Cognitive Development 46 (2018) 58–68

Contents lists available at ScienceDirect

Cognitive Development

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

Longitudinal evidence for 4-year-olds' but not 2- and 3-year-olds' false belief-related action anticipation

Charlotte Grosse Wiesmann[a,b,c,*], Angela D. Friederici[a], Denisse Disla[d], Nikolaus Steinbeis[b,e,1], Tania Singer[b,1]

[a] Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[b] Department of Social Neuroscience, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[c] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany
[d] Freie Universität Berlin, Germany
[e] Developmental and Educational Psychology, Institute of Psychology, Leiden University, the Netherlands

ARTICLE INFO

Keywords:
Theory of mind
False belief
Anticipatory looking
Longitudinal study
Replication study
Preschool age

ABSTRACT

Recently, infants younger than 2 years have been shown to display correct expectations of the actions of an agent with a false belief. The developmental trajectory of these early-developing abilities and their robustness, however, remain a matter of debate. Here, we tested children longitudinally from 2 to 4 years of age with an established anticipatory looking false belief task, and found a significant developmental change between the ages of 3 and 4 years. Children anticipated correctly only by the age of 4 years, and performed at chance at the ages of 2 and 3 years. Moreover, we found correct anticipation only when the agent falsely believed an object to be in its last rather than a previous location. These findings point towards the fragility of early belief-related action anticipation before the age of 4 years, when children start passing traditional false belief tasks.

1. Introduction

A milestone of Theory of Mind (ToM) development has long been assumed to occur around the age of 4 years, when children start passing traditional false belief tasks (e.g., Wellman, Cross, & Watson, 2001). In recent years, however, novel implicit false belief paradigms have shown that infants younger than 2 years of age already display correct expectations of how an agent with a false belief will act (e.g., Baillargeon, Scott, & He, 2010; Onishi & Baillargeon, 2005; Sodian, 2016; Southgate, Senju, & Csibra, 2007). These findings have caused an overhaul of our understanding of ToM abilities in infants, and have triggered one of the most controversial debates of developmental psychology: Why do children consistently fail traditional false belief tasks until the age of 4 years, if infants already understand others' false beliefs? The reasons for this discrepancy has been debated intensely. While some authors have argued for a continuity of the abilities measured by implicit and explicit false belief tasks (e.g., Baillargeon et al., 2010; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012; Sodian, 2016), others have argued that different processes with different developmental trajectories might underlie the tasks (Apperly & Butterfill, 2009; Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2016; Grosse Wiesmann, Schreiber, Singer, Steinbeis, & Friederici, 2017; Heyes, 2014; Perner & Roessler, 2012; Ruffman, 2014). The developmental trajectory of the early-developing abilities between infancy and preschool age, when children start passing the

* Corresponding author at: Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstr. 1a, 04103 Leipzig, Germany.
E-mail address: wiesmann@cbs.mpg.de (C. Grosse Wiesmann).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.cogdev.2017.08.007
Received 15 February 2017; Received in revised form 29 July 2017; Accepted 20 August 2017
Available online 07 September 2017

traditional false belief tasks, has not been studied to date, and could contribute to this debate. This question is all the more important, given that recent findings point to a certain fragility of the implicit false belief tasks (He, Bolz, & Baillargeon, 2012; Helming, Strickland, & Jacob, 2014; Kulke & Rakoczy, 2017).

So far, implicit false belief tasks have mainly been tested with infants and toddlers between the first and third year of life, but older children have rarely been tested on these paradigms. It remains unclear how performance on these tasks develops between toddlerhood at 2 years and preschool age around 3 and 4 years. Only a few studies have shown anticipatory looking in preschoolers (Clements & Perner, 1994; Grosse Wiesmann et al., 2016; Low, 2010), and, to our knowledge, no study has investigated how the same task develops from infancy until preschool age. Understanding the developmental trajectory of the early abilities from 2 years of age until 4, when children start passing the traditional false belief tasks, however, would inform about the nature of these abilities and their relation to the later-developing traditional false belief tasks. To address this question, we tested children longitudinally from 2 to 4 years with an established anticipatory looking false belief task. This allowed us to answer whether performance on the implicit false belief tasks remains stable once infants start passing these tests, or whether performance improves during early childhood, paralleling the performance breakthrough in the traditional explicit false belief tasks.

Moreover, in light of the recent debate about the fragility of implicit false belief tasks (He et al., 2012; Helming et al., 2014; Kulke & Rakoczy, 2017), the study allowed us to test how robust belief-related anticipatory looking is in 2-year-olds and in preschoolers. Current theoretical debates rely on the assumption that infants reliably show belief-congruent expectations (e.g., Apperly & Butterfill, 2009; Baillargeon et al., 2010; Heyes, 2014; Perner & Roessler, 2012; Ruffman, 2014). It is therefore crucial to know that these findings are robust, and to understand possible limitations of these findings. Potential fragility of performance in toddlers but not in older children that pass the traditional explicit false belief tasks could help bridge the gap between infants' early success in the implicit tasks and preschoolers' failure in the traditional tasks until the age of 4 years. Thus, both the developmental trajectory of implicit false belief tasks between infancy and preschool age and the robustness of performance on these tasks at different ages can contribute to solving the puzzle why these tasks are passed several years earlier than the traditional false belief tasks.

Arguably the most stringent support for infants' action anticipation based on attribution of false beliefs came from a particularly well-controlled anticipatory looking task by Southgate et al. (2007). In this study, 25-month-old children were shown to anticipate correctly where an agent who falsely believed an object to be in one of two empty boxes would search for the object. The authors constructed two false belief conditions (FB1 and FB2) that were orthogonal with respect to simpler non-belief-based strategies, such as, gazing at the first or last box the object had been in or at the last box the agent had attended to. Correct anticipation in both false belief conditions therefore ensured that children passed the test based on belief attribution and not due to these simpler associations. Similar anticipatory looking paradigms have been used at different ages between 18 months and 3 years (Clements & Perner, 1994; Gliga, Senju, Pettinato, Charman, & Johnson, 2014; Grosse Wiesmann et al., 2016; Low, 2010; Meristo et al., 2012; Senju, Southgate, Snape, Leonard, & Csibra, 2011; Surian & Geraci, 2012; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012; Wang & Leslie, 2016). In the present longitudinal study, we therefore used this anticipatory looking false belief task (Southgate et al., 2007). To increase the sensitivity of the measure, and to make sure that an individual child did not pass the test because of a simpler strategy, we presented every child with both false belief trials, one trial of each of the two original conditions FB1 and FB2. This also allowed us to compare performance between the two conditions within subjects. Analyzing only the first trial, in turn, allowed us to compare performance in our study to the original study where every child performed only a single false belief trial, condition FB1 or FB2 respectively (Southgate et al., 2007).

