A two-lab direct replication attempt of Southgate, Senju, & Csibra (2007)

Kampis, D.*[1], Kármán, P.[2], Csibra, G.[2,3], Southgate, V.[1], & Hernik, M.[4]

[1] University of Copenhagen, Denmark
[2] Central European University, Budapest, Hungary
[3] Birkbeck, University of London
[4] UiT The Arctic University of Norway

* corresponding authors: dk@psy.ku.dk, mikolaj.l.hernik@uit.no

Abstract

The study by Southgate, V., Senju, A., and Csibra, G. (Southgate et al., 2007) has been widely cited as evidence for false-belief attribution in young children. Recent replication attempts of this paradigm have yielded mixed results: several studies were unable to replicate the original finding, raising doubts about the suitability of the paradigm to assess non-verbal action prediction and Theory of Mind. In a preregistered collaborative study including two of the original authors, we tested 160 24- to 26-month-olds across two locations using the original stimuli, procedure, and analyses as closely as possible. We found no evidence for action anticipation: only roughly half of the infants looked in anticipation to the location of an agent's impending action when action prediction did not require taking into account the agent's beliefs and a similar number when the agent held a false-belief. These results and other non-replications suggest that the paradigm does not reliably elicit action prediction and thus cannot assess false belief understanding in 2-year-old children. While the results of the current study do not support any claim regarding the presence or absence of Theory of Mind in infants, we conclude that an important piece of evidence that has to date supported arguments for the existence of this competence, can no longer serve that function.

Key words: anticipatory looking, action anticipation, false belief attribution, replication, eyetracking

Introduction

Motivated by the seminal study of Onishi & Baillargeon (Onishi & Baillargeon, 2005), which reported an understanding of false beliefs in 15-month-old infants using a violation-of-expectation paradigm, Southgate, Senju & Csibra (Southgate et al., 2007) designed an eye-tracking study to test whether 25-month-old infants could anticipate actions on the basis of attributed false beliefs. The paradigm included two conditions which, together, controlled for the possibility that infants might correctly predict without considering the other's belief. Data was reported from 20 infants, 10 in each condition, and the majority of infants appeared to correctly anticipate the agent's action in accord with her false-belief. Since publication, this study has been used as one of several key pieces of evidence for the existence of cognitive mechanisms that track others 'beliefs in infancy (Baillargeon et al., 2010; Scott & Baillargeon, 2017)

The design of the study relied on anticipatory looking. It adopted the assumption that infants would be motivated to spontaneously anticipate through which of two windows an agent would reach for a ball. To elicit spontaneous anticipation, the study included two familiarization trials: the first showed an agent watching as a puppet placed a ball in one of two boxes, each located beneath one of two windows. Then, a light and sound cue was emitted simultaneously, after which the agent reached through the window above the box where she had seen the ball. Infants then received a second familiarization trial in which the agent observed the ball placed in the other box. If infants are motivated to predict which window the agent would open, the light and sound cues on this second familiarization trial would prompt them to look to the window above the box containing the ball before the agent would reach through the window. As the paradigm contained no 'task', and different participants may be motivated to attend to different aspects of the scene, only those infants who made a correct anticipation on this second familiarization trial were included in the analyses of anticipatory looking in the subsequent test trial. The reason for this criterion was that in order to interpret the data from the false-belief trial, evidence was needed for each infant that they were motivated and able to make a correct action prediction.

Due to the ease of eye-tracking with preverbal children and its ostensible immunity to factors that might vary across different laboratories, this paradigm has been employed extensively across different ages, populations and species. In later studies, either the original stimuli or versions of, have elicited belief-based action prediction in neurotypical adults and school-aged children, but not in adults and children diagnosed with ASD (Schuwerk et al., 2016; Senju et al., 2009, 2010). In other reports, anticipatory looking revealed differences in belief-based action prediction between typically-developing children and children at-risk for autism (Gliga et al., 2014) and deaf children growing up with impoverished language exposure (Meristo et al., 2012; Meristo & Strid, 2020). Versions with even younger infants revealed belief-based action anticipation in 18-month-olds (Senju et al., 2011; Thoermer et al., 2012). Most recently, new versions of the

paradigm have revealed belief-based action prediction in apes (Kano et al., 2019; Krupenye et al., 2016) and monkeys (Hayashi et al., 2020).

