Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis

Pamela Barone\(^a,b,*\), Guido Corradi\(^a,c\), Antoni Gomila\(^a,b\)

\(^a\)Human Evolution and Cognition Group (EvoCog), UIB, IFISC, Associated Unit to CSIC, Spain
\(^b\)Department of Psychology, University of the Balearic Islands, Palma, Spain
\(^c\)Departmento de Psicología, Facultad de Educación y Salud, Universidad Camilo José Cela, Spain

A B S T R A C T

Evidence obtained with new experimental paradigms has renewed the debate on the development of theory of mind in general and false belief ascription in particular. Namely, several studies contend to prove that infants already have the capacity to attribute false beliefs. The aim of the current meta-analysis is to review and summarize the empirical evidence about spontaneous-response false belief tasks in infants younger than 2 years old. Fifty-six false belief conditions using the violation-of-expectation, the anticipatory looking and interactive paradigms were included in this meta-analysis, including 1469 infants. The role of several moderators was examined, following Wellman et al.'s meta-analysis (2001). Results show that correct performance on spontaneous-response false belief tasks was about 1.76 times more likely than incorrect performance (β = 0.57, 95% CI 0.33; 0.80, \(p < .0001\)). Mediator analyses revealed that (i) year of publication had a significant influence on performance, reducing the average log odds of successful performance (β = -0.11, 95% CI: -0.16; -0.06, \(p < .0001\)); and (ii) correct performance was more likely than incorrect performance when the task was conducted in the violation-of-expectation paradigm (β = 0.75, 95% CI: 0.25; 1.26, \(p = .003\)). However, heterogeneity was high across the studies and the funnel plot revealed an asymmetric distribution suggesting that studies with small effect sizes were not published. These results cast doubt on the alleged robustness of the phenomenon: its effect size decreases as time passes, it seems to depend on the type of paradigm employed, and the variance across studies is not well understood yet.

1. Introduction

In our everyday social interactions, we frequently attribute mental states (e.g., beliefs, desires and intentions) to other people in order to predict or explain their behavior. For example, if you know that Ben wants to drink orange juice, and you know that Ben believes that the orange juice is in the fridge, you can predict that he will search for the orange juice in the fridge. This ability to impute mental states to others (and also to oneself) to make predictions about the future behavior of other agents is called theory of mind (Premack & Woodruff, 1978, p. 515). Competence in theory of mind consists of understanding agents as holders of beliefs and desires, which jointly give rise to intentions and goals, carried out in action.

False-belief ascription was chosen as diagnostic of theory of mind capacity (Bennett, 1978; Dennett, 1978; Harman, 1978) because an agent with a false belief holds an informational state incongruent with reality. Then, this ability requires that an individual understands the situation in the terms of the agent, which are necessarily different from how the individual himself conceives of it. As a result, developmental psychologists devised the false belief task (FBT) to investigate children’s ability to represent another person’s mistaken belief (Wimmer & Perner, 1983).

There are two main versions of the classical FBT: the “change of location” or the “unexpected transfer” version (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983), and the “unexpected contents” variant (Gopnik & Astington, 1988; Hogrefe, Wimmer,
Two additional true belief (TB) conditions were performed. In one of them, the agent was present while the watermelon moved from the green to a yellow box while the other was absent (FB-green condition). In another FB condition, the agent was present while the watermelon moved to the yellow box but then it returned to its original position as soon as the agent disappeared from the scene (FB-yellow condition) 1.

In the unexpected contents task, the child is shown a box whose content is not what the cues depicted on its surface suggest, and is asked, “What do you believe is in the box?” After revealing what is in fact inside, the participant is asked “What did you believe was in the box?” and “What another child will think is in the box when he first encounters it?” To pass the test, the child has to distinguish between her current and former perspective of the situation as well as the character’s perspective.

Correct performance on any version of the FBT indicates that the child understands that the agent will behave according to her beliefs about the reality (i.e., her perspective) which diverge from the reality itself (i.e., these beliefs are false) and from the child’s true beliefs about the situation. Using multiple variations of this basic test, a huge amount of converging evidence supported the conclusion that children consistently pass the FBT by the age of 4 (Wellman, Cross, & Watson, 2001). Interestingly, younger children consistently give the wrong answer in every version of the FBT, responding from their own point of view, not from the agent’s (Apperly, 2011).