By testing children longitudinally from 2 years until 4 years of age, we aimed at addressing the following questions: (1) How does belief-related anticipatory looking develop beyond the age of 2 years? Does performance remain comparable throughout preschool age, or is there an improvement on the implicit false belief tasks once children start passing the traditional explicit false belief tasks? (2) How robust is performance on implicit false belief tasks in 2-year-olds, and does robustness change with age? Is performance robust across different task conditions?

From the previous literature, we hypothesized that children would perform above chance at the age of 2 years. Further, we expected to find no difference in performance between the two different false belief conditions FB1 and FB2 based on the original study (Southgate et al., 2007). Concerning the developmental trajectory of performance between the ages of 2 and 4 years, the pattern was less clear from previous literature. Based on similar levels of performance in adults compared to previous infant studies (Senju, Southgate, White, & Frith, 2009), we expected to find either stable performance between the ages of 2 and 4 years or an increase in performance that we might be able to detect due to the higher sensitivity in our within-subject longitudinal design.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Eye-tracking data was acquired in three consecutive years, starting at the age of 2 years. In the first year, data was recorded from 52 toddlers (mean age: M = 2.55 years, SD = 0.32, range: 2.08–3.20 years), from which 6 data sets had to be excluded because the recording had to be interrupted before the first test trial (N = 3), data quality was insufficient to be analyzed (N = 2), or because of inattentiveness of the child on both test trials (N = 1). From these children, 26 returned in two consecutive years. In the second year, children had a mean age of M = 3.64 years (SD = 0.29, range: 3.07–4.12 years), and in the third year M = 4.49 years (SD = 0.34, range: 3.89–5.09 years). The mean difference between the measurement time points was 339 days (SD = 92 days). At the age of 3

years, one child had to be excluded because of insufficient data quality, and at the age of 4 years one child dropped out because of technical problems with the eye-tracker. Parental informed consent was obtained for every testing session, in accordance with the approval from the local ethics committee (No. 236-10-23082010).

Previous studies that used similar anticipatory looking false belief tasks in infancy and toddlerhood reported between 75% and 85% correct first saccades (Senju et al., 2009, 2011; Southgate et al., 2007). A power analysis for a binomial test assuming 80% correct anticipatory saccades showed that an error probability of 5% and a power of 90% required a sample size of 23 participants. With our final sample size of 46 children at the age of 2 years we therefore expected a power of 99.4% to detect correct anticipatory saccades in toddlers. Another measure of anticipation that has been used by previous studies is the differential looking score (DLS). These studies reported an effect size between d = 0.67 (Wang & Leslie, 2016) and d = 0.95 (Senju et al., 2009). Assuming an effect size of d = 0.65, a power analysis for a one-sample *t*-test with a power of 90 % required a sample size of 22 participants. The expected power with our sample of 46 children aged 2 years was 99.6 % percent to detect a looking bias in toddlers.

2.1.2. Procedure

Children were presented with the original stimuli from Southgate et al. (2007) on the integrated monitor of a Tobii T120 eye-tracker using the Tobii Studio software (version 2.0.8). Children were seated in a car seat approximately 60 cm from the monitor, while their parents were standing behind them. Before the start of the experiment, each child passed a five-point infant calibration available in Tobii Studio. Each child was then presented with two familiarization (FAM) trials (one left, one right) and then a false belief test trial (condition FB1 or FB2 from Southgate et al., 2007), followed by another two FAM trials (in opposite order to the first two) and a second false belief test trial. The test trials were presented in randomized order across participants (first FB1 and second FB2 in half of the participants). Until the first false belief trial, the setup was identical to the setup in Southgate et al. (2007), with the exception of the sound which was lower in our study (approximately 52–54 dB for the sound of the telephone). This discrepancy resulted from a format conversion of the original film clips.

The stimuli by Southgate et al. (2007) showed a female agent behind a panel with two windows, and an opaque box in front of each window. In all trials, a bear puppet appeared and hid a colored ball in one of the two boxes. In the FAM trials, the bear then left again, the two windows were illuminated, and a bell sounded. After a delay of 1750 ms after the lights turned off, the agent put her hand through the window close to the box that contained the ball, and retrieved the ball from the box. The FAM intended to teach the child the agent's goal and that one of the windows was about to open after the bell sound and illumination of the windows.

In the false belief test trials, the bear moved the ball from the first box to the other box while the agent was either watching (condition FB1) or turning her back to the scene in response to the sound of a telephone (condition FB2; see Fig. 1). The bear then left the scene, and, in condition FB1, the agent now turned away in response to the telephone. Then, while the agent was turned away in both conditions and the telephone kept ringing, the bear removed the ball from the box and left the scene with it. The telephone then stopped ringing and the agent turned back to the scene, the windows were illuminated and the bell sounded. This marked the beginning of the critical test period during which the children's gaze was analyzed. In the false belief trials, no outcome was shown to prevent children from learning from the first false belief trial for the second. The two conditions FB1 and FB2 were orthogonal with respect to simpler non-belief-based rules, such as, gazing at the first or last box the ball had been in or the last box the agent had attended to (see Southgate et al., 2007), and therefore ensured that children did not pass the test based on these simpler cues.

2.1.3. Traditional explicit false belief tasks

At time point 2 and 3 of the study, that is, at the ages of 3 and 4 years, children additionally received two traditional explicit false belief tasks – a standard false location and a false content task. The procedure for these tasks was exactly as described in Grosse Wiesmann et al. (2017). The two traditional false belief tasks were assessed in randomized order, however, always following the

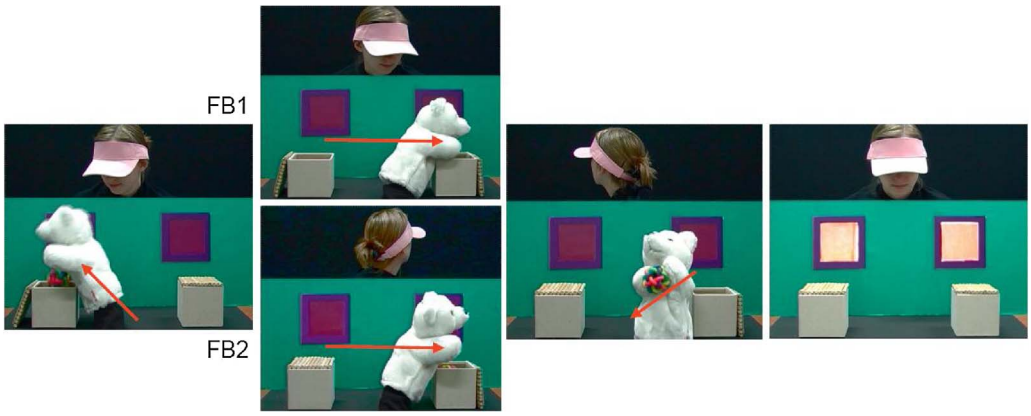

Fig. 1. Sequence of action in the two false belief conditions FB1 and FB2. In condition FB1, the agent observed how a bear puppet transferred a ball from the left to the right box, but turned away while the bear then removed the ball from the scene. In condition FB2, the agent turned away before the bear transferred the ball from the left to the right box, and only turned back after the bear had removed the ball from the scene. The two conditions respectively controlled for simpler non-belief related strategies, such as, gazing at the first or last box the ball had been in.

anticipatory looking task, which was administered first.

