,  $\eta_p^2 = 0.01$ , suggesting that the overall looking pattern between conditions did not change when narration was added to the stimuli.

### First saccade

In the FB condition, a binomial test showed that the number of correct ( $n = 23$ ) and incorrect ( $n = 24$ ) saccades did not significantly differ,  $p = 1$ ,  $BF = 0.343$ . In the TB condition, the number of correct ( $n = 28$ ) and incorrect ( $n = 17$ ) saccades did not significantly differ,  $p = .135$ ,  $BF = 1.134$ .

Fisher's exact tests showed no significant difference of the effect in the false belief condition between the condition recorded here and the previously recorded condition (using original stimuli, correct:  $n = 24$ , incorrect:  $n = 29$ ),  $p = .841$ ,  $d = 0.08$ ,  $BF = 0.262$ .

### Correlations between the tasks

There was a significant correlation of the DLS in the Southgate/Senju task ( $M = 0.23$ ,  $SD = 0.62$ , 95% CI = [0.03, 0.43]) with the DLS in the Surian and Geraci FB condition ( $M = -0.31$ ,  $SD = 0.66$ , 95% CI = [-0.52, -0.10]),  $r(40) = .415$ ,  $p = .008$ ,  $d = 0.91$ ,  $BF = 6.050$ , but no significant correlation of the Southgate/Senju task with the Surian and Geraci TB condition ( $M = 0.07$ ,  $SD = 0.69$ , 95% CI = [-0.15, 0.29]),  $r(40) = .112$ ,  $p = .490$ ,  $d = 0.23$ ,  $BF = 0.375$ , or of the Surian and Geraci TB and FB condition,  $r(40) = -.026$ ,  $p = .873$ ,  $d = -0.54$ ,  $BF = 0.312$ .

## Discussion

The current study aimed to investigate the effect of verbal narration on anticipatory looking behavior in implicit false belief tasks. To this end, the original stimuli of two completely non-verbal AL tasks that previously could not be replicated were modified by adding verbal narration. The main findings were the following: First, in the new version of the Southgate/Senju task, DLS patterns indicate that participants looked significantly more often to the belief-congruent location in the FB1, but not in the FB2 condition. Second, a comparison of this modified condition including narration with findings from a previous replication using original stimuli only (Kulke et al., 2017, 2018) showed no difference between both studies in the FB1 condition. However, DLS in the FB2 condition was significantly less negative when the narration was introduced. Third, in the narratively modified Surian and Geraci task in the present study, DLS patterns showed that, in general, participants tended to look at the object location rather than the empty location in both the TB and FB condition. Fourth, direct comparison with the Kulke et al. (2018) study showed that participants looked even more at the object location in the current study using a narration than in the previous study. Fifth, first saccades were not significantly more often directed to the belief-congruent than to the alternative location in any condition of the current and the previous study in any task. As first saccades are a binary measure, they may be less sensitive to detect potential differences. Sixth, correlation analyses show a significant correlation of the Surian and Geraci FB condition with the Southgate/Senju DLS.

In summary, with regard to the original findings by Southgate et al. (2007), only the FB1 condition of the Southgate/Senju task could be replicated. In relation to identical replication attempts without narration in Kulke et al. (2018), the narration led to slightly improved DLS patterns in the FB2 condition and a novel correlation between DLS scores in both paradigms that could not be detected in previous research (Kulke et al., 2018, 2018). Note that although the findings are now more similar to the original study by Southgate et al. (2007) compared to the findings by Kulke et al. (2017), the study still did not fully replicate it.

There are several possible explanations for these effects. Firstly, the novel measure might be more sensitive, due to the narrative structure enhancing spontaneous perspective-taking (see Rubio-Fernández & Geurts, 2013). Secondly, the narrative additions might actually transform the task into an explicit task. The outcome measure itself (gaze) was implicit, as it did not require explicit statements regarding belief tracking. Therefore, only the procedure but not the dependent measure could have become explicit due to the narration. This is impossible to test in the present study. Although the children tested here were old enough for solving explicit tasks, they cannot be asked explicitly like adults in debriefing protocols (see Schneider, Lam, Bayliss, & Dux, 2012). In the verbal narration in the current study, no information regarding the false belief of the actress was provided to avoid an explicit statement in the previously implicit task. This method avoids explicit belief statements and is therefore more conservative, potentially weakening effects of explicitness. Future research could investigate the possibility of explicit task solutions by testing younger children who do not solve standard explicit tasks yet and therefore should be less affected by explicit hints or older children and adults who can be asked about explicit ToM processing in a verbal debriefing procedure. Although the original tasks have been used across the lifespan (see e.g., Senju et al., 2009; Southgate et al., 2007), it would

further be interesting to study the effect of narration in other age groups than the 4- to 5-year-olds investigated in the current study.