However, doubts remained that passing the classical FBT involved additional abilities than just theory of mind, both in terms of linguistic abilities and executive control requirements (Ozonoff & McEvoy, 1994; Russell, Mauthner, Sharpe, & Tidswell, 1991). Further support for this suspicion came from an experiment in which 35-month-old children, who were allowed to answer by looking instead of verbally answering a question, passed the test (Clements & Perner, 1994). More recently, Rubio-Fernández and Geurts (2013), who also used the classical change of location task but required the participants to move the puppet character themselves instead of verbally answering questions, found that 36 months old passed the test. In addition, Setoh, Scott, and Baillargeon (2016) revealed that 30 months old also succeeded in the change of location task if they received two previous practice trials resembling the structure of the test question and if the target object was removed from the scene, so the child was uncertain about the object’s last location. As the structure of the classical FBT is made simpler, in terms of verbal competence and executive control requirements, attribution of false beliefs is found earlier than previously thought.

During the last decade, this suspicion gave rise to a new set of simpler, spontaneous-response FBT, developed in order to find out how early attribution of false beliefs, and therefore, theory of mind, can be said to appear in development (Scott, Baillargeon, Song, & Leslie, 2010). These new experimental paradigms involve violation-of-expectation (VOE) (Onishi & Baillargeon, 2005), and anticipatory looking (AL) measures (Southgate, Senju, & Csibra, 2007), as well as a variety of interactive designs (Buttelmann, Carpenter, & Tomasello, 2009).

The VOE paradigm exploits the fact that when an individual’s expectations are violated, she is surprised and surprise involves looking longer at the spot where the relevant event is taking place. Accordingly, it involves two steps: first, a habituation or familiarization phase, in which the infants are experimentally induced to form context-dependent expectations by being repeatedly exposed to the same event. Secondly, in the test trials, they are presented with either an expected or an unexpected event, given the induced expectation. The prediction is that infants will look longer to unexpected events as a sign of surprise. By experimentally modifying the event, authors try to determine the content of the infants’ expectations formed during the habituation trials (Jacob, 2013).

Onishi and Baillargeon (2005) designed the first spontaneous-response study applying this paradigm in false belief understanding as a variation of the classical change of location task. In the familiarization phase, infants first watched as an actor hid a slice of watermelon in a green box and retrieved it from there twice. Next, a sort of magic change occurred. In one false belief (FB) condition, the watermelon moved by itself (in fact, the agent moved it using a magnet under the table) from the green to a yellow box while the agent was absent (FB-green condition). In another FB condition, the agent was present while the watermelon moved to the yellow box but then it returned to its original position as soon as the agent disappeared from the scene (FB-yellow condition) 1. In the experimental trials, infants watched the actor searching in one of the two boxes. The authors predicted that infants would look for a longer period of time when the agent acted unexpectedly (according to her false belief): if she went to the yellow box in the FB-green condition or to the green box in the FB-yellow condition. After this seminal paper, the VOE paradigm has spread over the field as a measure of belief understanding 2.

On the other hand, other researchers have chosen infants’ first look in the test trial as the dependent variable. This first look generally takes place before the display is shown. This anticipatory looking method is a measure of what an infant expects to happen, given what happened before in his/her presence (Heyes, 2014). Clements and Perner (1994) were the first in using the AL measure in the context of a standard change of location FBT. Instead of directly asking where the agent would look for her toy upon her return, participants heard the experimenter say to himself “I wonder where she’s going to look?” Most 35 month-old children correctly anticipated the spot where the agent would search according to her belief looking first at that location.

Southgate et al. (2007) updated the AL methodology using an eye-tracker and removing any verbal interaction in order to test 25-

---

1 Two additional true belief (TB) conditions were performed. In one of them, the agent was present while the watermelon moved from the green to the yellow box (TB-yellow condition). In the other TB condition, the yellow box moved forward and then back to its original position but there was no change in the watermelon’s location (TB-green condition).