2.1.4. Analysis

Children's looking behavior was coded within two Areas of Interest each of which covered one window and the lid of the box below that window during the critical period of anticipation. This period was defined to start with the onset of the illumination and ended 1750 ms after the end of the illumination, according to the delay the children had been familiarized with until one window opened. Two measures of action anticipation were coded. First, we coded the DLS, that is, the difference in looking duration at the correct versus the incorrect window, divided by the total looking duration at either of the windows. The DLS scales from −1 to 1, and positive values indicate that children gazed longer at the correct window. It is a well-established measure of looking bias and considered to be highly reliable and sensitive (e.g., Gliga et al., 2014; Low, 2010; Senju et al., 2009; Thoermer et al., 2012; Wang & Leslie, 2016). Second, the children's first saccade to one of the windows after the onset of the illumination was coded as a measure of their anticipation where the agent would search for the ball. This corresponded to the principal measure in Southgate et al. (2007).

The children's attention was rated by 2 independent raters, and trials were excluded on a trial by trial basis if children had missed out on important parts of the course of action. This criterion led to the exclusion of 10 individual trials because of inattentiveness in the first year, and no exclusions in the second or third year. The raters agreed on 100 percent of the excluded trials.

Southgate et al. (2007) additionally applied a strict exclusion criterion according to which 11 out of 31 children were excluded because they did not anticipate correctly in the second FAM trial. In the present study, based on this criterion, we also had to exclude 31 out of 96 measurements, which corresponds to a similarly high exclusion rate of approximately one third. To avoid unnecessarily high exclusion rates and a potentially resulting sampling bias for more mature children, we analyzed and report both the results based on the strict original exclusion criterion and the results of the full sample. These results were highly similar (see Appendix A2), and there was no difference in false belief performance between children who passed and who did not pass the second FAM (see Appendix A1). We therefore report the results of the full sample in the main paper, and the reduced sample based on the strict original exclusion criterion in Appendix A2.

2.2. Results

2.2.1. Results of full dataset

The children's action anticipation was coded with two different measures, their DLS to the correct compared to the incorrect window and their first saccade.

2.2.1.1. DLS. For the DLS, a repeated measures ANOVA with two within-subject factors—time point (2, 3, and 4 years of age) and condition (FB1 and FB2)—showed a significant main effect of age (Greenhouse-Geisser test of within-subjects effects: $F(1.331) = 4.224$, $p = 0.049$) and a significant main effect of condition ($F(1) = 8.812$, $p = 0.013$) with better performance in condition FB1 than in condition FB2 (see Fig. 2). The interaction between time point and condition was not significant. Post-hoc pairwise comparisons showed that the children performed significantly better at the age of 4 years than at the ages of 2 years ($p = 0.008$) and 3 years ($p = 0.001$), but there was no significant difference in performance between 2 and 3 years of age ($p = 1.000$; see Fig. 3).

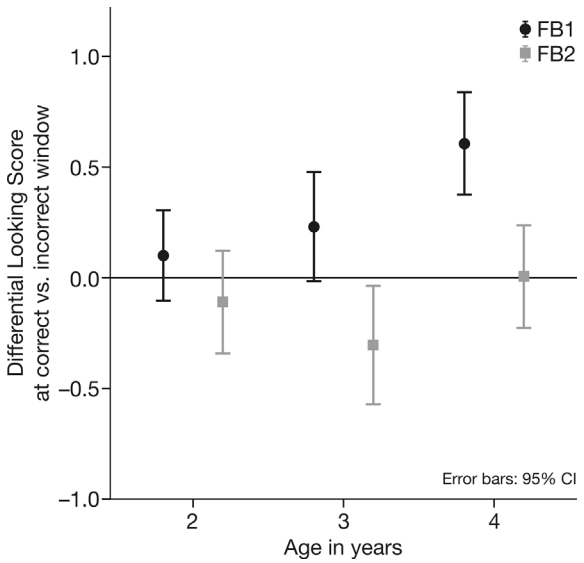

Fig. 2. Mean difference in looking times (DLS) at the correct compared to the incorrect window for conditions FB1 and FB2 separately. Children performed significantly better when the agent believed the ball to be in its last location (FB1) than when he believed it to be in its previous location (FB2).

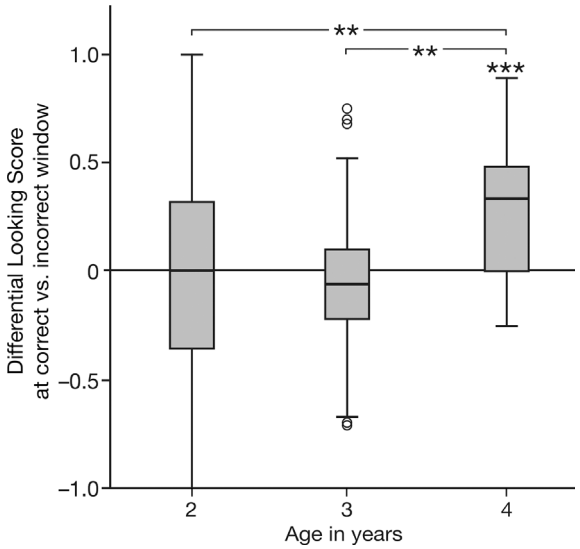

Fig. 3. Mean relative difference in looking times (DLS) at the correct compared to the incorrect window. Only by the age of 4 years did children gaze significantly longer at the correct than the incorrect window.

