Received: 11 August 2021 | Revised: 17 December 2021 | Accepted: 20 December 2021

DOI: 10.1111/desc.13224

Developmental Science WILEY

RESEARCH ARTICLE

Evidence for goal- and mixed evidence for false belief-based action prediction in 2- to 4-year-old children: A large-scale longitudinal anticipatory looking replication study

Larissa J. Kaltefleiter[1] | Tobias Schuwerk[1] | Charlotte Grosse Wiesmann[2] | Susanne Kristen-Antonow[1] | Irina Jarvers[1] | Beate Sodian[1]

[1] Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

[2] Minerva Fast Track Group Milestones of Early Cognitive Development, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Correspondence
Larissa J. Kaltefleiter, Department of Psychology, Ludwig-Maximilians-Universität München, Leopoldstraße 13, München 80802, Germany.
Email: Larissa.Kaltefleiter@psy.lmu.de

Abstract

Unsuccessful replication attempts of paradigms assessing children's implicit tracking of false beliefs have instigated the debate on whether or not children have an implicit understanding of false beliefs before the age of four. A novel multi-trial anticipatory looking false belief paradigm yielded evidence of implicit false belief reasoning in 3- to 4-year-old children using a combined score of two false belief conditions (Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. [2017]. *Developmental Science*, *20*(5), e12445). The present study is a large-scale replication attempt of this paradigm. The task was administered three times to the same sample of $N = 185$ children at 2, 3, and 4 years of age. Using the original stimuli, we did not replicate the original finding of above-chance belief-congruent looking in a combined score of two false belief conditions in either of the three age groups. Interestingly, the overall pattern of results was comparable to the original study. Post-hoc analyses revealed, however, that children performed above chance in one false belief condition (FB1) and below chance in the other false belief condition (FB2), thus yielding mixed evidence of children's false belief-based action predictions. Similar to the original study, participants' performance did not change with age and was not related to children's general language skills. This study demonstrates the importance of large-scaled replications and adds to the growing number of research questioning the validity and reliability of anticipatory looking false belief paradigms as a robust measure of children's implicit tracking of beliefs.

KEYWORDS
action prediction, anticipatory looking, early childhood, false belief, replication, theory of mind

1 | INTRODUCTION

Human behavior is driven by beliefs, intentions, desires, and emotions. Theory of Mind, the ability to attribute such mental states to oneself and others (Premack & Woodruff, 1978), allows us to explain and predict behavior (Frith & Frith, 2012). In the past decades, research on Theory of Mind has predominantly focused on the comprehension of false beliefs and the development of false belief understanding in childhood. When asked to predict where an agent who originally hid an object in location A and did not witness the transfer of the object to location B will search for the object, children between the ages of three and five begin to take into consideration that the agent holds a false belief about the object's location when predicting their action (Wellman et al., 2001; Wimmer & Perner, 1983).

In contrast to the traditional view that children only develop false belief understanding around the age of four, recent studies utilizing novel task formats have yielded evidence of false belief understanding at a much earlier age. Here, we will focus on anticipatory looking paradigms. Such paradigms make use of humans' tendency to anticipate actions while observing them which already develops until the end of the first year of life (e.g., Falck-Ytter et al., 2006; Flanagan & Johansson, 2003). Within anticipatory looking paradigms, actions that are performed based upon certain mental states (such as intentions or true/false beliefs) are presented and participants' action anticipation is observed to figure out whether participants take the agent's mental state into consideration when predicting their action. Using such an intriguing new task design, Clements and Perner (1994) were the first to measure children's false belief understanding employing a nonverbal paradigm. Enacting a standard change-of-location false belief task, they tracked children's anticipatory looks while prompting them to anticipate the agent's behavior. After the anticipatory phase, the explicit false belief action prediction question was uttered. Analyzing children's anticipatory looks, Clements and Perner (1994) found that children from 2;11 years on reliably looked at the old location of the target object in false belief trials and the current location of the target object in the true belief trials, thereby indicating an implicit understanding of false belief. Strikingly, the same children were unable to correctly verbally predict the agent's searching behavior in the false belief trials, exhibiting a lack of explicit false belief understanding (for an overview of the conceptual and terminological implicit-explicit distinction, see Perner & Roessler, 2012). Since then, a number of studies using anticipatory looking paradigms have contributed supporting evidence for the view that even 2-year-old children and infants possess implicit false belief understanding (e.g., Garnham & Ruffman, 2001; Ruffman et al., 2001; Senju et al., 2011; Southgate et al., 2007; Surian & Geraci, 2012; Surian & Franchin, 2020; Thoermer et al., 2012; Wang et al., 2012).

In the past years, competing accounts of early, implicit false belief understanding have been formulated based on these important findings. These accounts have tried to reconcile the new findings with the traditional results from standard explicit false belief tasks. On the one hand, proponents of the conceptual continuity view assume that Theory of Mind abilities are present from infancy on. They attribute the failure of younger children in explicit false belief tasks to children's struggle with the task demands of these explicit tasks, such as their limited inhibitory control (e.g., Wang & Leslie, 2016) or their insufficient pragmatic skills when interpreting the test question (Siegal & Beattie, 1991). Therefore, in spontaneous-response tasks in which inhibitory and pragmatic demands are reduced children can succeed at a much younger age (Baillargeon et al., 2010; Scott, 2017). On the other hand, supporters of a conceptual-change view assume that only the application of behavioral rules (Perner & Roessler, 2012; Perner & Ruffman, 2005; Ruffman & Perner, 2005) or infants' well-developed statistical learning skills (Ruffman, 2014) lead to successful performance in spontaneous-response tasks and that Theory of Mind abilities only develop later. Further, Heyes (2014a, 2014b) argues that children's success on implicit tasks stems from domain-general processes such as perceptual novelty. Dual-systems accounts assume that humans are equipped with an implicit and very efficient system for processing mental states which is already present in infants and a second more flexible, but less efficient, explicit system which only develops in the preschool period and is tied to the presence of language abilities and executive functions (Apperly & Butterfill, 2009; Grosse Wiesmann et al., 2017, 2020).

Research highlights

- We investigated early false belief tracking through a large-scaled longitudinal replication study
- In a multi-trial anticipatory looking paradigm, we did not replicate the original finding of above-chance belief-congruent looking in 3- and 4-year-olds with a combined measure of two false belief conditions
- We found above-chance false belief-congruent anticipatory looking in 2-, 3-, and 4-year-olds in false belief condition FB1, but below-chance performance in false belief condition FB2
- Our findings suggest that 2- to 4-year-olds track an agent's goal, but not reliably its false belief in an anticipatory looking task

If implicit task formats indeed measure early false belief competencies, continuity between false belief performance in infancy and childhood would be expected (Setoh et al., 2016). In a multi-measure longitudinal study, significant developmental relations between implicit false belief reasoning at 18 months and explicit false belief understanding at 48 months (Thoermer et al., 2012), and at 50, 60, and 70 months (Kloo et al., 2021) as well as belief-based intention understanding at 60 months (Sodian et al., 2016) were observed. In contrast, a study by Poulin-Dubois et al. (2020) did not find longitudinal relations between implicit false belief understanding in infancy and later implicit and explicit false belief understanding at 4 to 5 years.

Regarding concurrent relations between implicit and explicit false belief reasoning, on the one hand, Poulin-Dubois et al. (2020) did not find concurrent relations between performance in two explicit false belief tasks and in an anticipatory looking false belief task in 4- and 5-year-old children, and also Grosse Wiesmann et al. (2017) did not find such relations in 3- and 4-year-old children. On the other hand, Low (2010) detected relations between implicit and explicit false belief reasoning in 3- and 4-year-olds when using the same task design for implicit and explicit false belief, and Grosse Wiesmann et al. (2018) found a tentative concurrent relation between an explicit false belief task and some measures of anticipatory looking (first look but not relative looking duration).

While numerous findings of implicit false belief understanding in infants fueled the controversial debate on children's early mental state reasoning abilities, the past few years have yielded a fast-growing number of partial or failed replication attempts of anticipatory looking tasks assessing implicit false belief understanding (Kulke & Rakoczy, 2018) leading to what some researchers consider a replication crisis

(Poulin-Dubois et al., 2018). According to a meta-analysis by Barone et al. (2019), results in spontaneous-response paradigms are dependent on the type of paradigm used, with higher performance levels being obtained for violation-of-expectation paradigms than for anticipatory looking or interactive paradigms. In line with this finding, infants' performance often did not correlate across different types of paradigms and using different gaze measures (Dörrenberg et al., 2018; Poulin-Dubois & Yott, 2018). There were even meaningful performance differences in the same sample of participants depending on which gaze measure—first look or differential looking score—was analyzed (Burnside et al., 2018). The meta-analysis by Barone et al. (2019) also found that year of publication as well as sample size influenced children's performance: Positive findings of implicit false belief understanding in infants mainly stem from early studies that assessed only small samples of infants (Barone et al., 2019). Another problem with anticipatory looking tasks is high exclusion rates (Schuwerk et al., 2018; Southgate et al., 2007) which often lead to decreases in sample sizes. Because of the single-trial nature of most anticipatory looking false belief tasks[1], a large number of children was excluded from the analysis due to lacking correct anticipatory looking already in the familiarization phase.