However, in recent years, there have also been a growing number of papers reporting non-replications of belief-based action anticipation in toddlers (Schuwerk et al., 2016), children (Dörrenberg et al., 2018; Grosse Wiesmann et al., 2018; Kulke, Reiß, et al., 2018) and even adults (Kulke, Reiß, et al., 2018; Kulke, von Duhn, et al., 2018) in addition to non-replications of other anticipatory looking false-belief tests (Burnside et al., 2018; Kulke, von Duhn, et al., 2018) as well as non-verbal false-belief paradigms using other methods (for an overview, see Baillargeon et al., 2018). These reports are troubling, especially given that the original results were central in contributing to the reassessment of perspective-taking and mental attribution abilities of young children that challenged the long-held consensus that these abilities emerge relatively late in development. The non-replications have also highlighted a more fundamental problem with the paradigm. Specifically, that the paradigm may not reliably elicit the behavior – action prediction – on which measuring attributions of false belief depends (Baillargeon et al., 2018).

Nevertheless, only one of the reported non-replications was, strictly speaking, a direct replication of the original study. Specifically, one of the studies (Experiment 2) reported in Grosse-Wiesmann et al. (2018) was the only one that used the same number of familiarization and test trials, employed the original stimuli, tested approximately the same age group (24-27 months olds, instead of the original 24-26 months age range), and was not run in conjunction with multiple anticipatory looking paradigms. However, the main analyses they reported did not include the original criteria of correct anticipation in the second familiarization trial.

In the current preregistered study, we aimed to carry out a direct replication of the Southgate, Senju & Csibra (2007) study, following the original methods and analyses as closely as possible. The study was conducted with 160 25-month-old infants in collaboration across two locations: Copenhagen and Budapest. The final sample for the main analysis was approximately 2.5 times the original sample, as recommended for replication studies (Simonsohn, 2015)

Methods

The study closely followed the design of the original paradigm, with the exception that (1) more participants were recruited for the sample, (2) the experiment was performed at two different locations (both different from the original), (3) the first look was operationalised as fixation on either a window or a box to approximate the original analyses that were done based on video coding, (4) the sound of a bell was used while the agent was turning away, (5) areas of interest for all gaze analyses and the periods for inclusion criteria based on infants attending to the critical events were specified, (6) a

different brand of eye tracker was used, and (7) the data-processing pipeline was implemented differently (see Dependent measures below and Supplementary information). The details of the data-processing pipeline and the analyses were preregistered (see https://aspredicted.org/blind.php?x=u5zj3k.), and all materials are available at https://osf.io/86wq2/ .

Participants

In the original study 36 infants were tested, and 20 of them were included in the analysis. In this replication study, we aimed to get at least 40 infants who would provide data in the test trials after exclusions. As a pilot study indicated a high attrition rate, we decided to fix the full sample of infants at the number that would produce such a sample even if the exclusion rate was as high as 75%. All exclusion criteria were preregistered (see SI).

In the current study 160 two-year-old infants participated (mean age: 24 months and 24 days; range: 721–808 days; 95 males, 65 females; 124 monolinguals, 35 multilinguals [15% of the time is exposed to at least one other language besides the primary language – among the participants these were: Danish, Hungarian, Icelandic, Swedish, English, Faroese, Dutch, Arabic, French, Romanian, Spanish, Russian, Ukrainian, Portuguese, Italian, German, Chinese, 1 N/A). Eighty infants were tested at the University of Copenhagen (mean age: 24 months and 22 days; range: 721–808 days; 48 males, 32 females; 58 monolinguals, 21 multilinguals, 1 N/A) and 80 infants were tested at the Central European University in Budapest (mean age: 24 months and 25 days; range: 728–788 days; 47 males, 33 females; 66 monolinguals, 14 multilinguals).

Of the 80 participants tested in Copenhagen, 25 infants were included in the main analysis (mean age: 24 months and 20 days, range: 735–784 days; 15 males, 10 females; 16 monolinguals, 8 multilinguals, 1 N/A), and 55 infants were excluded for the following reasons: procedural errors (6), looking at the scene for less than 2 cumulative seconds during the key periods in familiarization (7), looking away during the key periods at test for more than 1.5 consecutive seconds (26), not looking at the scene during the cue in second familiarization trial (1), or not anticipating correctly on the second familiarization trial (15).