Children performed at chance level at the ages of 2 years (M = −0.03, SD = 0.51; one-sample *t*-test against 0: t(45) = −0.386, p = 0.701) and 3 years (M = −0.05, SD = 0.41; t(24) = −0.592, p = 0.559), and significantly above chance only by the age of 4 years (M = 0.30, SD = 0.36; t(24) = 4.094, p < 0.0005). Further, children only performed above chance in condition FB1 (M = 0.27, SD = 0.63; t(87) = 4.084, p < 0.0005), but not in condition FB2 (M = −0.13, SD = 0.67; t(89) = −1.883, p = 0.063).

2.2.1.2. First saccade. As a second measure we coded the children's first saccade after the onset of the illumination. Because of the binary nature of the first saccade data, we computed a binomial logistic regression with time point (2, 3, and 4 years of age) and condition (FB1 and FB2) as within-subject factors. This showed that, for the first saccade, there was no significant difference between the time points (Wald $Chi^2(2) = 2.089$; p = 0.352), but a marginally significant effect of the condition (Wald $Chi^2(1) = 3.327$, p = 0.034 one-sided) with better performance in FB1 than in FB2 (see Fig. 4). Only performance in condition FB1 was significantly above chance (55 out of 88, i.e. 63% correct first saccades, binomial test: p = 0.025), whereas performance in condition FB2 was at chance level (46 out of 90, i.e. 51% correct first saccades, binomial test: p = 0.916). There was no significant interaction between

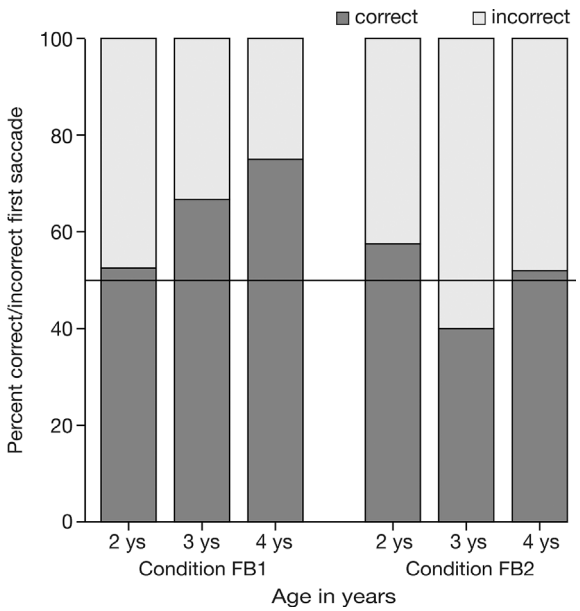

Fig. 4. Percent of children with correct first saccade in condition FB1 and FB2 respectively. Performance was above chance in condition FB1, and at chance level in condition FB2. There was no significant main effect of age.

time point and condition (Wald Chi^2(2) = 3.433, p = 0.180). To compare with the results of Southgate et al. (2007), we additionally tested whether children's first saccades were above chance at the different ages. Similar to the DLS, children performed at chance level at the ages of 2 (one-sample Wilcoxon signed rank test: p = 0.394) and 3 years (p = 0.564), and significantly above chance only at the age of 4 years (p = 0.035).

In sum, the DLS and first saccades showed above chance performance not until the age of 4 years, in contrast to the original study (Southgate et al., 2007) that showed above chance performance at the age of 25 months. A second discrepancy was that, across both measures of anticipation, we found a significant difference between conditions FB1 and FB2 with above chance performance only in condition FB1.

2.2.2. Only first trial

A potential reason for these deviations of our results from the original study could have resulted from the fact that, in our study, every child passed 2 false belief trials—the FB1 and the FB2 trial—whereas in the original study half of the children passed FB1 and the other half FB2 in a between-subjects design. In order to exclude this explanation, we performed a second analysis in which we only analyzed the first false belief trial (either FB1 or FB2 depending on the trial order a child had been presented with), thus reproducing the between-subjects design of Southgate et al. (2007). This analysis confirmed our findings that children only performed above chance by the age of 4 years, and only in condition FB1 and not FB2.

For the DLS, a repeated measures ANOVA with time point (2, 3, and 4 years of age) as within-factor and condition (FB1 and FB2) as between factor confirmed the significant main effect of time point (Greenhouse-Geisser test of within-subjects effects: F(1.362) = 7.473, p = 0.008) and of condition (F(1) = 22.536, p < 0.0005) with better performance in FB1 than in FB2. There was no significant interaction. As before, children performed significantly better at the age of 4 than at the ages of 2 years (p = 0.002) and 3 years (p = 0.004), and there was no significant difference between 2 and 3 years of age (p = 0.24). As before, performance was at chance level at the ages of 2 years (M = 0.07, SD = 0.67; one-sample *t*-test against 0: t(44) = −0.656, p = 0.516) and 3 years (M = −0.02, SD = 0.63; t(24) = −0.198, p = 0.848), and significantly above chance only by the age of 4 years (M = 0.37, SD = 0.66; t(23) = 2.768, p = 0.011). Likewise, children only performed above chance in condition FB1 (M = 0.35, SD = 0.62; t(45) = 3.783, p < 0.0005), but at chance in condition FB2 (M = −0.12, SD = 0.62; t(46) = −1.341, p = 0.187).

For the first saccade, as before, the effect of time point did not become significant (Wald Chi^2(2) = 2.944; p = 0.229) and there was a marginal effect of FB condition (Wald Chi^2(1) = 3.327, p = 0.034 one-sided) with better performance in FB1 than FB2. Only performance in condition FB1 was significantly above chance (32 out of 46, i.e. 70% correct, binomial test: p = 0.011), whereas performance in condition FB2 was at chance level (25 out of 47, i.e. 53% correct, binomial test: p = 0.771). Again, children's first saccade was at chance level at the ages of 2 (25 out of 45, i.e. 56% correct, binomial test: p = 0.551) and 3 years (15 out of 25, i.e. 60% correct, binomial test: p = 0.424), but significantly above chance by the age of 4 years (18 out of 24 correct, i.e. 75% correct first saccades, binomial test: p = 0.023).

2.2.3. Relation with traditional false belief tasks

At the age of 3 and 4 years, children additionally received two traditional false belief tasks – a false location and a false content task. As expected, there was a significant developmental change between the ages of 3 and 4 years (Wilcoxon signed rank test, false location: Z = −3.560, p < 0.0005, false content: Z = −3.232, p = 0.001). At the age of 3 years, children performed significantly below chance on the false location task (one-sample Wilcoxon signed rank test, median = 0.00, Z = −2.431, p = 0.015) and at chance-level on the false content task (median = 0.33, Z = −1.270, p = 0.204), whereas, at the age of 4 years, the children performed above chance on both tasks (false location: median = 1.00, Z = 3.337, p = 0.001; false content: median = 0.67, Z = 3.384, p = 0.001).