2 In fact, few studies use both FB conditions to study infants’ looking times. Following experiments mostly relied on and built upon the FB-green condition but it is more easily explained by shallow changes in the display (i.e., the agent always searched in the green box during the familiarization trials but she searched in the yellow box in the test; thus, infants looked longer). However, the FB-yellow condition is a more conservative test of belief understanding. Thanks to an anonymous reviewer for pointing this issue out.
month-old infants in a spontaneous-response FBT. They predicted that infants would look in anticipation where an agent with a false belief will search for an object, on the grounds of the visual information previously available to both. Interestingly, the object was not just changed of location but removed from the scene. After Southgate and colleagues' positive results, other authors started to use infants' first look as a dependent measure of false belief understanding.

In addition, a new set of spontaneous-response FBT emerged that took advantage of infants' intentional interaction abilities, such as active helping and referential communication. These tasks typically involve simple language and introduce participants to a scene in which an agent does or does not witness an event. Children are then given a verbal prompt—but one that only indirectly taps their representation of the agent's epistemic state. Authors measure young children spontaneous behaviors in response to the prompt, such as helping and pointing, as a proxy for belief understanding.

In the first FBT using the interactive paradigm, infants watched as an agent placed a toy inside a box, and later as another person switched the toy from one box to a second one while the agent either witnessed the switch (true belief condition) or not (false belief condition). Then, the agent unsuccessfully attempted to open the box that originally contained the toy and infants, who were previously instructed on how to open the boxes, spontaneously helped her. Buehlmann et al. (2009) expected that infants helped the agent to open the other box, where the toy is actually hidden, only in the false belief condition. So far, many other studies have employed infants' helping or pointing behavior to measure false belief ascription.

Each of the previous studies on early false belief understanding provided positive results showing that infants, from 15 months old, can attribute false beliefs to another agent. This new picture of young children tracking another agent's false beliefs gave rise to a paradox in development: on the one hand, infants succeed in different non-verbal and spontaneous versions of the FBT but, on the other hand, children cannot reliably pass the verbal, explicit versions of the FBT until they are 4 years old.

Different theoretical approaches tried to explain this developmental paradox. For instance, advocates of a nativist or an early, full mentalistic capacity account take the results of spontaneous-response FBT at face value, claiming that they prove genuine false belief understanding (Carruthers, 2013; Jacob, 2013; Leslie, 1994; Leslie, Friedman, & German, 2004; Scott & Bialargeon, 2017). They argue that 3 year-olds' difficulties with the explicit FBT are related to processing demands but not to their understanding of beliefs, thus applying the strategy first proposed by Fodor (1992) of an innate, full competence but limited performance due to the task's requirements. Working memory and related executive control abilities change throughout children's development but not their ability to understand propositional attitudes in general, which is assumed to be complete, universal, and innate or it appears very early in ontogeny. In other words, 2 year-old children do attribute full-blown beliefs—beliefs with propositional contents—as part of their understanding of others as intentional agents.

Empiricists, on the other side, deny that infants' performance on spontaneous-response FBT show that they attribute beliefs in first place. From their point of view, infants' behavior can be explained by low-level processes such as a novelty preference (Heyes, 2014) or simpler behavioral rules (Perner & Ruffman, 2005). For example, Heyes (2014) has forcefully argued that the novelty generating the surprise in the VOE and the AL procedures appears at a lower representational level than the true or false belief representations. Infants' looking behavior is a function of the degree to which the observed (perceptual novelty) and remembered or expected (imaginial novelty) low-level properties of the test stimuli—their colors, shapes, and movements—are novel with respect to the earlier events encoded by the infants in the experiment. On Perner's view, on the other hand, infants' performance in these tasks can be accounted in terms of learned or innate behavioral rules, such as that "people tend to look for an object where they last saw it and not necessarily where the object actually is" (Perner & Roesler, 2012).