Of the 80 participants tested in Budapest, 24 infants were included in the main analysis (mean age: 24 months and 22 days, range: 728–788 days; 10 males, 14 females; 21 monolinguals, 3 multilinguals), and 56 infants were excluded for the following reasons: procedural errors (5), looking at the scene for less than 2 cumulative seconds during the key periods in familiarization (1), looking away during the key periods at test for more than 1.5 consecutive seconds (20), no fixation to a window or box within 2750 ms from cue onset in the test trial (4) or not anticipating correctly on the second familiarization trial (26)

Altogether, 49 participants who passed all the inclusion criteria were included in the main analyses (mean age: 24 months and 21 days, range: 728–788; 25 males, 24 females; 37 monolinguals, 11 multilinguals, 1 N/A). Out of these 49 participants, 31 infants remained in the FB1 condition and 18 in the FB2 condition.

Apparatus

Eyelink1000 Plus eye trackers were used to collect gaze data in both locations. In order for the eye tracker to be able to follow the participants' gaze, infants wore a target sticker on their forehead or cheek. The eye tracker tracked the movement of the participant's right eye. The stimuli were presented on a 17" monitor (Copenhagen) or on a 24" monitor (Budapest) via a PC running Eyelink's Experiment Builder software (SR Research). In the original study, a Tobii 1750 eye tracker with an integrated 17" monitor was used, while the stimuli were presented using the Clearview AVI presentation software.

Procedure

When the child entered the lab, an animation with rotating colourful objects was played on the screen, accompanied by music. This animation was presented during the seating and preparation. Infants were seated on their parents' lap, 60 cm from the screen. The parents either wore opaque glasses or were asked to close their eyes during the entire study. The whole experiment was video recorded for each participant. (Videos were not recorded for 3 participants: 2 from Copenhagen, 1 from Budapest. However, as exclusions and all analyses were performed on gaze data, these infants were included in our sample). After the initial animation, Eyelink's 5-point calibration was carried out for each infant. After successful calibration, the experiment began.

We used the original stimuli (Southgate et al., 2007) with one modification. In the test trials, when the agent turned away, instead of the sound of a ringing phone, the sound of a bell was played. We reasoned that infants in 2019 may not be familiar with the phone sound used in the 2007 version. As in the original study, there were no attention getters between trials.

Each infant was presented with three video clips: two familiarization trials and one test trial. In all three trials, a female agent appeared behind a panel containing two windows, below which were two opaque boxes. All three trials began with a bear puppet coming into view from the bottom of the screen with a colourful ball and then placing this ball into one of the boxes. In the first familiarization trial the puppet hid the ball in the left-hand box, while in the second familiarization the puppet hid it in the right-hand box. After hiding the ball, the puppet left the scene. After this, a sound and light cue (the latter on

the windows) were played for 1 s, and following a 1750 ms delay, the agent reached through the left-hand window. In the first familiarization, she opened the box lid, and retrieved the ball. In the second familiarization trial she reached through the right window and the trial ended at the point where she made contact with the box. As in the original study, infants then saw one of two test trials, depending on condition.

In the False Belief 1 (FB1) condition, infants saw the puppet put the ball first into the left-hand box, then take it out and place it into the right-hand box. Then the sound of a bell was played and the agent turned away as if attending to the source of the sound. While she was turned away, the puppet reappeared, took the ball out of the right-hand box, and left. The bell stopped ringing when the puppet left and the agent turned back towards the child.

In the False Belief 2 (FB2) condition, infants saw the puppet put the ball into the left-hand box and leave. At this point the bell started ringing and the agent turned away. While she was facing away, the puppet reappeared, took the ball from the left-hand box, and transferred it into the right-hand box. Then the puppet retrieved the ball from the right-hand box and left with it. When the puppet left the scene, the bell stopped ringing and the agent turned back.

In both test trials, after the agent turned back towards the child, the same sound and light cue were presented as in familiarization, followed by an approximately 5 second period where the movie froze. Unlike in familiarization trials, the agent remained stationary and did not reach through either window.

After the lab session, parents filled out a basic demographic questionnaire (adapted from the ManyBabies1 project, link: https://osf.io/56fze/). The questionnaire as well as answers for each participant can be found in the supplementary information.