Because of the significant developmental change that we found in the anticipatory looking task between the ages of 3 and 4 years, at an age when children also start passing traditional false belief tasks, we were interested in the relation of anticipatory looking with the traditional false belief tasks. We computed a linear regression including the subject as a random effect predicting the children's DLS score in the anticipatory looking task with either the false location or the false content task as a fixed factor. This showed no significant correlation between anticipatory looking and the traditional false location task (Wald Chi^2(3) = 5.542, p = 0.136), nor between anticipatory looking and the false content task (Wald Chi^2(4) = 5.502, p = 0.240). For the first saccades, we computed an analogous multinomial logistic regression (taking into account the ordinal nature of the data with 0, 1, or 2 correct trials). This showed that children's first saccades in the anticipatory looking task were significantly related to the traditional false location task (Wald Chi^2(3) = 8.131, p = 0.043), but not to the false content task (Wald Chi^2(3) = 1.136, p = 0.889).

2.3. Discussion

We found a significant developmental change in an established anticipatory looking false belief task between the ages of 3 and 4 years and correct anticipation not until the age of 4 years. This finding was consistent across different measures of anticipation and analyses. Moreover, the analyses consistently show that children anticipated above chance only in condition FB1, where the agent believed the ball to be in its last location, but not in condition FB2. These results were in contrast to the original study that showed correct anticipation already at 25 months and no difference between the two FB conditions. What differences in our study compared to the original study could have explained this discrepancy? We showed that our results were independent of the application of the original performance-based inclusion criterion that excluded all participants who did not show correct anticipatory looking in the second FAM trial. Moreover, we replicated our results when analyzing only the first FB trial. This showed that the deviations from the

original study did not result from presenting repeated trials in our within-subjects design compared to the single trial presented in the original study.

What differences in our study compared to the original study could have explained this discrepancy? Apart from the trial repetition in our study, the setup of our study was exactly as in the original study, except for one additional small difference: In the original study, there had been a salient ringing of a telephone while the agent turned her back to the scene. This underlined the fact that the agent turned away and did not witness or hear the change of location. In contrast to the original study, in our study, the volume of this ring tone was very low. The more salient ring tone in the original study might have enhanced infants' performance in several ways. First, it gave a plausible reason why the agent turned away, and underlined that she did not pay attention to the scene and could not witness the relocation of the ball. Second, the salient ringing of the telephone centered the child's attention on the agent rather than the bear that was relocating the ball. This might have helped children to focus on the agent and keep track of her perspective in the original study (e.g., Helming et al., 2014; Rubio-Fernández & Geurts, 2013, 2015). Third, Heyes (2014) suggested that the ring tone might have entirely drawn the children's attention to the agent and distracted them from the relocation of the ball, so that the children themselves might have missed the relocation. The children could then have gazed correctly in the original study based on their own representation of the ball rather than on their understanding of the agent's false belief. If the salience of the ring tone indeed had an impact on children's performance in the FB trials, this would therefore have important implications for the theoretical interpretation of early ToM findings in infants.

Therefore, we tested and excluded this hypothesis systematically in a second experiment by manipulating the sound while the agent turned her back to the scene. The setting was exactly as in Southgate et al. (2007), except for the fact that for half of the children there was no sound while the agent turned away, while the other half heard a salient telephone ringing as in the original study. The results of this experiment are reported in the next section.

3. Experiment 2

3.1. Methods

3.1.1. Participants

In experiment 2, eye-tracking data from 57 toddlers (mean age M = 25.2 months, SD = 0.70, age range: 24.2–27.0 months, 19 female) was collected, corresponding to the age range in Southgate et al. (2007). From these, 7 had to be excluded because data quality was insufficient to be analyzed (N = 5), or because of inattentiveness of the child on the test trials (N = 2). Consequently, 27 children saw the original videos with the sound of the telephone (*with-sound condition*: mean age M = 25.27 months, SD = 0.65, 12 female), and 23 children saw the videos with no sound while the agent turned her back to the scene (*no-sound condition*: mean age M = 25.02 months, SD = 0.67, 12 female). Parental informed consent was obtained for every testing session, in accordance with the approval from the local ethics committee (No. 236-10-23082010).

3.1.2. Procedure

As before, children were presented with the original stimuli from Southgate et al. (2007) on the integrated monitor of the eye-tracker used in experiment 1. The setting and procedure of experiment 2 were exactly as before with the exception that, now, each child was presented with two FAM trials followed by only a single FB trial (FB1 or FB2), instead of two FB trials as in experiment 1. This was done to keep the procedure as similar as possible to the original study. Crucially, one group of children (N = 27) watched the false belief video with the salient sound of a telephone ringing (approx. 62–64 dB) while the agent turned her back to the scene as in the original study, whereas the other group (N = 23) heard no sound during this period of the video. The children were equally distributed to condition FB1 and FB2 within each group.

3.1.3. Analysis

Children's first saccades and their DLS were analyzed as described in experiment 1. Because we showed that children's performance was independent of the inclusion criterion applied in the original study, we analyzed the data of the full sample in experiment 2.

3.2. Results

For the DLS, a 2 by 2 ANOVA with sound and FB condition as between-subject factors showed no significant main effect of sound ($F(1,46) = 0.016$, $p = 0.833$), but confirmed the significant main effect of FB condition ($F(1,46) = 8.872$, $p = 0.005$), with significantly better performance in FB1 than in FB2 (see Fig. 5). The interaction between the sound and FB condition was not significant ($F(1,46) = 1.802$, $p = 0.186$). The children performed marginally significantly above chance in condition FB1 (M = 0.25, SD = 0.63, one-sample *t*-test p = 0.028 one-sided), and significantly below chance in FB2 (M = −0.28, SD = 0.57, one-sample *t*-test p = 0.025), in total confirming the chance-level performance of toddlers found in experiment 1.

For the first saccade, a binary logistic regression with sound and FB condition as between-subjects factors showed no significant effect of sound (Wald $Chi^2(1) = 0.095$, $p = 0.758$), nor of FB condition (Wald $Chi^2(1) = 0.154$, $p = 0.694$), and there was no significant interaction (Wald $Chi^2(1) = 1.339$, $p = 0.247$). Toddlers gazed at chance level (25 out of 50, i.e. 50%, correct first saccades, binomial test p = 1.0), confirming the results of experiment 1.