More recently, though, a two-systems account of belief-ascription has been put forward (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Low, Apperly, Butterfill, & Rakoczy, 2016). According to this dual-system view, overcoming the spontaneous-response FBT depends on a fast and implicit but inflexible system. Such system reasons about "belief-like" states, called "registrations", and develops early in ontogeny. A registration is a non-propositional, extensional state that holds relations to objects and properties and has its effects directly on action. Someone who reasons about registrations would track beliefs, true and false, but in a limited range of situations. For example, infants can attribute registrations regarding the location of an object and thus succeed in many spontaneous-response FBT (Apperly & Butterfill, 2009). However, they cannot track beliefs that involve quantifiers, indefinitely complex combinations of properties, or aspects, that is, situations in which a protagonist refers to an object under one aspect in contrast to another (Apperly & Butterfill, 2009; Rakoczy, Bergfeld, Schwarz, & Fizke, 2015; Rakoczy, 2017). The fact that registrations have signature limits explains why young children fail spontaneous tasks in which the protagonist was mistaken about the identity of an object (Fizke, Butterfill, van de Loo, Reindl, & Rakoczy, 2017). Since registrations are relational states, they fail to capture the intentionality characteristic of propositional attitudes and the aspectuality of the agent's mental state (Rakoczy, 2017).

On the other hand, a late-developing and flexible system supports the attribution of genuine beliefs. This system is effortful and explicit, but also inefficient, and supports children's success in explicit FBT. Beliefs, unlike registrations, are propositional, intentional states, which engage in holistic and complex interactions with other propositional attitudes. Such a flexible system allows children to overcome the limitations posed by the early system and track more complex beliefs.

Lately, Tomasello (2018) has argued that insofar as the previous theories are based on individual cognition, the developmental puzzle will not be solved. The key to understanding beliefs requires an account of the processes of social and mental coordination with other agents and their perspectives. According to Tomasello (2018), the origins of this coordination lie in infants' joint attentional activities around their first birthdays, in which they relate two perspectives. To pass the spontaneous-response FBT, infants just need to display their ability to triangulate, the same one that they use in joint attention, which involves tracking what the agent sees or has seen in the test, and how this information will affect the agent's behavior. But such epistemic tracking does not imply that the child should understand that the agent's belief is incorrect; in other words, the infant does not need to compare the agent's perspective with his own view of the situation or with the objective situation (Tomasello, 2018).
While three-year-old children answer the false belief question focusing on the objective perspective, 4- to 5-year-old children pass the FBT because they come to coordinate the three perspectives involved: the child’s, the agent’s and the objective perspective. Much social and communicative interaction with others is required before such achievement, which is mainly characterized by communicative exchanges involving joint attention to mental content (Tomasello, 2018).

Almost fifteen years after the seminal study from Onishi and Baillargeon, there is no consensus about the capacities displayed by the infants in the non-verbal tasks. After the positive evidence obtained in the pioneer studies, new tasks either confirmed a similar pattern of results or failed to replicate the original findings (most of them were not published; see Kulke & Rakoczy, 2018). Given the mixed pattern of the accumulated evidence, it is time for a meta-analysis of the literature on implicit theory of mind. Specifically, this meta-analysis explores how correct performance in spontaneous-response FBT has revealed and whether (and how) other variables mediate this effect.

2. Method

2.1. Inclusion and exclusion criteria

We used the following criteria to select studies to include in the meta-analysis.

1. We only included studies testing participants’ capacity to attribute false beliefs in which their mean age was 26 months old or younger. Thus, we excluded all false belief conditions testing older children.
2. We only included tasks that employed one of these three paradigms for implicit false belief ascription: the violation of expectation methodology, the anticipatory looking and the interactive paradigms (grouping both tasks that require infants’ helping and pointing behaviors).
3. We only included implicit false belief conditions; thus, we excluded true belief, ignorance and control conditions that might be present in the studies as well.
4. We only included false belief conditions in which typical-developmental infants were tested. Then, we excluded false belief tests carried out on samples with atypical development.
5. We included all the tasks published from Onishi and Baillargeon seminal paper (2005) until 4th February 2019. Although we checked for unpublished data in the overview of studies done by Kulke and Rakoczy (2018), most of these data are now published and we have already included in the current meta-analysis 3.