Dependent measures

We defined two different pairs of areas of interest (AOIs, see SI for details on AOIs). The window+box AOIs included both the window and the box on one side of the screen, while the window-only AOIs were defined only around the window on each side.

The time window of interest was defined as the 2750 ms interval starting from the onset of the acoustic and visual cues that indicated the agent's impending action. In the original paper, the time window of analysis is described as 1750 ms from the cue onset and it was said to correspond to the time-lag between the cue onset and window opening in familiarization. In fact this time-lag lasts 2750 ms. After inspecting the original stimuli and data as well as consulting among the authors, we concluded that the length of 1750 ms was reported incorrectly in the original methods and we corrected it to 2750 ms for the current replication. We do, however, also report analyses for the time window lasting 1750 ms from the cue onset (see SI).

For measurement, we used two dependent variables. The primary dependent variable was binary: it expressed whether the first window+box AOI that the participant fixated within the time window of interest was on the correct side. Window+box AOIs were used in order to approximate the operationalization in the original study, where the first *saccade* to a window was coded from movies showing graphic representation of gaze generated with Tobii's Clearview software. Given that saccades were coded from video replay of gaze location, it is possible that the original coding could have included saccades leading through the window towards the box below. We additionally report first look to a window-only AOI in the SI.

The secondary dependent variable was a Differential Looking Score (DLS) calculated for each participant as cumulative fixation time to correct window-only AOI divided by the sum of cumulative fixation times to the two window-only AOIs during the time window of interest. For this analysis we analysed fixations to the window only, as this approximates well the coding of looking times used in the original study. Analysing the DLS measure is preferred to comparing looking durations to the two AOIs because looking durations are not independent from each other. Nevertheless, in the Supplementary Information we also report the same statistical analyses on looking durations as those reported in the original paper.

Results

From the full sample, 70 infants were excluded for procedural errors, inattention or lack of anticipatory fixations at test. Of the 90 remaining infants, 49 (54%) anticipated correctly in the second familiarization trial. This is a somewhat lower proportion than was reported in the original study (20 of 31, 65%).

Our primary dependent measure was whether infants anticipated correctly (i.e., whether after the cue onset, the first window or box on which they fixated was on the side where the agent should have expected to find the ball, given her false belief). Twenty-two of the 49 infants anticipated correctly, a proportion that did not differ from chance (p = .568, two-tailed exact binomial test). Correct anticipations were found in 15 of 31 infants in the FB1 Condition and in 7 of 18 infants in the FB2 condition. The proportion of correct responses did not differ significantly between the conditions (p = .565, two-tailed Fisher's exact test).

Our secondary dependent variable was the DLS score calculated from the time infants looked at one or the other window-only AOI. The mean DLS score was .47 (95%CI [0.36 0.57]; a score above .5 would have indicated longer looking to the correct AOI). A one-sample t-test on the DLS scores against chance (.5) indicated that infants did not look longer at the correct AOI then the incorrect AOI ($t(47) = -0.63$, p = .532). One infant

was not included in this analysis because they fixated first a box but never fixated a window. An independent-samples t-test with condition (FB1 vs. FB2) as a grouping variable showed that DLS was higher in the FB1 (M = .54) than in the FB2 (M = .33) condition (t(35.76) = 2.04, p = .049, 95% CIdiff [0.001 0.41]). This effect is illustrated on Figure 1, and was primarily due to the fact that infants tended to look longer at the right window (which is correct in FB1 and incorrect in FB2, thus leading to a condition effect). DLS was only numerically higher than chance level (.5) in FB1 (t(30) = 0.64, p = .527), and numerically lower than chance in FB2 (t(16) = 2.11, p = .051; both tests are one-sample t-tests on the DLS scores against chance (.5)).