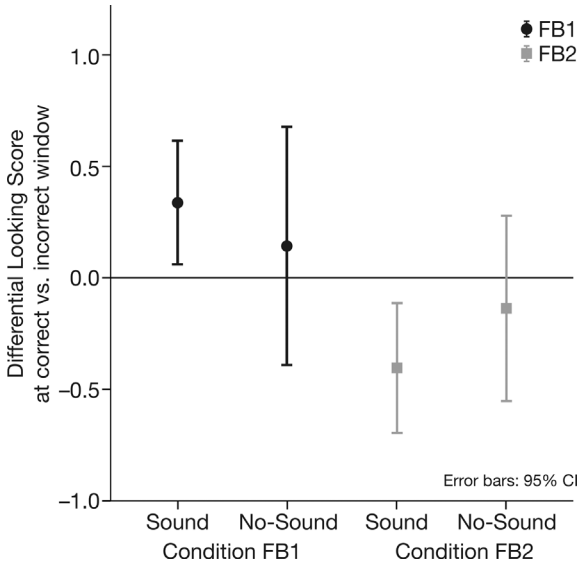

Fig. 5. Difference in looking duration (DLS) at the correct compared to the incorrect window for children who saw the videos with the sound of a telephone or without, for the two FB conditions separately (FB1 and FB2). Children performed significantly better in condition FB1 than in FB2. There was no difference between the sound and the no-sound condition.

3.3. Discussion

In experiment 2, we confirmed our finding of chance-level performance of 2-year-olds on a false belief task by Southgate et al. (2007) in an independent sample of 50 toddlers aged 25 months. Moreover, by manipulating the sound of a telephone ringing while the agent turned her back to the scene, we showed that children's performance was independent of the presence of this sound in the critical phase where the agent acquired a false belief. This manipulation eliminated a potential explanation for the deviation of our findings compared to the original study (Southgate et al., 2007) that had reported correct false-belief-related anticipation in 20 toddlers aged 25-months. That is, our results show that the low volume of the telephone in experiment 1, compared to the salient ringing of a telephone in the original study, does not seem to have caused 2- and 3-year-olds to fail the task in our study. Instead, we show that toddlers fail to anticipate correctly, independently of the presence of a salient sound or not. Further, by studying precisely the same age range than the original study and replicating the exact procedure of the original study, we excluded any other difference between the studies that might have explained the discrepancy between our findings and the original study.

Moreover, experiment 2 also confirmed the significant difference in performance between the two different false belief conditions of the original study. As in experiment 1, children performed significantly better if the agent believed the object to be in its last location (condition FB1) rather than a previous one (condition FB2).

4. General discussion

The present study addressed how early ToM abilities observed in infants' anticipatory looking develop between 2 and 4 years of age, when children start passing traditional explicit false belief tasks, and how robust these abilities are during early childhood. While previous studies had mainly focused on infants and toddlers or adults, we tested 2- to 4-year-old children longitudinally with an established anticipatory looking false belief task (Southgate et al., 2007) in three consecutive years. This longitudinal design allowed us to study developmental change in belief-related anticipatory looking during early childhood. Additionally, this design allowed us to test how robust implicit false belief tasks are in 2-year-old children, and whether potential fragility of task performance changes in the course of preschool age. We found that children only anticipated correctly where an agent with a false belief about the location of a ball would search for this ball at the age of 4 years and not yet at the ages of 2 and 3 years, in contrast to the original study which had found correct anticipation at 25 months (Southgate et al., 2007). In our study, there was a significant developmental change between the ages of 3 and 4 years, at a similar age when children start passing traditional explicit false belief tasks. Moreover, we showed a significant difference in performance between two different false belief conditions that respectively controlled for simpler non-belief-related strategies, such as gazing at the last location that had contained the ball. Children performed significantly better when the agent believed the ball to be in the last location where it had been before it was removed from the scene (condition FB1), rather than when the agent believed it to be in a previous location (condition FB2). Both patterns–the significant developmental change between 3 and 4 years and the difference between the two false belief conditions were consistent across different measures of anticipation (first saccade or looking duration to the correct compared to the incorrect location), in different subsamples of the data (in the total sample and in a subsample of children who correctly anticipated the agent's search in a familiarization phase), and both within-subjects and between-subjects. Moreover, we replicated the chance-level performance at the age of 2 years as well as the

difference between the two false belief conditions in two independent samples of 46 toddlers in experiment 1 and 50 toddlers in experiment 2.

These results stand in contrast to the original study that had found correct anticipation in 20 toddlers aged 25 months, and no difference between the two false belief conditions (Southgate et al., 2007). A few other studies had replicated correct anticipation in 2- and 3-year-old children (Gliga et al., 2014; Senju et al., 2011; Wang & Leslie, 2016), however, most of these studies only used false belief condition FB1, which we also replicated in our study. Only one study replicated condition FB2 (Wang & Leslie, 2016), which we failed to replicate in the present study.

What factors could have led to the discrepancy between the present and the original findings from Southgate et al. (2007)? First, we showed that our results were independent of an inclusion criterion based on children's performance in the familiarization phase before the test trials that had been applied by the original study. This inclusion criterion, therefore, cannot explain differences in the results. One difference in our study compared to the original study was that every child performed two consecutive false belief trials, an FB1 and an FB2 trial in counterbalanced order, whereas, in the original study, children had only received a single trial (FB1 or FB2) in a between-subject design. To exclude the possibility that differences in performance might have resulted from carry over effects from the first to the second trial, we additionally analyzed only the first false belief trial. This replicated our results of a significant increase in performance between the ages of 3 and 4 years with above chance performance only by the age of 4 years. Further, it replicated the significant difference in performance between the two false belief conditions. This analysis, therefore, shows that the discrepancy of the results of our study compared to Southgate et al. (2007) was not due to the fact that children performed multiple trials.

A second difference in our setup compared to the original study was that the sound in our study was lower. In particular, the sound of the telephone ringing while the agent turned her back to the scene and thus obtained a false belief was very low. In experiment 2, we therefore tested whether the sound of the telephone had an impact on toddlers' anticipation by systematically manipulating the sound across subjects. Half of the children saw the original videos with the sound of the telephone, while the other half heard no sound while the agent turned her back to the scene and missed the relocation of the ball. All other settings were exactly as in the original study, including the exact age range of the participants. This showed that children's anticipatory looking was independent of the sound of the telephone. Furthermore, experiment 2 replicated our findings of experiment 1 that 2-year-olds performed at chance-level in the anticipatory looking false belief task by Southgate et al. (2007), and confirmed the significant difference between the two false belief conditions (FB1 and FB2) in an independent sample of 50 toddlers aged 25 months exactly as in the original study. This shows that the discrepancy between our results and the original study could not be explained by the volume differences in experiment 1 compared to Southgate et al. (2007).