2.2. Moderator and mediator analyses

Each condition included in the meta-analysis was coded for the following dependent variables, most of them following Wellman et al. (2001) 4:

1. Author
2. Year of publication
3. Paradigm: the paradigm employed in each condition, that is, the violation of expectation (VOE), the anticipatory looking (AL) or the interactive paradigm.
4. Age: the mean age of participants in each condition (expressed in months). In some cases, the authors reported the mean age of participants from the whole group (grouping together either true and false belief conditions or two false belief conditions).
5. Sample size: the number of participants in a condition.
6. Dropped participants: the number of participants excluded from each condition. In many cases, the authors reported the total amount of excluded participants in the whole study, grouping together false and true belief conditions, or two false belief conditions.
7. Familiarization trials: the number of familiarization trials performed in each condition. When the authors employed warm-up trials, instead of familiarization trials, we coded the familiarization trials as 0.
8. Belief: the type of belief attributed to the agent. We coded four levels, distinguishing beliefs about:
   a. an object location (L);
   b. a non-obvious property of an object (NP);
   c. the identity of an object (I); or
   d. (un)expected contents (C).

---

3 There is only one missing study from the survey published by Kulke and Rakoczy (2018) that satisfies the points 1 to 4 of our inclusion criteria and that is still unpublished (e.g., a study from Surian and Franchin). Unfortunately, we could not get more unpublished results. However, the meta-analysis is a technique that draws conclusions from the available results and estimates how many articles were not published.

4 We did not include three variables from Wellman and colleagues’ meta-analysis: the nature of the target object, the type of question and the temporal marker. The first one was difficult to code in our data. It distinguished three levels—a real object, a toy or a videotaped object—and it was not clear if the target real object was also a toy or not in many conditions. The remaining two variables corresponded to the linguistic version of the test.
9 Agent: three levels that described the nature of the agent as:
   a a real present person;
   b a videotaped person; or
   c a virtual agent (including cases of a human-like individual, a geometric shape or an animal).
10 Real presence of the target object: whether, at the test trial, the target object was:
   a real and present (i.e., the object was inside one of the boxes); or
   b absent (i.e., the object had been removed from the scene).
11 Object movements: the number of displacements of the object that the agent did not see before the object arrived at its final position.
12 Motive of the transformation: two levels capturing whether the key transformation (e.g., the change of location or the substitution of unexpected contents) was done:
   a to trick: to explicitly trick the agent, or
   b for other reason: for some other reason including no explicit reason at all.
13 Salience of the agent’s mental state: we distinguished four levels:
   a absence: the false belief state had to be tracked from the agent’s absence during the key events;
   b back turned: the false belief state had to be tracked due to the agent’s back turned during the key transformation;
   c first-person experience: the false belief experience was demonstrated initially on the children themselves; and
   d blindfold: the agent used a blindfold during the transformation.
14 Interaction: we coded whether there was interaction between the child and the agent during the test or not.
15 Design: indicates the design employed in each condition. We distinguished between:
   a a within-subjects design (WS): participants were tested in both test events (congruent and incongruent events in the VOE case) or in both conditions (false and true belief conditions); and
   b a between-subjects design (BS): participants were tested in only one test event (in the VOE case) or in only one false belief condition.
16 Test trials: the number of test trials that were conducted with each child.

Depending on the paradigm employed, we coded the dependent variable in each case. In the VOE paradigm, we coded the mean looking time (in seconds) that infants spent looking to (a) the expected event, that is, when the agent acted according to her false belief; and (b) the unexpected event, that is, when the agent acted in a way that was not according to her false belief. In the AL and the interactive paradigms, we coded the percentage of passers, that is, the number of infants that showed the correct anticipatory behavior (in the AL paradigm) or the correct helping or pointing behavior (in the interactive paradigms).

2.3. Search strategies & coding procedures

The literature on implicit false belief tasks is relatively recent (the first study was published in 2005), so new studies are known in the area and are cited as new articles appear. However, we checked the databases EbscoHost – PsycINFO, SCOPUS, and Web of Science (WoS) to ensure we were taking into account all the studies published up to date. The search terms employed were false belief, implicit false belief tasks, false belief + infants, spontaneous response false belief tasks, violation of expectation + false beliefs, looking paradigm + false beliefs, theory of mind.

We addressed studies published in English and we only included published studies. All the studies were coded by the first author, after agreement on the codes with the third author. Table 1 shows the final list of references included in the current meta-analysis.

2.4. Statistical methods

We calculated the effect sizes based on the data described in the articles and on the responses from the authors to requests for further information⁶.