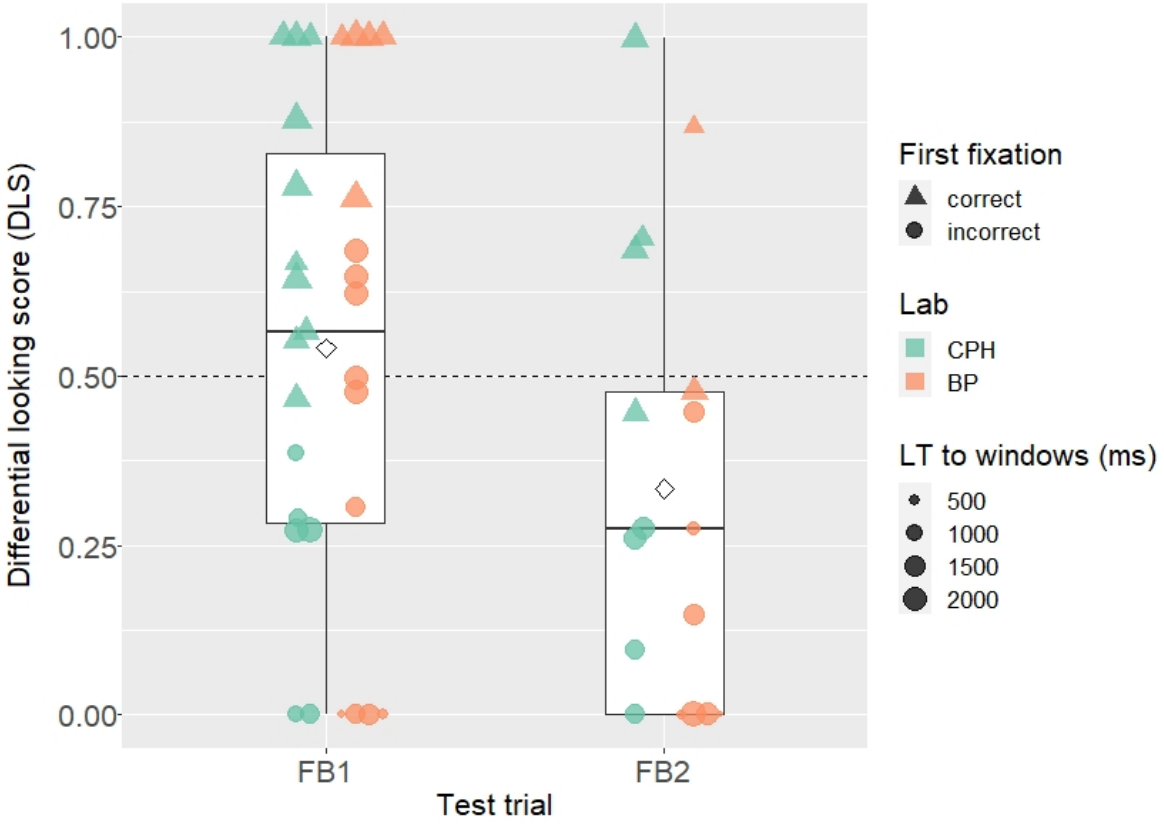

Figure 1. Differential looking score (DLS) in the two conditions, in the group of infants who passed all inclusion criteria, and correctly anticipated in the second familiarization trial. Note that one infant did not provide a DLS score as they did not fixated on a window. Boxplots show median, first and fourth quartile, minimum and maximum values. White diamonds indicate conditions means. Individual data points are plotted per lab: left side of each box plot – Copenhagen, right side – Budapest. Shapes indicate whether the first fixated window or box was on the correct (triangle) or incorrect side

(disc). The size of individual data points scale with the amount of looking time (LT) to both windows contributed by each participant.