Together, the analyses in experiment 1 and 2 systematically rule out methodological differences between the present and the original study as contributing factors to the discrepant results. Our findings, therefore, question previously reported early false-belief-related action anticipation in 2-year-olds. In contrast, we find that children do not anticipate correctly until the age of 4 years.

The observed developmental change between the ages of 3 and 4 years took place at a similar age as the well-established developmental breakthrough in traditional explicit false belief tasks. This raises the question of how development in these two task types is related. By studying the correlation of the anticipatory looking task with two traditional false belief tasks, we found some indication for a relation of 3- and 4-year-olds anticipatory saccades with a traditional false location task. The correlation, however, was not robust across different measures of anticipation and traditional false belief tasks. Previous studies had found no correlation between implicit and explicit false belief tasks (Grosse Wiesmann et al., 2016) and a dissociation on the neural level (Grosse Wiesmann et al., 2017; Schneider, Slaughter, Becker, & Dux, 2014). This question, therefore, requires further investigation by future research.

Another discrepancy in our results compared to the original study was the significant difference between performance in the two false belief conditions FB1 and FB2 that were orthogonal with respect to simpler non-belief-based strategies. The children anticipated significantly worse when the agent did not notice that the ball was transferred to the other box before it was entirely removed from the scene (condition FB2), in line with recent findings by Kulke & Rakoczy (2017). Children performed at chance level in condition FB2 across all ages. These results were in contrast with the original study that had found no significant difference between the two conditions (Southgate et al., 2007). However, the sample size in the original study was too small to test whether performance in condition FB2 individually was above chance (10 children per condition). The other studies that used the same paradigm only tested one of the two conditions (Gliga et al., 2014; Senju et al., 2009; Wang & Leslie, 2016) so that performance on the two conditions could not be compared. Kulke & Rakoczy (2017), in turn, found a very similar difference between the conditions FB1 and FB2 in a large sample of 2- to 6-year-old children.

Why might condition FB2 have been more difficult than condition FB1? In condition FB2, children had to remember the agent's belief about the object for a longer period of time while they were updating their own knowledge about the actual location of the ball at least twice (transfer to the other box and removal from the scene). In contrast, in condition FB1, children only had to remember that the agent believed the object to remain in the last box while they updated their own knowledge that it was being removed from the scene. It is therefore plausible that condition FB2 made higher working memory and conflict inhibition demands to maintain the agent's belief for a longer time against one's own conflicting perspective. Condition FB2 was constructed to control for simple non-belief-related strategies or for paying attention to different things other than anticipating the agent's action, such as, gazing at the last location of the ball. Therefore, only above chance performance in both conditions together, FB1 and FB2, showed that children anticipated the agent's actions, and were not paying attention to some other aspect of the task, such as, the last ball location. However, it is unlikely that 4-year-olds passed condition FB1 simply because they gazed towards the last ball location because such a strategy would have led to below chance rather than the observed chance level performance in condition FB2. Moreover, 4-year-olds

performed above chance when both conditions were collapsed, which they could not have achieved with a simpler non-belief-based strategy. Finally, at the age of 4 years, the majority of children passed the traditional explicit false belief tests, which indicates that they did not fail in condition FB2 because of difficulties to attribute false beliefs. Instead, it is more likely that 4-year-olds failed in this condition because of lacking working memory or executive abilities.

These limitations of condition FB2 do not affect our conclusion of a developmental change between 3 and 4 years and correct anticipation not until 4 years because this pattern was also present in condition FB1–even in condition FB1, children did not perform above chance until the age of 4 years. This supports our conclusion of our lack of evidence for early false belief-related action anticipation.

5. Conclusions

We show a developmental change in children's anticipation of the actions of an agent with a false belief between the ages of 3 and 4 years in an established anticipatory looking task (Southgate et al., 2007). Children consistently only anticipated correctly at the age of 4 years, and not yet at the ages of 2 and 3 years, in contrast to previous studies that had shown correct anticipation already at 2 years. Furthermore, children only anticipated correctly when the agent believed the ball to be in its last location rather than in the previous location, which was likely due to high working memory demands of the latter condition. Our results contradict previous conclusions of early false-belief-related action anticipation, and, instead, indicate that this might only develop around the age of 4 years, when children also start explicitly attributing false beliefs in the traditional tasks.

Declaration of interest

The authors declare no conflict of interest.

Acknowledgements

The research was supported by a grant from the German National Academic Foundation to CGW and a grant from the European Research Council (ERC-2010-360 AdG 20100407) awarded to AF. We would like to thank Victoria Southgate for sharing her stimulus material and very helpful discussions, Louisa Kulke for comments to the manuscript, as well as Hung Nguyen Trong, Elisabeth Muche, Christian Kliesch, Christine Schipke, and Anna Strotseva-Feinschmidt for their help with data acquisition, Ulrike Kachel for data coding, and Katja Kirsche for recruitment and coordination of the testing.

Appendix A1 No performance difference between children who passed and did not pass the original inclusion criterion

The original study had a strict exclusion criterion, according to which only children were included into the analysis that showed correct first saccades in the second familiarization trial. This exclusion criterion led to an exclusion rate of approximately one third of the children in the original as well as in our study. To avoid high exclusion rates and a potential sampling bias, we therefore tested the effect and necessity of this exclusion criterion by comparing false belief performance of children who passed and who did not pass the criterion. This showed that there was no difference in false belief performance between children who passed and who did not pass the criterion, neither for the DLS nor for their first saccades.

For the DLS, this was shown with a linear model that included time point (2, 3 and 4 years of age) as repeated measure and inclusion criterion (pass or fail) as between-subjects factor. This showed that there was no main effect of the inclusion criterion ($F(1,88.03) = .884$; $p = 0.447$) nor a significant interaction with the time point ($F(2,87.32) = 0.653$; $p = 0.523$). This lack of a difference in the DLS between children who passed and did not pass the inclusion criterion was confirmed by direct comparisons of the groups at each time point (at age 2: children who passed: $M = -0.03$, $SD = 0.56$, children who failed: $M = -0.05$, $SD = 0.48$, independent samples *t*-test $t(43) = 0.16$, $p = 0.87$; at age 3: children who passed: $M = -0.10$, $SD = 0.41$, children who failed: $M = 0.19$, $SD = 0.38$, independent samples *t*-test $t(23) = -1.30$, $p = 0.21$; at age 4: children who passed: $M = 0.30$, $SD = 0.41$, children who failed: $M = 0.29$, $SD = 0.23$, independent samples *t*-test $t(23) = 0.03$, $p = 0.97$) as well as collapsed across all time points (children who passed: $M = 0.04$, $SD = 0.50$, children who failed: $M = 0.07$, $SD = 0.43$, independent samples *t*-test $t(93) = -0.346$, $p = 0.73$).