On the one hand, the dependent variable in the VOE paradigm is measured on a quantitative scale (mean looking time). Then, we used the standardized mean difference (d) as the effect size. The formula of d is equal to (m1i-m2i)/spi, where m1i and m2i represent the means of the two groups, and spi the pooled standard deviation of the two groups (for which the standard deviations of the scores

⁵ AL tasks also report the differential looking score (DLS). This a continuous variable that considers the total amount of looking time at both locations (correct and incorrect areas of interest). This score results from the difference between the looking time to the correct window and the looking time to the incorrect window divided by the total looking time to both windows (Kulke et al., 2018). We coded the DLS when authors reported it, or we calculated the DLS if authors reported the mean looking times to each window. For the calculation we used the same formula as in Kulke et al. (2018): DLS = (time(correct AOI) - time(incorrect AOI)) / (time(correct AOI) + time(incorrect AOI)). We could calculate it in 11 out of the 15 conditions. However, we lacked the information to calculate the effect sizes and standard errors. This is the reason why we did not use this quantitative measure.

⁶ In one study, authors reported that 25 out of 50 infants did a correct first saccade for both false belief 1 and false belief 2 conditions (Grosse Wiesmann et al., 2018). Separate information of passers in each condition is not reported and could not be obtained via personal communication. However, as these conditions differ in the number of movements the object makes outside of the agent’s view (a variable that we coded), then we split half of the sample for FB1 and the other for FB2 in order to maintain the same overall performance and coding the two different conditions.
in the two groups and both sample sizes should be reported). The variance of \( d \) is equal to

\[
\sigma^2 = \frac{n_1 + n_2}{n_1 \cdot n_2} + \frac{d^2}{2(n_1 + n_2)}
\]

For within-subjects designs, in which infants were tested on both congruent and incongruent test trials, the standardized mean change was calculated. Its formula is equal to \((m_{1i} - m_{2i})/S_{diff}\), where \(m_{1i}\) and \(m_{2i}\) refer to the observed means at the two measurement occasions. To calculate the standard deviation of the paired differences, the corresponding standard deviations of the two measurement occasions and the correlation coefficient between both should be reported. No condition that employed a within-subject design reported the correlation between participants’ two measures. Then, we fixed a medium correlation of 0.5 for these cases.

Another formula for the standardized mean change consists of \((m_{1i} - m_{2i})/S1\) (Ausina & Meca, 2015), requiring no correlation coefficient to perform the calculus. We compared the results of this formula to the ones obtained with the previous formula with a correlation of 0.5 and the values were very close.

On the other hand, studies conducted with the AL and the interactive paradigms provide data for single groups with respect to a dichotomous dependent variable:

1. correct anticipatory look (it means, looking in anticipation to the place where the agent should go according to her false belief), or incorrect anticipatory look; or
2. correct helping or pointing behavior (i.e., opening the correct box, giving the appropriate object, or pointing to the correct location), or incorrect helping or pointing behavior.

Therefore, we used the logit transformed proportion (log odds) as an outcome measure for the studies from both paradigms. The log odds for single groups is equal to \(\log((xi/ni)/(1-(xi/ni)))\) where \(xi\) denotes the number of individuals experiencing the event of interest and \(ni\) refers to the total number of individuals in the group. The variance of log odds is equal to \(1/xi + 1/(ni-xi)\).

Finally, we converted \(d\) scores into log odds by applying the formula \(d \cdot (\pi/\sqrt{3})\); and the variance of \(d\) into the variance of log odds.