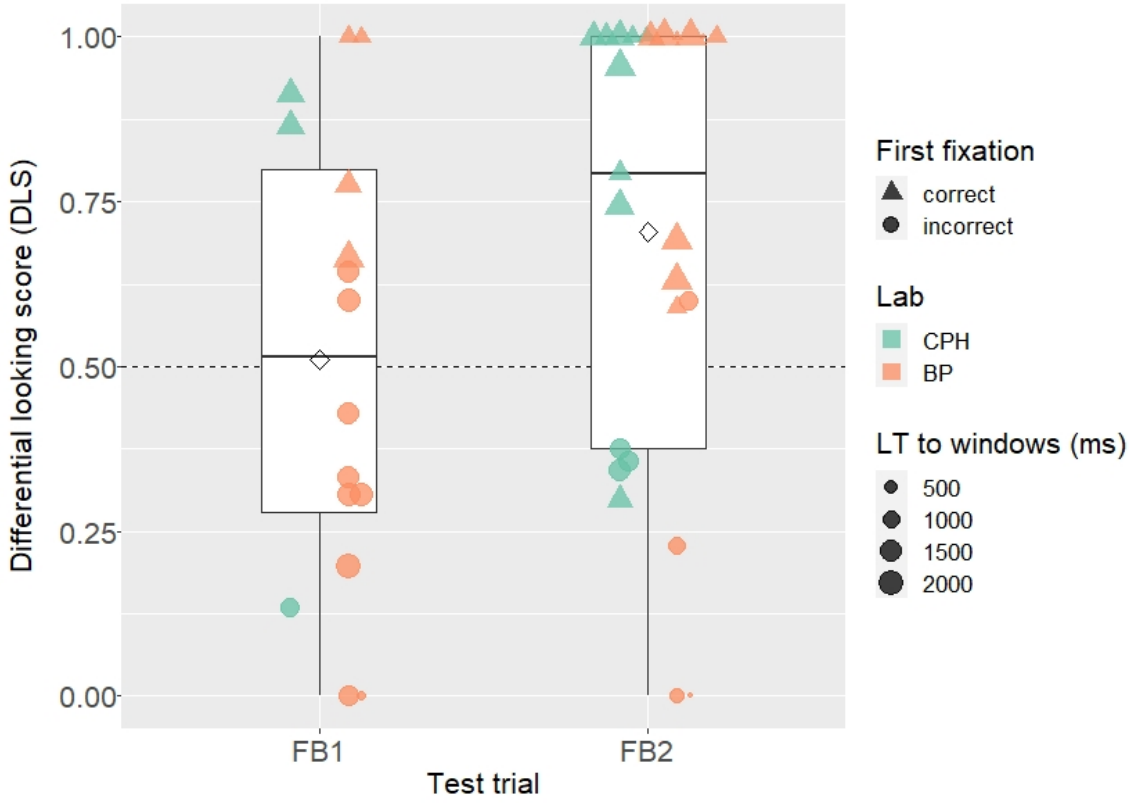

Figure 2. Differential looking score (DLS) in the two conditions, in the group of infants who passed all inclusion criteria, except that they anticipated incorrectly in the second familiarization trial. Boxplots show median, first and fourth quartile, minimum and maximum values. White diamonds indicate conditions means. Individual data points are plotted per lab: left side of each box plot – Copenhagen, right side – Budapest. Shapes indicate whether the first fixated window or box was on the correct (triangle) or incorrect side (disc). The size of individual data points scale with the amount of looking time (LT) to both windows contributed by each participant.

The above analyses only included infants who, on the second familiarization trial, looked first to the correct window or box. We considered post hoc that the difference in DLS between the two conditions could be driven either by a general right-side bias (because it was the location of the ball in the last familiarization trial, or because it was the last location of the ball in test) or by an individual side bias, i.e., infants looking to the same side in two consecutive trials (last familiarization, and test trial). The results of the included infants alone could not differentiate between these alternatives. In a follow-up analysis performed on all infants who passed all other exclusion criteria and in the second familiarization trial produced either the 'correct '(N = 49, included infants) or the 'incorrect '(N = 41, excluded infants) first fixation to window+box AOI, we asked whether either of these tendencies might be visible. We conducted an ANOVA with DLS as a dependent variable and condition (FB1 vs. FB2) and inclusion (included vs. excluded) as between-subjects factors. As a DLS above .5 reflects longer looking to the correct AOI, it indicates longer looking to the right window in FB1 and to the left window in FB2. An overall bias to the right side, therefore, in this extended sample would manifest in a main effect of condition. On the other hand, a potential individual side-bias to look to the same location in the second familiarization and in test trial should manifest in a main effect of inclusion (as exclusion and inclusion correspond to looking left or right in second familiarization, respectively).

This analysis resulted in a significant effect of inclusion ($F(1,85) = 4.77$, $p = .032$), as the included infants had overall higher DLS than the excluded ones. In addition, there was also a significant interaction between condition and inclusion ($F(1,85) = 7.00$, $p = .010$), see Figure 2. Thus overall, in the test trials, infants tended to look to the same side as where they fixated in a window+box AOI first in the preceding familiarization trial, and this pattern was more manifest in the FB2 than in the FB1 condition. This finding indicates an individual side bias. Since this effect is irrelevant for the purpose of our study, we will not discuss it further.