Moreover, several anticipatory looking studies found above-chance performance in one false belief condition but not in another false belief condition. In a seminal anticipatory looking study by Southgate et al. (2007), two different false belief conditions (FB1 and FB2) were implemented. In FB1 trials, an agent observed the transfer of a target object from location A to location B. The target object was then removed from the scene in the absence of the agent, leading to the agent's false belief about the target's location. In FB2 trials, the agent was already absent during the transfer of the target object from location A to location B and also missed the removal of the target from the scene leading to their false belief about the target's location. In FB1 trials, the agent thus believed the object to be in the last location (B), whereas in FB2 trials, the agent believed the object to be in the first location (A). In FB1 trials, anticipatory looking at the belief-congruent location coincides with looking at the last location the object was. In FB2 trials, however, looking at the belief-congruent location coincides with looking at the first location the object was. Thus, in combination, the two false belief conditions mutually serve as controls for each other to rule out the possibility that participants solve the task using alternative strategies (Southgate et al., 2007). Compared to FB1 trials, FB2 trials pose increased processing and memory demands due to the added intermediate events. A study comparing FB1 and FB2 performance in the paradigm by Southgate et al. (2007) found that 2- to 4-year-olds performed significantly better in FB1 than in FB2 trials, indicating that FB2 trials might be harder to solve (Grosse Wiesmann et al., 2018). In following replication attempts, researchers were usually only able to replicate the above-chance performance in children and infants in FB1 trials, but not in FB2 trials (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Kulke, von Duhn et al., 2018). This is problematic because only if children pass both FB1 and FB2 trials, their performance can be interpreted as solid evidence for implicit false belief understanding (Baillargeon et al., 2018).

The difficulties in replicating the original findings of false belief-congruent looking in infants and young children are worrisome and call for novel paradigms which can reliably and robustly assess infants' early false belief understanding. A recent promising study by Grosse Wiesmann et al. (2017) addressed the issue that most paradigms rely on only a single trial to measure belief understanding. In their anticipatory looking change-of-location task, each child watched six FB1 and six FB2 trials. In each trial, children anticipated the behavior of an animal agent who was searching for a mouse. Aggregated over trials and over both conditions (FB1 and FB2), the authors found belief-congruent looking in 3- and 4-year-old children (3-year-olds: $M = 54\%$ correct, $SD = 11\%$, $N = 26$; 4-year-olds: $M = 54\%$ correct, $SD = 11\%$; $N = 31$). This finding is important since it indicates that above-chance belief-congruent looking in anticipatory looking tasks can be found in preschool children when aggregating performance over several trials and two false belief conditions. Thus, anticipatory looking tasks might yield evidence for implicit false belief understanding in young children, but single trials might not be reliable enough. To corroborate this assumption, we conducted a large-scale replication of the anticipatory looking paradigm by Grosse Wiesmann et al. (2017). Systematic replication studies are a highly relevant and desirable part of scientific progress and a useful tool for evaluating effects (Frank et al., 2017; Nosek & Errington, 2020a). Particularly in the context of implicit Theory of Mind, replicability of findings has a great importance since findings of robust implicit Theory of Mind would have important theoretical consequences about the onset and acquisition of a Theory of Mind.

In the present longitudinal study, we conducted a large-scale replication attempt of the multi-trial anticipatory looking task by Grosse Wiesmann et al. (2017) in 27-, 36-, and 52-month-old children. First, we aimed at closely replicating the original finding of above-chance false belief performance in 3- and 4-year-old children. While children's mean age in the original study was 39.6 and 51.6 months, the age of the 3- and 4-year-olds in the present study was within the originally tested age range. In addition, we assessed 2-year-olds to explore whether the paradigm is also sensitive towards implicit tracking of beliefs in children below the age of three. Second, we were interested in longitudinal performance trajectories in the age range from 2 to 4 years. As in Grosse Wiesmann et al. (2017), we analyzed relations with children's general language abilities and for the two older age groups relations with explicit false belief understanding.

2 | METHOD

This study was preregistered using the replication recipe by Brandt et al. (2014). The preregistration and the eye tracking data can be found at OSF (https://osf.io/eyvsr/). We report how we determined the sample size, all data exclusions, all manipulations, and all measures in this

[1] Note that also violation-of-expectation and interactive helping paradigms make use of only a single trial to measure children's implicit false belief understanding. Thus, the single-trial nature is a limitation of most implicit tasks.

study. The individual demographic information cannot be shared for data protection reasons.

2.1 | Participants

The present study was part of a large longitudinal research project assessing the role of language in Theory of Mind development from 2 to 4 years. We report data from three measurement points. Children were tested at 27, 36, and 52 months of age. The total sample consisted of $N = 185$ children. From these children, $N = 173$ participated at the age of 27 months ($M_{age} = 27.3$ months, $SD = 0.32$ months), $N = 142$ at the age of 36 months ($M_{age} = 36.2$ months, $SD = 0.44$ months), and $N = 71$ at the age of 52 months ($M_{age} = 53.2$ months, $SD = 0.99$ months). Due to dropouts, only $N = 62$ children participated in all three measurement points. We provide details on how many children participated in how many assessments in the Supplemental Material (S1). Another $N = 24$ children were initially invited but were excluded from the study due to insufficient German skills assessed via a standardized language assessment at 24 months. All children were typically developing at the time of the measurements. They were German natives, or German was the main language in their daily routine. The children were recruited via birth registries from the local registration office and the laboratory's database. The local ethics committee approved the study based on the ethical principles of the European Federation of Psychologists' Associations. $N = 12$ of the 36-month-olds and $N = 28$ of the 52-month-olds participated in the false belief tasks and in the majority of the language tasks remotely via a video-conferencing tool due to the outbreak of the Covid-19 pandemic and subsequent laboratory shutdowns from March 2020 onwards. Control analyses revealed that the children assessed via the video-conferencing tool did not perform significantly better or worse in the language tasks (SETK3: $t(20.17) = -1.59$, $p = .127$; SETK4: $t(67.22) = 0.79$, $p = .435$; both: two-sided Welch tests) or in the explicit false belief tasks (low-inhibition false belief task: $\chi^2(1) < 0.01$, $p > .999$, Chi-square test; Wellman & Liu false belief tasks: $W = 406$, $p = .866$, Wilcoxon test) than the children assessed in the laboratory.

The sample size of this replication attempt was determined by the sample of a larger longitudinal project. We aimed to include as many children as possible from the overall sample. The original study reported data from 26 3-year-olds ($M_{age} = 39.6$ months, range = 36–43 months) and 31 4-year-olds ($M_{age} = 51.6$ months, range = 48–54 months). Comparing this to our study, our sample of 3-year-olds (36 months) is 5.5 times larger. Our sample of 4-year-olds (52 months) is approximately 2.3 times larger. Our sample of 27-months-olds is 5–7 times larger than the (albeit older) samples of the original study. The small telescopes approach by Simonsohn (2015) recommends that replication attempts should have sample sizes large enough to find an effect the original study had 33% power to detect. That means, an approximately 2.5 times larger sample size than the sample of the original study is necessary to detect such effects with sufficiently high power (i.e., with 80% power). While our oldest sample was slightly below this criterion, our two younger samples substantially exceed this recommendation. Moreover, the longitudinal combination of these data sets additionally increases our study's power (Vickers, 2003). We followed the small telescopes approach (Simonsohn, 2015) for determination of our sample size, since it was not possible to determine a reliable size of the effect under investigation (i.e., of implicit false belief understanding measured in anticipatory looking paradigms) based on previous research. Thus, it was not possible to reliably conduct an a-priori power calculation since effect sizes obtained in previous studies vary greatly among individual studies.

2.2 | Tasks and procedure

The anticipatory looking false belief task was performed at the ages of 27, 36, and 52 months. Children's general language abilities were assessed at 24, 36, and 52 months. Assessments of children's explicit false belief understanding were performed at 36 and 52 months. At all measurements, further tasks not relevant for the aims of this study were performed and the experimenters always allowed for flexible breaks between the individual tasks to keep the child motivated and attentive. The session at 27 months lasted approx. 45 min and the remaining sessions lasted between 60 and 90 min. The anticipatory looking task was performed last at 27 months and was preceded by some tasks not relevant for this study. At 36 months, the utilized tasks were preceded by other tasks not relevant for this study and were conducted in the following order: language assessment, low-inhibition false belief task, anticipatory looking task. At 52 months, the language assessment took place on one day together with several other tasks. The anticipatory looking task followed by several other unrelated tasks and by the explicit false belief tasks took place on another day.