Interestingly, the current study found a correlation between the Southgate/Senju and the Surian and Geraci FB conditions. This is in contrast with previous research that found no relations between different AL paradigms (Kulke et al., 2018, 2018), no relations between different types of paradigms (Poulin-Dubois & Yott, 2018) or randomly scattered correlations (Dörrenberg et al., 2018) between paradigms. At the current state of inquiry, we can only speculate about potential reasons for these diverging findings. One possibility is that the added narration in the current task has indeed increased the similarity between tasks. This may either be related to increased belief processing measured here, or due to the narration increasing another, confounding factor, for example cognitive load or verbal skills.

Crucially, however, even after adding verbal narration, original findings could still not reliably be replicated, particularly in the Surian and Geraci task—much in line with previous large-scale failed replication attempts of anticipatory looking false belief tasks (Kulke et al., 2018). A lack of looking patterns in line with false belief processing in the Surian and Geraci (2012) task may be because the effect is fragile and hence difficult to measure. Recent preliminary findings point out that the time-course of gaze may play a crucial role, with the original time window possibly not representing the ideal time-frame for observing belief processing effects (Rubio-Fernandez, 2018). Even more, in contrast to the original, non-verbal task replicated by Kulke et al. (2018), the current study showed a marginal decrease in the false belief congruent gaze pattern. Possibly, the additional narration may have increased cognitive load and thus impaired performance. Cognitive load, in form of object interference, has already been shown to interfere with belief tracking (Wang & Leslie, 2016). The more difficult Surian and Geraci task that involves inhibition of the true ball location, as well as belief tracking in the false belief condition may thus have been more negatively affected by the narration. To speculate, narration may improve performance in simple tasks, like the Southgate et al. FB2 condition in the current study, while it impairs performance in more difficult tasks, possibly when a limit of processing capacity is met. Systematic future research is needed to put this speculation to test.

Note that the current study tested children between 4 and 5 years, an age at which children already show explicit forms of Theory of Mind. They may therefore find the task boring or too slow. However, the Southgate et al. paradigm has been used in many age groups, including infants (Southgate et al., 2007), children between 6 and 9 years (Senju et al., 2010) and adults (Senju et al., 2009) by the original authors, suggesting that it should be equally suitable across the life span. Testing children between 4 and 5 years in the current study (an age at which children show explicit Theory of Mind) ensured that a failure to pass the task should not be related to a lack of Theory of Mind, but rather to the task itself. It furthermore ensured that an identical age range was tested as in the study by Kulke et al. (2018). Although the comparison of the data from Kulke et al. (2018) and the current study should be treated with caution, children were sampled from the same population (same town and age group). Therefore, as the Southgate/Senju paradigm uses a between-participant design, the different samples could only have effects on the Surian and Geraci task.

In conclusion, the present study indicates a very modest effect of adding narration to AL FB tasks such that narration slightly improves false belief processing as measured through anticipatory looking. However, this effect was very restricted and only present for one, in itself ambiguous, condition in one out of the two tasks; the other, less ambiguous, conditions and tasks did not reveal belief-processing even with narration. In the end, anticipatory looking FB tasks may simply not be sensitive and reliable measures of spontaneous Theory of Mind; or if some versions of these tasks turn out to be, some other factor than verbal narration may be crucial for making them robust, reliable and replicable.