As some methodological choices were implemented slightly differently in this study than in the original, we ran additional pre-registered analyses on the data we collected (See SI). None of these additional analyses yielded substantially different results.

Discussion

This paradigm was originally intended to serve as a non-verbal means of assessing whether young children anticipate others' actions in a way that takes into account their false beliefs. A positive answer to this question would serve as evidence of false belief attribution. However, having sampled 160 25-month-old infants across two sites, we found no evidence for either action anticipation, or belief-based action anticipation. Of the participants who were not excluded for other reasons, only 54% of

infants anticipated the impending action in the second familiarization trial correctly, where anticipation did not require belief attribution. Furthermore, on the critical test trial, where an action could be anticipated by considering the agent's false belief, only 45% of infants anticipated correctly.

Thus, in the present study, we were unable to replicate the original finding of Southgate et al. (2007). This replication attempt, unlike most of the earlier ones (for a discussion see Baillargeon et al., 2018), was very close to a direct replication (see Methods for the exceptions). Where we did deviate from the procedure as described in the original paper, the changes aimed at approximating (e.g., when specifying AOIs), improving (e.g., when pre-defining the critical periods infants should attend to), or clarifying (e.g., by correcting the description of the anticipatory time window to 2750 ms from cue onset) the original methods, and we have no reason to suspect that these differences would systematically account for infants 'failure to anticipate actions in the present study. Overall, we conclude that this paradigm and stimuli do not reliably elicit action anticipation. That is, the paradigm does not elicit the behavior upon which it relies to assess the ability to attribute false beliefs, and therefore it does not provide evidence for this ability in 2-year-old children.

While some follow-up versions of the original paradigm reported high levels of correct anticipation on the second familiarization trial when no belief consideration was required (Low & Watts, 2013; Senju et al., 2009, 2011), several other studies found that participants anticipated either at or even below chance (see Baillargeon et al., 2018 for a review). Why do children (and even adults, see Kulke, Reiß, et al., 2018) not make correct anticipations on familiarization trials? Other paradigms, including those using anticipatory looking, show that infants can generate the kind of goal-based action predictions required to look to the correct location on the second familiarization trial (Brandone et al., 2014; Cannon & Woodward, 2012; Falck-Ytter et al., 2006; Fawcett & Liszkowski, 2012; Kanakogi & Itakura, 2011; but see Ganglmayer et al., 2019). A possible answer to this question is that either this task does not motivate infants to predict others 'actions, or that action prediction in this kind of event does not reliably recruit the kind of eye movements that the paradigm was designed to elicit. An indication that participants may not produce saccade and fixation patterns that clearly reveal their predictions comes from a replication and extension of the original study. Wang & Leslie (2016) included two false belief conditions; a low-demand condition in which the object is removed from the scene in order to reduce the so-called 'pull of the real 'and a high-demand condition in which the object is transferred from the original location to the other location. While both young children and adults showed correct action prediction on the low-demand condition, neither group showed correct action anticipation on the high-demand condition. Assuming that adults should easily be able to make the correct prediction on the high-demand condition, the most likely explanation for their apparent failure is simply that actual prediction is not reflected in their saccade pattern.

A further way in which our findings differed from those of the original study is that we had a higher exclusion rate due to inattention (34% vs. 6%). However, direct comparison of exclusion rates may not be appropriate given the differences in implementation of the exclusion criteria across the two studies. For example, in the original study, looking was coded from movies showing graphic representation of gaze generated with Tobii's Clearview software, whereas in the current study it was measured by exported fixation data. Given that we used the same stimuli as in the original study, it is possible that what was sufficient to motivate toddlers to attend to the stimuli in 2006 is different from what is sufficient to motivate them fourteen years later. Moving forward, based on such considerations, new projects such as the multi-lab collaborative ManyBabies2 will present toddlers with more modern, newly devised stimuli that are designed to be more appealing (Kampis & Schuwerk, 2020).

Besides not providing evidence for goal-based action anticipation, our study also failed to reproduce the original finding according to which most of the 25-month-olds who appeared to anticipate during familiarization also did so when action prediction could have been made from attributed false beliefs. While in the original study, infants were included based on the assumption that they had shown correct anticipation on the second familiarization trial, it should be acknowledged that what was interpreted as correct anticipatory looking may have been randomly distributed exploration or reaction evoked by the lighting up of the windows (20 of 31 infants appeared to anticipate correctly). Thus, it is possible that the original criteria did not select children who were genuinely predicting an action from the goal attributed to the agent. While the majority of those children did then look correctly again on test, the current failure to replicate this pattern with a larger sample indicates that they may have done so by chance. We therefore conclude that the effect reported by Southgate et al. (2007) is not reliable and should not be used as evidence for false belief attribution in children of this age.