Anticipatory looking false belief task

To replicate Grosse Wiesmann et al.'s (2017) task as closely as possible, we implemented the task and eye tracking procedure following advice from the original author. In the task, children's looking behavior was recorded while watching an agent search for a mouse to assess their implicit tracking of others' beliefs. To this end, the children watched 10 familiarization trials and 12 false belief trials (six false belief trials in condition one and six false belief trials in condition two) as in Grosse Wiesmann et al. (2017). The original aminated video clips were used. The true belief trials from the original study were left out for reasons of time constraints in the overarching study and because the main goal was to replicate the above-chance performance in the false belief trials. In the original study, the true belief trials aimed at keeping up children's anticipatory looking by showing an action outcome of the trial which was not provided in the false belief trials. Further, the true belief trials were meant to provide a performance baseline for children's anticipatory looking which we reasoned could also be provided by the familiarization trials.

In each trial, children watched a mouse enter the scene, followed by another agent (one of eight other animals). Subsequently, the mouse entered a y-shaped tunnel and exited it into one of two boxes situated

at the tunnel's arms. The agent witnessed these events. In the familiarization trials (FAM), the agent immediately followed the mouse through the tunnel and opened the box in which the mouse was hiding. The content of the FAM trials should clarify for the participants that it was always the agent's goal to try to find the mouse when entering the tunnel. Once the agent had entered the tunnel, the tunnel's endings and the corresponding boxes were illuminated to elicit children's anticipatory looking. The test phase in which children's anticipatory gaze was recorded commenced 540 ms before this light effect and ended 40 ms before the first part of the animal was visible exiting the tunnel. The test phases of the FAM trials lasted 2500 ms each.

In the false belief trials, the mouse transferred from the box in which it was initially hiding to the other box and then left the scene. Two types of false belief trials were used, and they differed with regards to whether the agent watched the transfer of the mouse (FB1) or not (FB2). In neither the FB1 nor the FB2 trials, the agent watched that the mouse finally left the scene after this transfer. Thus, in both types of false belief trials, the agent held a false belief regarding the mouse's current location. In the FB1 trials, the agent assumed that the mouse was in the final hiding location although it was actually gone. In the FB2 trials, the agent thought the mouse was in the initial hiding location although it was actually gone. Once the mouse had left the scene, the agent re-appeared and entered the tunnel. The agent had tracked the mouse's prior movements with respective head turns. In combination with the events in the FAM trials, the participants should assume that the agent was trying to find the mouse. Next, the tunnel's endings and the boxes were illuminated. The test phase in which children's anticipatory gazes were recorded commenced 540 ms before this light effect and ended 80 ms before the end of the trial as in the original study. In the false belief trials, the agent did not re-appear at either end of the tunnel. The test phases in the false belief trials lasted 2940 ms each.

In Figure 1, the events in the FAM, FB1, and FB2 trials are displayed. All trials were arranged in two different randomizations, of which each child watched only one per measurement point. The trials were spread out over two blocks with a short break in-between. While in the original study, two FAM or true belief trials depicting the outcome of the trial were conducted before the first FB trial, in the present study, only one FAM trial showing the outcome was presented before the first FB trial.

As in Grosse Wiesmann et al. (2017), a Tobii T60 eye tracker (60 Hz sampling rate) was used to record children's eye gaze, and the built-in software Tobii Studio 3.2.2 was utilized to present the stimuli. The participants sat approximately 60 cm in front of the 17-inch TFT flat screen. They sat in a children's car chair that was mounted on a revolving chair or sat on their parent's lap. The parent wore blackened sunglasses if the child completed the task on the parent's lap. The eye tracker was mounted on a flexible monitor arm and was therefore individually adjustable. Before the start of the task, participants completed the built-in five-point-calibration procedure. If fewer than three points had been calibrated correctly, the procedure was repeated for the missing points. The calibration was repeated between the two blocks.

Participants' faces were recorded within Tobii Studio 3.2.2 using a webcam capturing the child from front right. This video recording was used to decide post-hoc for each trial of each participant whether it needed to be dropped from the analysis. Single trials of children had to be excluded from the analysis if children were not paying attention to vital events during the trials such as the mouse's initial hiding location, the transfer of the mouse to the other location, or the time at which the agent or the mouse left the scene. Furthermore, trials were dropped from the analysis if children looked away during or at the start of the test phase. By doing so, we followed the same procedure as in Grosse Wiesmann et al. (2017). One coder watched 84% of all trials at 27 months. A second coder watched all trials at 36 and 52 months as well as the remaining 16% of all trials at 27 months. Further, the second coder double-coded 52% of the coding done by coder one. The two codings were significantly correlated by $r = .84$. The coders decided for each trial whether it had to be excluded based on the above-mentioned pre-defined criteria, which can also be found in the Supplemental Material (S2) and deleted the trials within Tobii Studio. 8.3% of all available trials were not presented as the child refused to watch the second block of the task or the presentation of stimuli was stopped at an earlier point due to the child losing concentration. Of all trials that were actually presented to participants, 15.2% were excluded due to the child missing relevant information (e.g., transfers or hiding locations of the mouse) during the trial or due to the child not looking at the screen throughout the test phase. Another 0.9% of the presented trials were excluded due to technical problems (e.g., faulty calibration). Data from 2.5% of the remaining valid trials are missing as no gaze of the child had been recorded during the test phase. Lastly, scores from 3.5% of the remaining trials with gaze data are missing as the child did not show anticipatory gaze during the test phase but merely looked at other parts of the screen. The different exclusion criteria resulted in 79.0% of all presented trials being used in the analysis[2]. Separate exclusion rates for each measurement point can be found in the Supplemental Material (S3).

After deciding which trials were kept in the analysis, two areas of interest (AOIs) were defined, which were identical throughout all trials. One area covered the upper right corner and one the upper left corner of the screen. The corresponding squares were defined horizontally by a tangent to the circle of the light effect and vertically by a perpendicular line in the middle between the vertical tangent to the circle of the light effect and the symmetry axis of the tunnel. These two AOIs were identical to the ones used in Grosse Wiesmann et al. (2017). An overview of the two AOIs can be found in the Supplemental Material (S5).

The raw gaze data exported from Tobii Studio was preprocessed within R 3.4.3 (R Core Team, 2020). See the Supplemental Material (S4) for a detailed data processing script developed in agreement with the original author. Next, for each trial, a first fixation score was calculated as in Grosse Wiesmann et al. (2017). A score of 1 was assigned for trials in which the first gaze was shifted to the correct AOI. A score of 0 was assigned for trials in which the first gaze was shifted to the incorrect AOI. Furthermore, a longer look score was created as in Grosse Wiesmann et al. (2017): If the child looked longer at the correct AOI,

[2] Note that the above-described percentages due not sum up to 21% due to different base rates the percentages depend on.