For the first saccade in the false belief trials, an ordinal logistic regression with time point (2, 3 and 4 years of age) and inclusion criterion (pass or fail) showed no main effect of the inclusion criterion (Wald Chi-Square = 1.18, $p = 0.28$), nor a significant interaction with the time point (Wald Chi-Square = 0.52, $p = 0.773$). The lack of a significant difference in first saccades between children who passed and did not pass the inclusion criterion at every time point (independent-samples Mann-Whitney U Test, at age 2: $p = 0.37$; at age 3: $p = 0.41$; and at age 4: $p = 0.93$) as well as collapsed across all time points (independent-samples Mann-Whitney U Test: $p = 0.28$) confirmed that the inclusion criterion had no impact on children's false belief performance.

Thus, because children's false belief performance was independent of whether they passed the second familiarization or not, we have reported the results of the full sample in the Results. Applying the original inclusion criterion, however, yields comparable results, as is shown in Appendix A2.

Appendix A2 Results with original inclusion criterion

In a second analysis, all children were excluded who did not pass the second familiarization trial following the inclusion criterion of the original study (Southgate et al., 2007). According to this criterion, we were left with 27 out of 46 children in the first year (2 years of age), 21 out of 25 children in the second year (3 years of age), and 17 out of 25 children in the last year (4 years of age). The total drop-out rate of 32% due to this criterion was comparable to the original study. Because according this criterion different children had to be excluded at every time point, data for both FB conditions from all three time points was only available from 6 children. We therefore analyzed every time point individually. This confirmed the pattern of results found for the full sample.

For the DLS, children performed at chance level at the ages of 2 ($M = -0.03$, $SD = 0.56$, one-sample t-test: $p = 0.813$) and 3 years ($M = -0.10$, $SD = 0.41$, $p = 0.30$), and above chance only at the age of 4 years ($M = 0.30$, $SD = 0.41$, $p = 0.009$). As in the full sample, children performed significantly better in condition FB1 than in FB2, marginally at 2 years (FB1: $M = 0.25$, $SD = 0.55$; FB2: $M = 0.25$, $SD = 0.55$; paired samples t-test: $p = 0.035$ one-sided), and significantly at 3 years (FB1: $M = 0.24$, $SD = 0.60$; FB2: $M = -0.40$, $SD = 0.63$; $p = 0.005$ two-sided) and at 4 years (FB1: $M = 0.65$, $SD = 0.62$; FB2: $M = -0.03$, $SD = 0.66$; $p = 0.013$ two-sided).

Similarly, for the first saccade, children performed at chance level at the ages of 2 (one-sample Wilcoxon signed rank test: $p = 0.166$) and 3 years ($p = 0.317$), and marginally significantly above chance at the age of 4 years ($p = 0.030$, one-sided). For the first saccade, children performed significantly better in condition FB1 than FB2 at the age of 3 years ($N = 10$ passed FB1 but failed FB2, and only $N = 2$ passed FB2 but failed FB1, McNemar's test $p = 0.039$), and marginally at 4 years ($N = 8$ passed FB1 but failed FB2, and only $N = 2$ passed FB2 but failed FB1, McNemar's test $p = 0.05$ one-sided), but their was no difference between FB conditions at 2 years ($N = 7$ passed FB1 but failed FB2, and $N = 7$ passed FB2 but failed FB1).

Taken together, these analyses show that the pattern of results reported in this paper remain similar, independently of whether the original inclusion criterion is applied or not. That is, children performed at chance level at the ages of 2 and 3 years and not above chance until the age of 4 years, and performance in condition FB1 was significantly better than in condition FB2

References

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

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

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

Gliga, T., Senju, A., Pettinato, M., Charman, T., & Johnson, M. H. (2014). Spontaneous belief attribution in younger siblings of children on the autism spectrum. *Developmental Psychology, 50*(3), 903–913. http://dx.doi.org/10.1037/a0034146.

Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2016). Implicit and explicit false belief development in preschool children. *Developmental Science,* 1–15. http://dx.doi.org/10.1111/desc.12445.

Grosse Wiesmann, C., Schreiber, J., Singer, T., Steinbeis, N., & Friederici, A. D. (2017). White matter maturation is associated with the emergence of Theory of Mind in early childhood. *Nature Communications, 8*, 14692. http://dx.doi.org/10.1038/ncomms14692.

He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-year-olds succeed at a verbal anticipatory-looking false-belief task. *British Journal of Developmental Psychology, 30*, 14–29. http://dx.doi.org/10.1111/j.2044-835X.2011.02070.x.

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*(4), 167–170. http://dx.doi.org/10.1016/j.tics.2014.01.005.

Heyes, C. M. (2014). False belief in infancy: A fresh look. *Developmental Science, 17*(5), 647–659. http://dx.doi.org/10.1111/desc.12148.

Kulke, L., & Rakoczy, H. (2017). How robust and replicable are implicit Theory of Mind tasks? *Paper presented at the BCCCD, 2017*.

Low, J. (2010). Preschoolers' implicit and explicit false-belief understanding: Relations with complex syntactical mastery. *Child Development, 81*(2), 597–615. http://dx.doi.org/10.1111/j.1467-8624.2009.01418.x.

Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., et al. (2012). Belief attribution in deaf and hearing infants. *Developmental Science, 15*(5), 633–640. http://dx.doi.org/10.1111/j.1467-7687.2012.01155.x.

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

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

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

Rubio-Fernández, P., & Geurts, B. (2015). Don't mention the marble! The role of attentional processes in false-belief tasks. *Review of Philosophy and Psychology, 7*(October), 835–850. http://dx.doi.org/10.1007/s13164-015-0290-z.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review, 34*(3), 265–293. http://dx.doi.org/10.1016/j.dr.2014.04.001.

Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *Neuroimage, 101*, 268–275. http://dx.doi.org/10.1016/j.neuroimage.2014.07.014.

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

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

Sodian, B. (2016). Is false belief understanding continuous from infancy to preschool age? In D. Barner, & A. S. Baron (Eds.). *Core Knowledge and Conceptual Change*Oxford University Press p. 301.

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

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *The British Journal of Developmental Psychology, 4*, 30–44. http://dx.doi.org/10.1111/j.2044-835X.2011.02046.x.

Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. *The British Journal of Developmental Psychology, 30*(Pt 1), 172–187. http://dx.doi.org/10.1111/j.2044-835X.2011.02067.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 and Language, 31*(2), 147–176. http://dx.doi.org/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.