Nevertheless, it is important to note here that this, and other replication failures of this paradigm, cannot be used as evidence for the opposite conclusion either. Rather than showing the absence of competence in belief attribution, the fact that infants did not anticipate actions even when they did not have to think about the other's belief renders these current null results uninterpretable. Other paradigms may also support the conclusion that infants and toddlers track others 'attention and build expectations based on what others have or have not seen (Scott & Baillargeon, 2017; Southgate, 2020). Similarly, our failure to replicate the original finding does not nullify the positive results obtained with versions of this paradigm in different populations, such as non-human primates. Thus, the current results should be taken as damning of this particular version of the paradigm with this particular age group, rather than the entire body of evidence for false belief understanding in infants or in other populations. Nevertheless, we conclude that an important piece of evidence that has to date supported the existence of early competence of mental state attribution should no longer serve this function.

References

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., & 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

Brandone, A. C., Horwitz, S. R., Aslin, R. N., & Wellman, H. M. (2014). Infants' goal anticipation during failed and successful reaching actions. *Developmental Science*, *17*(1), 23–34. https://doi.org/10.1111/desc.12095

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

Cannon, E. N., & Woodward, A. L. (2012). Infants generate goal-based action predictions. *Developmental Science*, *15*(2), 292–298. https://doi.org/10.1111/j.1467-7687.2011.01127.x

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

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

Fawcett, C., & Liszkowski, U. (2012). Infants Anticipate Others' Social Preferences. *Infant and Child Development*, *21*(3), 239–249. https://doi.org/10.1002/icd.739

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. https://doi.org/10.1037/a0034146

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

Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., Kato, S., Hori, Y., Nagai, Y., Iijima, A., Someya, T., & Hasegawa, I. (2020). Macaques Exhibit Implicit Gaze Bias Anticipating Others' False-Belief-Driven Actions via Medial Prefrontal Cortex. *Cell Reports*, *30*(13), 4433-4444.e5. https://doi.org/10.1016/j.celrep.2020.03.013

Kampis, D., & Schuwerk, T. (2020, July). *ManyBabies 2: Theory of mind in infancy*. vICIS 2020 (International Congress on Infant Studies), Glasgow/online.

Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature Communications*, *2*, 341. https://doi.org/10.1038/ncomms1342

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, *116*(42), 20904–20909. https://doi.org/10.1073/pnas.1910095116

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

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

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is Implicit Theory of Mind a Real and Robust Phenomenon? Results From a Systematic Replication Study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Low, J., & Watts, J. (2013). Attributing False Beliefs About Object Identity Reveals a Signature Blind Spot in Humans' Efficient Mind-Reading System. *Psychological Science*, *24*(3), 305–311. https://doi.org/10.1177/0956797612451469

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

Meristo, M., & Strid, K. (2020). Language First: Deaf Children from Deaf Families Spontaneously Anticipate False Beliefs. *Journal of Cognition and Development*, *21*(4), 622–630. https://doi.org/10.1080/15248372.2020.1749057

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Schuwerk, T., Jarvers, I., Vuori, M., & Sodian, B. (2016). Implicit Mentalizing Persists beyond Early Childhood and Is Profoundly Impaired in Children with Autism Spectrum Condition. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01696

Scott, R. M., & Baillargeon, R. (2017). Early False-Belief Understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

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

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., White, S., & Frith, U. (2009). Mindblind Eyes: An Absence of Spontaneous Theory of Mind in Asperger Syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

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

Southgate, V. (2020). Are infants altercentric? The other and the self in early social cognition. *Psychological Review*, *127*(4), 505–523. https://doi.org/10.1037/rev0000182

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

Wang, L., & Leslie, A. M. (2016). Is Implicit Theory of Mind the 'Real Deal'? The Own-Belief/True-Belief Default in Adults and Young Preschoolers. *Mind & Language*, *31*(2), 147–176. https://doi.org/10.1111/mila.12099