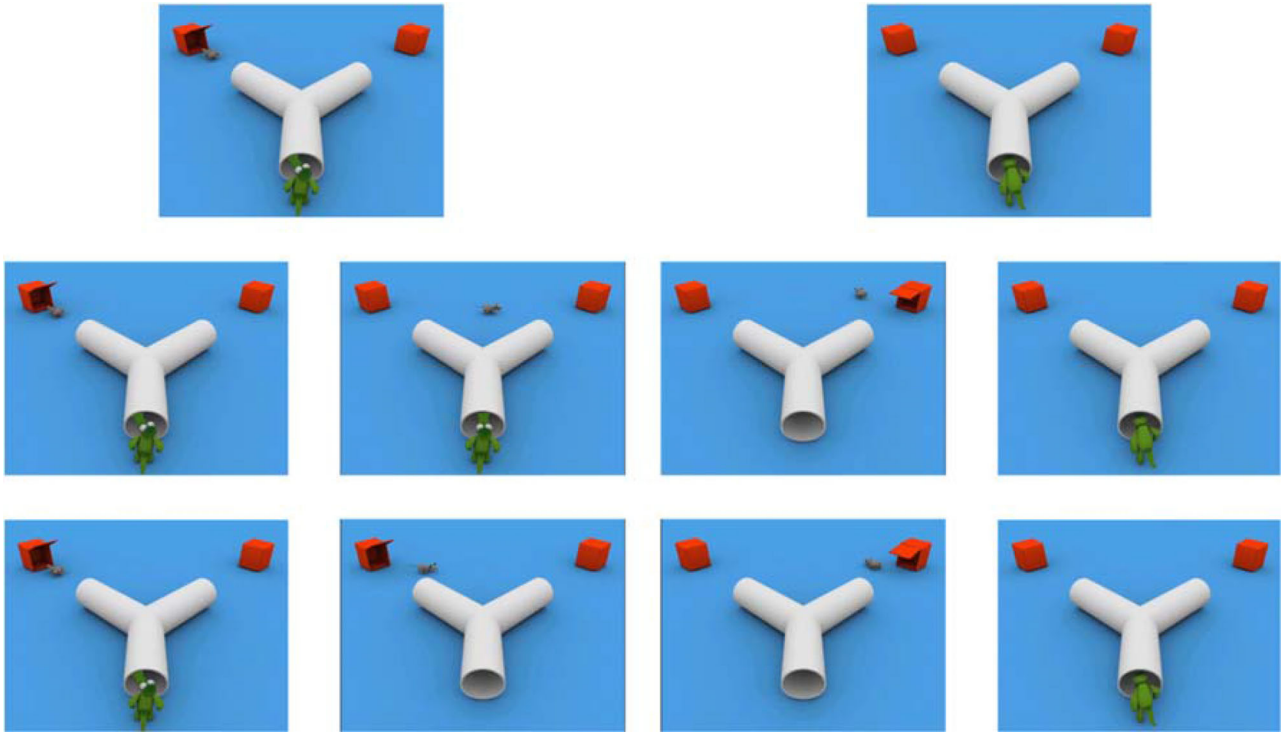

FIGURE 1 Stimuli: Still frames from familiarization trials (first row), false belief condition 1 trials (second row), and false belief condition 2 trials (third row)

a score of 1 was assigned and if the child looked longer at the incorrect AOI, a score of 0 was assigned. If the looking durations were equal for both AOIs, a score of 0.5 was assigned. A mean of first fixations and longer looks across all FAM, FB1, or FB2 trials was calculated to determine children's FAM, FB1, and FB2 first fixation and longer look scores. For the replication attempt, the score used by Grosse Wiesmann et al. (2017) was calculated as the mean of the average first fixation score and the average longer look score for all three conditions for each child. In the Supplemental Material (S6), we also report all main analyses using the average differential looking score (DLS).

Explicit false belief understanding

As measures of children's explicit false belief understanding, we conducted the change-of-location low-inhibition false belief task by Setoh et al. (2016) at 36 months and the two standard explicit false belief tasks from the Theory of Mind scale by Wellman and Liu (2004) at 52 months.

Low-inhibition false belief task

The low-inhibition false belief task was only conducted at the age of 36 months following the procedure described in Setoh et al. (2016). In this task, a typical change-of-location story was presented using a picture book. In the story, the protagonist Lilli finds an apple in a bucket and transfers the apple to a basket. While Lilli is outside playing with a ball, her brother finds the apple and takes it away. When Lilli returns, children are prompted to answer the question "Where will Lilli search for her apple?" by pointing either at the picture of the basket or at the picture of the bucket. This question format was practiced twice throughout the story to familiarize children with 'where'-questions. Children's reply to the final test question was scored as correct (basket) or as incorrect (bucket). As in Setoh et al. (2016), children were excluded from the task if they failed to answer the practice questions correctly. In the remote assessments, the picture book was presented picture-by-picture on the computer screen in front of the child using the screen-sharing mode. Parents reported children's pointing gestures in response to the practice and test questions if these were not clearly perceptible for the experimenter.

Standard explicit false belief tasks

At the age of 52 months, only the two explicit false belief tasks from the Theory of Mind scale by Wellman and Liu (2004) were conducted to measure explicit false belief understanding. A sum score of both tasks was used.

In the *contents false-belief* task, children were shown a Smarties box and were asked to guess the content of the box. Once the children had guessed that the box contains Smarties, the true content (a piglet figurine) was revealed. The piglet was put back into the box, then, children were asked to name the true content as a memory control. Next, the figurine Lucas was presented, and children were told that Lucas had never seen the content of the box. Then children were asked the test question ("What does Lucas think what is inside the box? Smarties or a piglet?") and the control question ("Has Lucas looked inside the box before?").

Children were credited with one point if they answered both questions correctly. For the remote assessments, the Smarties tube, the piglet, and the figurine Lucas were presented to the child via the webcam of the computer and the child replied verbally to the experimenter's questions.

In the *explicit false-belief* task, children were shown a picture of a backpack and a picture of a closet and they were told that the figurine Paul is searching for his gloves which could be in the backpack or in the closet. Next, children were told that Paul's gloves are really in his backpack, but that Paul thinks that they are in his closet. Then, children were asked where Paul will search for his gloves (test question) and where Paul's gloves really are (control question). Children were credited with one point if they answered both questions correctly. For the remote assessments, the picture of the backpack and the closet as well as the figurine Paul were presented to the child via the webcam of the computer. The child replied verbally to the experimenter's questions.

Language assessment

As assessments of children's general language abilities, the language development test for 2-year-olds (*Sprachentwicklungstest für Zweijährige;* Grimm et al., 2000; SETK 2) was conducted at the age of 24 months and the language development test for 3- to 5-year-olds (*Sprachentwicklungstest für Drei- bis Fünfjährige*; Grimm et al., 2015; SETK 3–5) was conducted at the age of 36 and 52 months.

SETK 2

The SETK 2 consisted of two language comprehension and two language production subtasks. According to the age-specific norm table, the obtained raw values from each of these subtasks were transformed into standardized T-values. A mean of all four subtasks was used as an index of children's general language abilities.

SETK 3–5

At 36 months, children's encoding of semantic relations, their comprehension of sentences, their morphological rule formation, and their phonological working memory were assessed using the corresponding age-adequate subtasks. At 52 months, children's language memory, their language comprehension, and their morphological rule formation were assessed. All obtained raw values from each subtask were transformed into standardized T-values according to the age-specific norm table. At 36 and at 52 months, a mean of all corresponding subtasks was used as an index of children's general language abilities. The comprehension of sentences and language comprehension subtasks were always conducted in person since they required children to manipulate certain objects (pencils, buttons, etc.). The remaining tasks were conducted remotely for some of the children (see section Participants). For the encoding of semantic relations and morphological rule formation tasks, the stimulus materials were presented on the child's computer via the screen-sharing mode. For the phonological working memory task, the stimulus material was held into the webcam of the computer. The language memory tasks only required children to repeat sentences and words after the experimenter.

2.3 | Statistical analysis

All data preprocessing and data analysis were conducted in R 3.4.3 (R Core Team, 2020). Two-tailed testing and a significance level of .05 was used for all analyses. If not indicated otherwise, the original score by Grosse Wiesmann et al. (2017) which is a combination of the first fixation and the longer look score was used for data analysis. For the analysis of relations between general language and performance in the implicit false belief task, equivalence tests for correlations (Lakens et al., 2018) were used to corroborate the equivalence of the relation between general language and implicit false belief. In contrast to null-hypothesis significance testing, equivalence testing can be used to investigate "whether an observed effect is surprisingly small, assuming that a meaningful effect exists in the population" (Lakens et al., 2018, p. 259). Within this procedure, two one-sided *t*-tests are performed to be able to reject the null hypothesis that there is an effect at least as extreme as a pre-defined smallest effect size of interest. The absence of a meaningful effect can then be supported. Following one recommendation by Lakens et al. (2018), we determined the smallest effect size of interest such that we had at least 80% power to find it given our sample sizes.

Violin plots were created to visualize the distribution of the data within each measurement point and condition. Violin plots are similar to box plots but also depict the probability density of the data to represent the data distribution. The probability density function is calculated by a Kernel density estimator in a way such that the obtained function fits the observed data well. The thicker sections of the plot indicate a higher probability that members of the population take on a value in this range, whereas the thinner sections stand for a lower probability that members of a population fall into this value range.

3 | RESULTS

3.1 | Descriptive statistics and control analyses

In Table 1, descriptive statistics of the anticipatory looking false belief task at all three measurement points can be found. For a graphical display of the data distribution at each measurement point in each condition, see Figure 2. Independent samples *t*-tests on performance in the anticipatory looking false belief task were performed to rule out possible gender effects. No effects of gender were observed (all *p*-values > .05). Performance in the language tests and the explicit false belief tasks is displayed in Table 2.