## Highlights

- The effect of narration on implicit Theory of Mind was investigated
- Partial improvements occurred in an anticipatory looking task involving object removal
- No improvements occurred in a task involving object transfer
- Original findings of implicit Theory of Mind could not be replicated
- Anticipatory looking tasks may not be sufficiently sensitive to implicit Theory of Mind

## Acknowledgments

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 stimulus creation, testing and data processing for this project, particularly Marieke Wübker, Maj-Lis Ute Klüh, Jana Susanne Fabian and Simon Niedermeier, and to the children who volunteered to participate in this study and their parents.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## 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. doi:10.1037/a0016923
- 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*, 1, 79–150.
- Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*, 46, 4–11. doi:10.1016/j.cogdev.2017.08.006
- Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337–342. doi:10.1016/j.cognition.2009.05.006
- Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, 28(5), 606–637. doi:10.1111/mila.12036

- Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2), 141–172. doi:10.1111/mila.12014
- Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, 9(4), 377–395. doi:10.1016/0885-2014(94)90012-4
- Crivello, C., & Poulin-Dubois, D. (2018). Infants' false belief understanding: A non-replication of the helping task. *Cognitive Development*, 46, 51–57. doi:10.1016/j.cogdev.2017.10.003
- Dörrenberg, S., Liszkowski, U., & Rakoczy, H. (2018). How (not) to measure infant Theory of Mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*. doi:10.1016/j.cogdev.2018.01.001
- Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental Science*, 20(5), e12445.
- Heyes, C. (2014). Submentalizing I am not really reading your mind. *Perspectives on Psychological Science*, 9(2), 131–143. doi:10.1177/1745691613518076
- 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. doi:10.1111/inf.2012.17.issue-6
- Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830–1834. doi:10.1126/science.1190792
- Kulke, L., & Rakoczy, H. (2018). Implicit Theory of Mind—An overview of current replications and non-replications. *Data in Brief*, 16, 101–104. doi:10.1016/j.dib.2017.11.016
- Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2017). Implicit Theory of Mind across the life span – Anticipatory looking data. *Data in Brief*, 15(SupplementC), 712–719. doi:10.1016/j.dib.2017.10.021
- Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, 46, 97–111. doi:10.1016/j.cogdev.2017.09.001
- Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*. doi:10.1177/0956797617747090
- Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences*, 9(10), 459–462. doi:10.1016/j.tics.2005.08.002
- 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. doi:10.1111/cdep.2016.10.issue-3
- Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science*, 24(3), 305–311. doi:10.1177/0956797612451469
- Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, 78(2), 622–646. doi:10.1111/j.1467-8624.2007.01018.x
- Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: Adults fail nonverbal false-belief reasoning. *Psychological Science*, 18(7), 574–579. doi:10.1111/j.1467-9280.2007.01942.x
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255–258. doi:10.1126/science.1107621
- Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA, US: The MIT Press.
- Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science*, 21(4), e12600.
- Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46, 40–50.
- Rubio-Fernandez, P. (2018). *Spatial indexing in false-belief tasks: A continuous eye-tracking study*. Paper presented at the Budepest CEU Conference on Cognitive Development, Budapest, Hungary.
- Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, 24(1), 27–33. doi:10.1177/0956797612447819



- 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. doi:10.1037/a0025458
- Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, 23(8), 842–847. doi:10.1177/0956797612439070
- Schuwerk, T., Jarvers, I., Vuori, M., & Sodian, B. (2016). Implicit mentalizing persists beyond early childhood and is profoundly impaired in children with autism spectrum condition. *Frontiers in Psychology*, 7, 1696. doi:10.3389/fpsyg.2016.01696
- Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: The impact of experience. *Autism*, 19(4), 459–468. doi:10.1177/1362361314526004
- Scott, R. M. (2017). The developmental origins of false-belief understanding. *Current Directions in Psychological Science*, 26(1), 68–74. doi:10.1177/0963721416673174
- 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. doi:10.1017/S0954579410000106
- 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. doi:10.1126/science.1176170
- 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. doi:10.1111/j.1467-7687.2009.00946.x
- Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, 18(7), 587–592. doi:10.1111/j.1467-9280.2007.01944.x
- Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, 30(1), 30–44. doi:10.1111/j.2044-835X.2011.02046.x
- Wang, L., & Leslie, A. M. (2016). Is implicit theory of mind the 'Real Deal'? The own-belief/true-belief default in adults and young preschoolers. *Mind & Language*, 31(2), 147–176. doi:10.1111/mila.12099
- 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. doi:10.1080/15248372.2015.1086771