3.2 | Confirmatory analyses

In our confirmatory analyses, we mirrored the analysis plan of the original study for a close comparison. For the anticipatory looking task, chance performance lay at 0.50 for all measurement points. First, children's performance in the FAM trials was analyzed. They performed above chance level at all three measurement points using one-sample

TABLE 1 Descriptive statistics and results of one-sample *t*-tests on performance in the three conditions (FAM, FB1, and FB2) at all three measurement points using the score by Grosse Wiesmann et al. (2017)

Measurement point	Condition	*M*	*SD*	*n*	Test Statistic	*p*-value	Cohen's *d*
27 months	FAM	0.71	0.18	172	$t(171) = 15.51$	$p < .001$	$d = 1.18$
	FB1	0.60	0.22	172	$t(171) = 5.77$	$p < .001$	$d = 0.44$
	FB2	0.42	0.24	170	$t(169) = -4.36$	$p < .001$	$d = 0.33$
36 months	FAM	0.69	0.16	142	$t(141) = 13.71$	$p < .001$	$d = 1.15$
	FB1	0.57	0.20	139	$t(138) = 4.02$	$p < .001$	$d = 0.34$
	FB2	0.43	0.21	141	$t(140) = -4.09$	$p < .001$	$d = 0.34$
52 months	FAM	0.70	0.15	71	$t(70) = 11.06$	$p < .001$	$d = 1.31$
	FB1	0.55	0.20	71	$t(70) = 2.18$	$p = .033$	$d = 0.26$
	FB2	0.45	0.21	71	$t(70) = -2.20$	$p = .031$	$d = 0.26$

The Range of Possible Values was 0-1

Abbreviations: FAM = Familiarization trials; FB1 = false belief condition 1 trials; FB2 = false belief condition 2 trials.

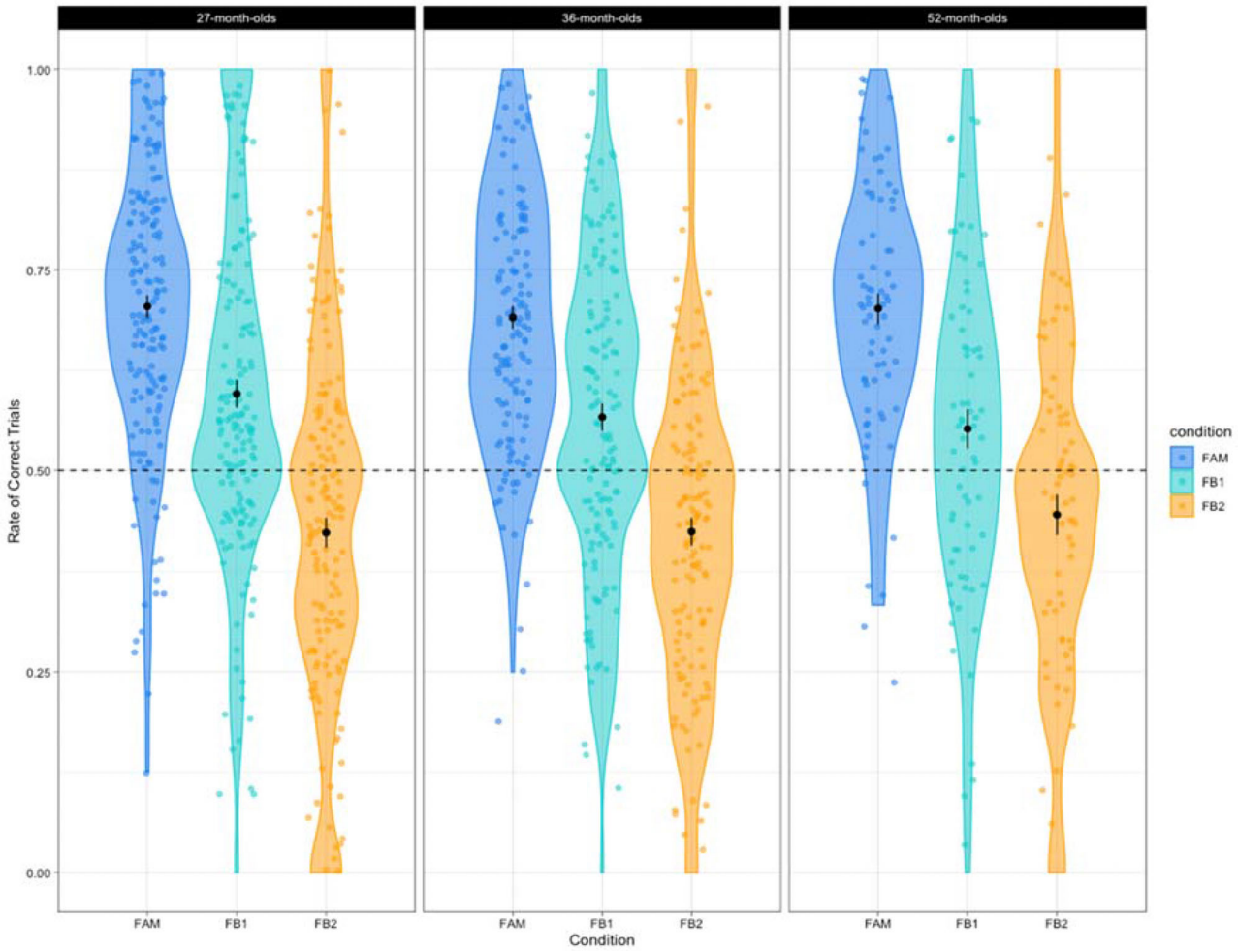

FIGURE 2 Graphical display of children's performance in the anticipatory looking task at all measurement points using the score by Grosse Wiesmann et al. (2017).

Note. The dots represent the mean performance of each participant (horizontally jittered for an illustration without overlaps). The colored areas show the probability density functions in each condition. The black dots depict mean performance across all participants in each condition (black vertical lines illustrate the standard error). The dashed line indicates chance level. FAM = familiarization trials, FB1 = false belief condition 1 trials, FB2 = false belief condition 2 trials

TABLE 2 Descriptive statistics of performance in the language and explicit false belief tasks

Measurement point	Task	*M*	*SD*	*n*	Range of possible values
24 months	SETK2	48.39	6.69	165	0–∞
36 months	SETK3	52.64	6.15	142	0–∞
	low-inhibition false belief task	0.68	0.47	111	0–1
52 months	SETK4	53.62	5.63	70	0–∞
	explicit false belief tasks	1.35	0.69	57	0–2

Abbreviations: SETK2, SETK3, and SETK4 represent performance in the language development test at 2, 3, and 4 years.

t-tests [27 months: $t(171) = 15.51$, $p < .001$; 36 months: $t(141) = 13.71$, $p < .001$; 52 months: $t(70) = 11.06$ $p < .001$].

In the attempt to replicate the finding of Grosse Wiesmann et al. (2017) of above-chance performance in a combined score of FB1 and FB2, we calculated the FB score as the mean of the FB1 and FB2 scores as in Grosse Wiesmann et al. (2017). Then, we conducted three one-sample *t*-tests on this score against the chance level of 0.50. Children performed at chance level using the combined FB score at 27 months ($M = 0.51$, $SD = 0.18$, $t(172) = 0.64$, $p = .527$), 36 months ($M = 0.50$, $SD = 0.15$, $t(140) = -0.05$, $p = .961$), and 52 months ($M = 0.50$, $SD = 0.14$, $t(70) = -0.06$, $p = .951$). Thus, we did not replicate the findings of false belief-congruent looking in 3- and 4-year-old children using the anticipatory looking task and the same score in a large sample[3].

In Grosse Wiesmann et al. (2017), no significant performance differences between the two age groups were found. Thus, we also investigated whether children's performance in the anticipatory looking false belief task improved with age by conducting a repeated-measures ANOVA with the within-participant factor measurement point for each of the three conditions. No significant effect of measurement point was found for the FAM trials ($F(2,122) = 0.05$, $p = .955$, $\eta_p^2 < .01$), the FB1 trials ($F(2,120) = 2.74$, $p = .069$, $\eta_p^2 = .04$), and the FB2 trials ($F(2,120) = 0.65$, $p = .523$, $\eta_p^2 = .01$). Concluding, in this regard, we replicated the result by Grosse Wiesmann et al. (2017) that children's performance did not change with age.

3.3 | Post-hoc analyses

To investigate possible reasons for not replicating the original study's finding, we investigated whether children performed significantly above chance level in FB1 and FB2 trials separately at each measurement point. The results can be found in Table 1: Children performed significantly above chance level in the FB1 trials at all measurement points. Yet, in contrast to our assumptions, children performed significantly below chance level in the FB2 trials at all measurement points.

[3] Note that these *p*-values remain non-significant when adjusting them for one-tailed analyses.

TABLE 3 Post-hoc dependent-samples *t*-tests on performance in the different conditions at all three measurement points

Measurement point	Compared conditions	Test Statistic	*p* – value	Cohen's *d*
27 months	FAM – FB1	$t(168) = 5.05$	$p_{adj} < .001$	$d = 0.39$
	FAM – FB2	$t(168) = 11.48$	$p_{adj} < .001$	$d = 0.88$
	FB1 – FB2	$t(168) = 7.38$	$p_{adj} < .001$	$d = 0.57$
36 months	FAM – FB1	$t(138) = 6.46$	$p_{adj} < .001$	$d = 0.55$
	FAM – FB2	$t(138) = 11.89$	$p_{adj} < .001$	$d = 1.01$
	FB1 – FB2	$t(138) = 6.18$	$p_{adj} < .001$	$d = 0.52$
52 months	FAM – FB1	$t(70) = 4.95$	$p_{adj} < .001$	$d = 0.59$
	FAM – FB2	$t(70) = 9.07$	$p_{adj} < .001$	$d = 1.08$
	FB1 – FB2	$t(70) = 2.98$	$p_{adj} = .012$	$d = 0.35$

Abbreviations: FAM = Familiarization trials; FB1 = false belief condition 1 trials; FB2 = false belief condition 2 trials; p_{adj} = *p*-values adjusted for multiple testing using Bonferroni correction.

To test whether children's performance differed between FB1 and FB2 trials, we conducted three repeated-measures ANOVAs with condition as the within-participants factor[4]. The Mauchly test revealed a violation of the sphericity assumption at 36 months. Thus, the Greenhouse-Geisser correction was used to adjust the degrees of freedom for this analysis. Performance differed significantly among the three different conditions at 27 months ($F(2,336) = 74.95$, $p < .001$, $\eta_p^2 = .31$), 36 months ($F(1.92,264.37) = 76.13$, $p < .001$, $\eta_p^2 = .36$), and 52 months ($F(2,140) = 33.31$, $p < .001$, $\eta_p^2 = .32$). Post-hoc paired-sample *t*-tests revealed significant results for all three pairwise comparisons at 27 months (all *p*-values < .001), 36 months (all *p*-values < .001), and 52 months (all *p*-values < .05). All results of the pairwise comparisons are displayed in Table 3.

[4] As suggested in the review process, we also calculated a more comprehensive 3×3 ANOVA investigating simultaneously the effects of measurement point and condition to fully exploit our longitudinal design. In doing so, we found the same pattern of results (see S7). We did not initially plan to conduct such an ANOVA since the rationale of the study was first to follow the analysis play in Grosse Wiesmann et al. (2017) (including checking for effects of age) and only in a second step to investigate possible reasons for not replicating the original finding (such as comparing performance between FB1 and FB2 trials).

3.4 | Comparison of looking durations at the initial and final hiding location

As a further investigation of differential looking patterns depending on the agent's belief in FB1 and FB2 trials, we calculated children's total looking durations at the initial and final hiding location of the mouse, separately for the FB1 and the FB2 condition at each measurement point. Then, we conducted three repeated-measures ANOVAs on the looking durations with the within-participants factors condition (FB1 and FB2) and hiding location[5] (initial and final) to analyze whether children's looking durations at the initial and final hiding location differed dependent on the agent's belief. At all three measurement points, there was no significant interaction between condition and location [27mth: $F(1,172) = 0.38$, $p = .539$, $\eta_p^2 < .01$; 36mth: $F(1,141) = 0.12$, $p = .728$, $\eta_p^2 < .01$; 52mth: $F(1,70) = 0.65$, $p = .422$, $\eta_p^2 = .01$], indicating that children looked longer at the final hiding location than at the initial hiding location [main effect of location: 27mth: $F(1,172) = 78.81$, $p < .001$, $\eta_p^2 = .31$; 36mth: $F(1,141) = 74.69$, $p < .001$, $\eta_p^2 = .35$; 52mth: $F(1,70) = 12.95$, $p < .001$, $\eta_p^2 = .16$] independently of the agent's belief. In general, children had longer looking durations in FB1 than in FB2 trials at 27 and 36 months [27mth: $F(1,172) = 4.27$, $p = .040$, $\eta_p^2 = .02$; 36mth: $F(1,141) = 11.39$, $p < .001$, $\eta_p^2 = .08$] but not at 52 months ($F(1,70) = 1.90$, $p = .173$, $\eta_p^2 = .03$).

In the Supplemental Material, we further provide a between-participants analysis of performance only in the first FB1 and FB2 trial to allow comparison of our findings with the results of other single-trial studies (see S8) - a procedure also adopted in the replication study by Dörrenberg et al. (2018). Moreover, we analyzed children's progression through the task on a trial-by-trial basis. These results can also be found in the Supplemental Material (S9). Lastly, we investigated based on an approach by Anderson and Maxwell (2016) whether the effect obtained in our study is consistent or inconsistent with the effect obtained in the original study and found that the effect was not inconsistent with the original study's findings. This analysis can also be found in the Supplemental Material (S10).

3.5 | Relations between the anticipatory looking false belief task and language

Based on Grosse Wiesmann et al.'s (2017) results and the assumptions of the dual-systems account (Apperly & Butterfill, 2009), we expected to find no relation between children's general language skills and their performance in the anticipatory looking false belief task. To investigate this hypothesis, we ran equivalence tests. We followed an approach described in Lakens et al. (2018) to choose as the smallest effect size of interest one for which we had 80% power to detect it and to set this smallest effect size of interest as the equivalence bounds to test against. This resulted in equivalence bounds of ± 0.21 for correlations between the SETK at 24 months and the anticipatory looking task at 27 months, bounds of ± 0.23 for correlations at 36 months, and bounds of ± 0.32 for correlations at 52 months. As a measure of children's false belief performance in the anticipatory looking false belief task, we again used the mean of the FB1 and FB2 score as in Grosse Wiesmann et al. (2017). We additionally ran analyses with children's performance in the FAM trials. Table 4 shows the results of the equivalence tests and Pearson correlations. The significant results of the equivalence tests indicate that the correlations between language and performance in the anticipatory looking task were equivalent and were not more extreme than the pre-defined equivalence bounds. In line with this, Pearson correlations revealed only non-significant, close-to-zero correlations. This pattern of findings suggests that the true relation between general language and performance in the anticipatory looking task was not more extreme than the pre-defined equivalence bounds.

3.6 | Relations between the anticipatory looking false belief task and explicit false belief

Finally, for the two older age groups, we analyzed relations between explicit false belief understanding and performance in the anticipatory looking task. At 36 months, there was a trend for a positive relation between children's performance in the FB1 trials and performance in the low-inhibition false belief task which closely failed to reach significance ($r(108) = .18$, $p = .067$, point-biserial correlation). Neither performance in the FB2 trials nor performance in the FAM trials was positively related with performance in the low-inhibition false belief task (FB2: $r(109) = -.05$, $p = .631$; FAM: $r(109) = -.00$, $p = .972$; both: point-biserial correlation).

At 52 months, there was a significant positive relation between children's performance in the FB1 trials and the sum of the two standard explicit false belief tasks ($r_S = .29$, $p = .031$, $N = 57$, Spearman's rank correlation). Performance in the FB2 trials and in the FAM trials was not positively related to performance in the standard explicit false belief tasks (FB2: $r_S = -.08$, $p = .571$, $N = 57$; FAM: $r_S = -.04$, $p = .765$, $N = 57$; both: Spearman's rank correlation). Correlations based on only those children who participated in the explicit false belief tasks in the laboratory can be found in the Supplemental Material (S11).

4 | DISCUSSION

In the present study, we attempted to replicate the finding of false belief-congruent anticipatory looking in young children by conducting the multi-trial, anticipatory looking false belief task by Grosse Wiesmann et al. (2017) in a large sample. As in the original study, we found above-chance performance in the familiarization trials in 2-, 3-, and 4-year-olds. However, we did not find the previously reported above-chance performance in either of the three age groups with the combined false belief score used in the original study (an average of performance in two different false belief conditions, FB1 and FB2). Further investigation of the data indicated that all three age groups performed

[5] Note that the final hiding location corresponds to the correct, belief-based location in FB1 trials, and the initial hiding location corresponds to the correct, belief-based location in FB2 trials.

TABLE 4 Results of equivalence tests and pearson correlations between the anticipatory looking false belief task and general language abilities

Relation between	*p*-value of upper equivalence test	*p*-value of lower equivalence test	correlation coefficient	*p*-value correlation
SETK2 and				
FAM 27mth	.032	< .001	.07	.390
FB 27mth	.072	< .001	.10	.210
SETK3 and				
FAM 36mth	.009	< .001	.03	.700
FB 36mth	.008	.001	.03	.748
SETK4 and				
FAM 52mth	.003	.004	−.01	.962
FB 52mth	< .001	.011	−.05	.663

Abbreviations: SETK2, SETK3, and SETK4 represent performance in the language development test at 2, 3, and 4 years; FAM = familiarization trials; FB = mean of false belief condition 1 trials and false belief condition 2 trials.

significantly above chance in FB1 trials but significantly below chance in FB2 trials. As in Grosse Wiesmann et al. (2017), children's performance did not change between 3 and 4 years of age, and importantly there were no age-related differences between 27 and 36 months either. Equivalence testing corroborated the finding by Grosse Wiesmann et al. (2017) that general language abilities were not related to children's performance in the anticipatory looking task, but there were tendencies for cross-sectional relations between performance in one of the two implicit false belief conditions (FB1) and explicit false belief reasoning.

The present study assessed children's implicit belief-tracking abilities using an anticipatory looking paradigm. We attempted to replicate the original study's finding of false belief-congruent looking in 3- and 4-year-olds in this multi-trial paradigm. We closely followed the data collection and data preparation procedure and measures described in Grosse Wiesmann et al. (2017) and utilized the same stimuli apart from two deviations: First, the true belief trials from the original study which intended to keep up action anticipation were left out due to time constraints and second, the first false belief trial was only preceded by one familiarization trial. Despite using very similar procedures, we did not find above-chance false belief-congruent looking in 27-, 36- and 52-month-old children. This finding is unlikely to be due to children not grasping the story presented in the task since children performed well above chance level in the familiarization trials, requiring simple goal-based action predictions. This above-chance performance indicates that children understood the agent's goal-directed behavior.

In the false belief trials, children additionally needed to consider that the agent held a false belief about the target's location when predicting the agent's actions. Further analyses on the false belief data yielded false belief-congruent looking even at the age of 27 months but only in the false belief condition FB1. In the other false belief condition, FB2, all age groups performed significantly below chance level. This pattern of findings resembles the results of other recent anticipatory looking studies (Dörrenberg et al., 2018; Kampis et al., 2021; Kulke, Reiß et al., 2018, Study 2b). In FB1 trials, the agent observed the displacement of the target object from location A to location B and was only absent while the target left the scene. In FB2 trials, however, the agent was already absent during the displacement of the target. Our results indicate that children might not have taken into consideration that the agent did not watch the target's transfer in the FB2 trials. Rather, they mostly looked at the last place where they themselves observed the mouse going, neglecting that the agent did not have this information. Many researchers argue that above-chance performance in both FB1 and FB2 trials is required to conclude that participants engaged in implicit false belief reasoning (Baillargeon et al., 2018; Southgate et al., 2007).

While the pattern of our results is comparable to the original study (Grosse Wiesmann et al., 2017), only in the large sample that we collected, the within-participant differences in FB1 and FB2 performance became pronounced enough to suggest that FB1 and FB2 trials might be processed differently. Our finding that children looked longer at the target's final than at the target's initial hiding location independent of the false belief condition suggests that children in all three age groups treated both false belief conditions equally. Thus, children might have neglected the absence of the agent during the last transfer in the FB2 trials, leading to above-chance performance in FB1 but below-chance performance in FB2 trials and overall longer looking durations at the final hiding location. This finding demonstrates the importance of large enough samples to find such performance differences with sufficient power.

Children's successful performance in FB1 trials could therefore also be explained by applying a strategy such as 'looking at the last location the target was at'. In other replication attempts of the anticipatory looking task, often chance performance in FB2 trials and low performance levels in the familiarization trials were observed. According to the original authors, this contradicts the idea that infants follow a last location strategy (Baillargeon et al., 2018). In our study, we found high success rates on FAM trials and below-chance performance in FB2 trials. However, we also observed positive correlations of explicit false belief understanding with FB1 performance, but not with FAM

performance, in the older two age groups. This provides a tentative indication that success in FB1 trials might be related with succeeding in a mental state reasoning task and therefore might tap a similar skill. Together with performance patterns observed in other replication studies, it seems unlikely that above-chance performance in FB1 trials can be solely explained by the child applying non-mentalistic behavioral rules (Baillargeon et al., 2018). However, without suitable control conditions, this possibility cannot be ruled out.

Not finding evidence for belief-congruent looking in FB2 trials is well in line with previous research using anticipatory looking false belief tasks (Baillargeon et al., 2018; Grosse Wiesmann et al., 2018; Kampis et al., 2021; Poulin-Dubois et al., 2018; Schuwerk et al., 2018). A direct comparison of FB1 and FB2 performance in our sample yielded that participants in all three age groups performed significantly better in FB1 than in FB2 trials, which constitutes a conceptual replication of Grosse Wiesmann et al.'s (2018) finding with a different paradigm. Further, as opposed to the FB1 trials, performance in the explicit false belief tasks was not related with performance in the FB2 trials. Structural features of the FB2 trials might explain the performance discrepancy between FB1 and FB2 trials. While in FB1 trials, children indeed only need to remember that the agent observed the target at the last location before the final displacement of the target, in FB2 trials, they need to keep track of an additional transfer before the displacement while at the same time having to remember the initial location. Thus, it is possible that children looked at the last location they saw the target going, because they incorrectly remembered that the agent had also observed these actions. Consequently, FB2 trials might draw more heavily on working memory, attention capacity, and inhibitory skills while tracking the target's actions and representing the agent's belief (Baillargeon et al., 2018; Grosse Wiesmann et al., 2018). Also, Senju et al. (2010) argue that participants must maintain the agent's epistemic state longer in FB2 than in FB1 trials which makes this condition more challenging. The finding that heightened cognitive load in a dual-task design hindered implicit false belief processing in adults (Schneider et al., 2012) are in line with this interpretation and indicate that even low-level, implicit processing of beliefs to some extent requires executive functions. In our study, we measured FB2 performance longitudinally and found no age-related improvement of performance restricting the argument that young children's memory limitations hindered successful FB2 performance. However, executive functions and working memory capacity still develop beyond the age of 4 years (e.g., Evers, 2019; Garon et al., 2008) such that the requirements of the FB2 tasks might still have been too challenging for the 52-month-olds. Concluding, FB2 trials may be a less reliable measure of implicit tracking of beliefs due to the additional demands they impose and may therefore not constitute a suitable control condition to assess false belief tracking (Baillargeon et al., 2018).

Despite finding only evidence of false belief-congruent anticipatory looking in one type of false belief trials (FB1) but not in the other (FB2), we observed that all three age groups performed well above chance level in the familiarization trials. Since these trials required children to understand the protagonist's goal (which was to follow the mouse), children's successful performance in these trials might indicate their ability to perform goal-based action predictions.

The high number of failed replications (Kulke & Rakoczy, 2018) of implicit false belief tracking in early childhood from the past years cast doubt on the view that implicit false belief understanding precedes explicit false belief understanding by about 2 years. Proponents of this view base their theory on findings of implicit false belief understanding, which often have been hard to replicate (Barone et al., 2019; Kulke & Rakoczy, 2018). A previous large-scale replication attempt of four anticipatory looking paradigms, for instance, did not replicate any one of the paradigms with adults despite applying the original stimuli and procedures (Kulke, von Duhn et al., 2018). This finding calls into question the reliability and validity of anticipatory looking paradigms as measures of implicit false belief understanding (Kulke, von Duhn et al., 2018; Poulin-Dubois et al., 2018). Our study extends this finding to preschoolers by not replicating above chance false belief-congruent looking in a multi-trial task in a large sample. While we do not claim that replication studies are more trustworthy per se, a failure to replicate the original finding with a large enough sample size casts doubt on the reliability of the original findings (LeBel et al., 2018; Poulin-Dubois et al., 2018) since in a proper replication attempt, "outcomes inconsistent with a prior claim would decrease confidence in the claim" (Nosek & Errington, 2020b, p. 2). Further, by means of a simulation study, Oakes (2017) was able to show that increases in sample size in infant looking time research lead to more reliable effect size measures.

Using the multi-trial paradigm, we observed the same pattern of above-chance FB1 and below-chance FB2 performance known from single-trial studies (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Kulke, Reiß et et al., 2018, Study 2b). Yet, in our study, we found tentative positive, concurrent relations between implicit and explicit false belief reasoning which lends some support to the view that there is conceptual-continuity between implicit and explicit mental state reasoning (Baillargeon et al., 2016; Sodian et al., 2020). These relations, however, only emerged for the FB1 condition and not for the FB2 condition. This again indicates that the FB2 condition may not be a suitable measure of implicit false belief understanding.

A limiting factor to our study is the percentage of trials that were excluded from the analysis. More than 20% of all presented trials were excluded due to the child not paying attention during vital moments of the trial, the child looking away during the test phase, or the child failing to anticipate. Parts of these exclusions can be explained by decreased motivation towards the end of the task and are inherent to the task's multi-trial nature and the young age groups we are assessing with it. Nevertheless, due to the paradigm's multi-trial design, we did not need to exclude any participant from the analysis and remained with a sufficient amount of data from each child and measurement point. Moreover, while the anticipatory looking task was administered first at 52 months, it was performed last at the earlier two measurement points due to the design of our underlying longitudinal study. The order of tasks might have influenced performance and motivation in the task differently. Nevertheless, the pattern of performance in all three

conditions was comparable across measurement points, indicating that, even if task order had an effect, its size was negligible. A further limitation lies within the considerable variance observed across the progression of the task (see S9) and the fact that participants contributing only few trials to a specific condition were treated identically to participants contributing all trials of a condition. In the Supplemental Material (S12), we provide information on how many participants contributed less than half of the trials to a condition. A last limiting factor to this study are two deviations from the original study: First, due to time constraints, the true belief trials from the original study were not conducted. These trials were intended to keep up children's motivation and action anticipation by showing the action outcome of the trial. We reasoned that the familiarization trials would also serve this purpose, but it might be possible that omitting the true belief trials caused decreased action anticipation towards the end of the task. Yet, the pattern of performance across the entire task was comparable to performance in only the first FB1 or FB2 trial (see S8) restricting this limitation. Second, in the present study, only one FAM trial depicting the action outcome of the trial was presented before showing the first FB1 or FB2 trial while in the original study two trials with action outcome were shown prior to the first false belief trial. Since children still succeeded in the FB1 trials, it seems unlikely that these procedural changes only affected performance in the FB2 trials.

5 | CONCLUSION

The present study is an attempted large-scale replication of the false belief-congruent looking behavior in young children in the anticipatory looking false belief task by Grosse Wiesmann et al. (2017). We conducted the task in three age groups, of which two fell into the same age ranges as in the original study. Further, we closely followed the original study's data preparation and analysis procedure and utilized the original stimuli and scores. Nevertheless, we did not replicate the finding of overall above chance false belief-congruent looking in either of the three age groups. Separate analyses of the two different false belief conditions (FB1 and FB2) yielded findings in line with previous replication attempts and cast doubt on the usability of FB2 trials as a measure of implicit false belief understanding. Summarizing, our results are in line with, and add to, the growing number of partial or failed replications of implicit false belief understanding. The increasing number of studies finding no clear evidence of implicit false belief understanding calls into question whether anticipatory looking paradigms exist which can robustly assess children's implicit Theory of Mind abilities (Poulin-Dubois et al., 2018).

FUNDING

This work was funded by grant SO213/33-1 from the Deutsche Forschungsgemeinschaft to the last author. The project is part of the research group Crossing the Borders: The Interplay of Language, Cognition, and the Brain in Early Human Development.

DATA AVAILABILITY STATEMENT

The eyetracking data is available at https://osf.io/eyvsr.

DECLARATIONS OF INTEREST

None.

ORCID

Larissa J. Kaltefleiter https://orcid.org/0000-0002-0921-9047
Tobias Schuwerk https://orcid.org/0000-0003-3720-7120

REFERENCES

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

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. https://doi.org/10.1037/a0016923

Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development*, *46*, 112–124. https://doi.org/10.1016/j.cogdev.2018.06.001

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

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

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*, 101350. https://doi.org/10.1016/j.infbeh.2019.101350

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, A., Perugini, M., Spies, J. R., & van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

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. https://doi.org/10.1016/j.cogdev.2017.08.006

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

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12–30. https://doi.org/10.1016/j.cogdev.2018.01.001

Evers, W. F. (2019). *Entwicklung und Struktur der Exekutiven Funktionen im Vorschulalter*. (Doctoral dissertation, Ruprecht-Karls-Universität Heidelberg). https://archiv.ub.uni-heidelberg.de/volltextserver/27306/

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*(7), 878–879. https://doi.org/10.1038/nn1729

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771. https://doi.org/10.1038/nature01861

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Garnham, W. A., & Ruffman, T. (2001). Doesn't see, doesn't know: Is anticipatory looking really related to understanding or belief?. *Developmental Science*, *4*(1), 94–100. https://doi.org/10.1111/1467-7687.00153

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Grimm, G., Aktaş, M., & Frevert, S. (2000). *Sprachentwicklungstest für zweijährige Kinder*. Göttingen, Germany: Hogrefe.

Grimm, G., Aktaş, M., & Frevert, S. (2015). *Sprachentwicklungstest für dreifünfjährige Kinder*. Göttingen, Germany: Hogrefe.

Grosse Wiesmann, C., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018). Longitudinal evidence for 4-year-olds' but not 2-and 3-year-olds' false belief-related action anticipation. *Cognitive Development*, *46*, 58–68. https://doi.org/10.1016/j.cogdev.2017.08.007

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. https://doi.org/10.1111/desc.12445

Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2020). Two systems for thinking about others' thoughts in the developing brain. *Proceedings of the National Academy of Sciences*, *117*(12), 6928–6935. https://doi.org/10.1073/pnas.1916725117

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

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

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct replication attempt of Southgate, Senju and Csibra (2007). *Royal Society Open Science*, *8*(8), 210190. https://doi.org/10.1098/rsos.210190

Kloo, D., Sodian, B., Kristen-Antonow, S., Kim, S., & Paulus, M.(2021). Knowing minds: Linking early perspective taking and later metacognitive insight. *British Journal of Developmental Psychology*, *39*(1), 39–53. https://doi.org/10.1111/bjdp.12359

Kulke, L., & Rakoczy, H. (2018). Implicit Theory of Mind – An overview of current replications and non-replications. *Data in Brief*, *16*, 101–104. https://doi.org/10.1016/j.dib.2017.11.016

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018a). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, *46*, 97–111. https://doi.org/10.1016/j.cogdev.2017.09.001

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018b). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Nosek, B. A., & Errington, T. M. (2020a). The best time to argue about what a replication means? Before you do it. *Nature*, *583*, 518–520. https://doi.org/10.1038/d41586-020-02142-6

Nosek, B. A., & Errington, T. M. (2020b). What is replication?. *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186

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

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep?. *Science*, *308*(5719), 214–216. https://doi.org/10.1126/science.1111656

Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science*, *21*(4), e12600. https://doi.org/10.1111/desc.12600

Poulin-Dubois, D., Azar, N., Elkaim, B., & Burnside, K. (2020). Testing the stability of theory of mind: A longitudinal approach. *Plos One*, *15*(11), e0241721. https://doi.org/10.1371/journal.pone.0241721

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B., Rafetseder, E., & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet–A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, *48*, 302–315. https://doi.org/10.1016/j.cogdev.2018.09.005

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

R. Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

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

Ruffman, T., & Perner, J. (2005). Do infants really understand false belief?: Response to Leslie. *Trends in Cognitive Sciences*, *9*(10), 462–463. https://doi.org/10.1016/j.tics.2005.08.001

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology*, *80*(3), 201–224. https://doi.org/10.1006/jecp.2001.2633

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. https://doi.org/10.1177/0956797612439070

Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*, *5*(5), 172273. https://doi.org/10.1098/rsos.172273

Scott, R. M. (2017). The developmental origins of false-belief understanding. *Current Directions in Psychological Science*, *26*(1), 68–74. https://doi.org/10.1177/0963721416673174

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. https://doi.org/10.1177/0956797611411584

Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., Osanai, H., & Csibra, G. (2010). Absence of spontaneous action anticipation by false belief attribution in children with autism spectrum disorder. *Development and Psychopathology*, *22*(2), 353–360. https://doi.org/10.1017/S0954579410000106

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences*, *113*(47), 13360–13365. https://doi.org/10.1073/pnas.1609203113

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition*, *38*(1), 1–12. https://doi.org/10.1016/0010-0277(91)90020-5

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Sodian, B., Kristen-Antonow, S., & Kloo, D. (2020). How does children's theory of mind become explicit? A review of longitudinal findings. *Child*

Development Perspectives, *14*(3), 171–177. https://doi.org/10.1111/cdep.12381

Sodian, B., Licata, M., Kristen-Antonow, S., Paulus, M., Killen, M., & Woodward, A. (2016). Understanding of goals, beliefs, and desires predicts morally relevant theory of mind: A longitudinal investigation. *Child Development*, *87*(4), 1221–1232. https://doi.org/10.1111/cdev.12533

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

Surian, L., & Franchin, L. (2020). On the domain specificity of the mechanisms underpinning spontaneous anticipatory looks in false-belief tasks. *Developmental Science*, *23*(6), e12955. https://doi.org/10.1111/desc.12955

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. https://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. *British Journal of Developmental Psychology*, *30*(1), 172–187. https://doi.org/10.1111/j.2044-835X.2011.02067.x

Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology*, *3*(1), 1–9. https://doi.org/10.1186/1471-2288-3-22

Wang, L. U., & 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. https://doi.org/10.1111/mila.12099

Wang, B., Low, J., Jing, Z., & Qinghua, Q. (2012). Chinese preschoolers' implicit and explicit false-belief understanding. *British Journal of Developmental Psychology*, *30*(1), 123–140. https://doi.org/10.1111/j.2044-835X.2011.02052.x

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

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. https://doi.org/10.1111/1467-8624.00304

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

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Kaltefleiter, L. J., Schuwerk, T., Wiesmann, C. G., Kristen-Antonow, S., Jarvers, I., & Sodian, B. (2022). Evidence for goal- and mixed evidence for false belief-based action prediction in two- to four-year-old children: A large-scale longitudinal anticipatory looking replication study. *Developmental Science*, e13224. https://doi.org/10.1111/desc.13224