### Table 1

| References | FB conditions used in the meta-analysis |
|------------|----------------------------------------|
| Buttelmann et al. (2009) | 2 |
| Buttelmann, Over, Carpenter, and Tomasello (2014) | 1 |
| Buttelmann, Suhrke, and Buttelmann (2015) | 1 |
| Crivello and Poulin-Dubois (2018) | 2 |
| Dörrenberg et al. (2018) | 4 |
| Fizek et al. (2017) | 2 |
| Grosse Wiesmann et al. (2018) | 2 |
| Király, Oláh, Csibra, and Kovács (2018) | 2 |
| Knudsen and Liszkowski (2012a) | 2 |
| Knudsen and Liszkowski (2012b) | 2 |
| Kulke et al. (2018) | 2 |
| Luo (2011) | 1 |
| Merigo et al. (2012) | 1 |
| Morinuchi, Ban, Ono, and Uchiyama (2018) | 1 |
| Onishi and Baillargeon (2005) | 2 |
| Poulin-Dubois and Yott (2018) | 1 |
| Poulin-Dubois, Poloni, and Yott (2013) | 1 |
| Powell et al. (2018) | 2 |
| Priewasser et al. (2018) | 1 |
| Schuwerk et al. (2018) | 1 |
| Scott and Baillargeon (2009) | 1 |
| Scott et al. (2010) | 2 |
| Scott, Richman, and Baillargeon (2015) | 2 |
| Senju, Southgate, Snape, Leonard, and Csibra (2011) | 1 |
| Song, Onishi, Baillargeon, and Fisher (2008) | 2 |
| Southgate, Chevallier, and Csibra (2010) | 3 |
| Southgate et al. (2007) | 2 |
| Surian and Geraci (2012) | 2 |
| Surian, Cadi, and Sperber (2007) | 2 |
| Thoermer et al. (2012) | 1 |
| Yott and Poulin-Dubois (2012) | 1 |
| Yott and Poulin-Dubois (2016) | 1 |
| Zmyj, Prinz, and Daum (2015) | 1 |
| Total | 56 |
following the formula \( v(d) \cdot (\pi^2/\sqrt{3}) \). Then, we calculated the effect size standard error (SE) which is equal to the square root of the variance.

95% confidence intervals were calculated in the statistical program R Core Team (2013). The lower bound of the confidence interval results from applying the formula \( \log \text{odds} - (1.96 \cdot SE) \), while the upper bound of the interval is determined by \( \log \text{odds} + (1.96 \cdot SE) \).

We employed a random effects model in which we assume that the true effects are normally distributed (Borenstein, Hedges, Higgins, & Rothstein, 2009). Under this model, two sources of variability are assumed: intra-study variability (due to sampling error) and inter-study variability (each study estimates its own parametric effect) (Ausina & Meca, 2015).

Heterogeneity was assessed by calculating different statistics. First, we used the Q statistic, which is the weighted sum of squares (WSS) on a standardized scale, and its p-value as a test of significance. As a standard score, it can be compared with the expected WSS (on the assumption that all studies share a common effect) to yield a test of the null and also an estimate of the excess variance (Borenstein et al., 2009). We also quantify \( T^2 \) that reflects the variance of the true effects; and \( I^2 \) that reflects the proportion of observed dispersion that is due to true heterogeneity and is expressed as a ratio. Higgins, Thompson, Deeks, and Altman (2003) suggest that values on the order of 25%, 50%, and 75% might be considered as low, moderate, and high, respectively. Finally, when we add moderators to the model, we report \( R^2 \) which reflects the amount of heterogeneity accounted for the variables included in the model.

Possible publication bias was assessed with a funnel plot. The funnel plot is a graphical display of effect sizes against their standard errors and shows that, in the absence of publication bias, studies will be symmetrically distributed around the mean effect size, since the sampling error is random. In the presence of publication bias, the studies are expected to follow the model, with symmetry at the top, a few studies missing in the middle, and more studies missing near the bottom (Borenstein et al., 2009, p. 284).

We used the metafor package (Viechtbauer, 2010) in R Core Team (2013) to calculate effect sizes, heterogeneity estimates, to run the meta-analysis, and to perform the forest, the funnel, and the meta-regression plots. 8

3. Results

3.1. Description of studies

A PRISMA flow diagram illustrates the search strategy and study selection (see Fig. 1). The search procedure yielded 44 references for which additional information was obtained. Two of these references were duplicates (the VOE task with 18-months-old from Poulin-Dubois & Yott, 2018 with the one from Yott & Poulin-Dubois, 2016; and the AL study of Sodian et al., 2016 with the one of Thoermer, Sodian, Vuori, Perst, & Kristen, 2012), thus we removed them. Of these 42 papers, 2 studies were carried out with another paradigm (Kovács, Téglás, & Endress, 2010; Southgate & Vernetti, 2014) and, in 5 of them, the target group was constituted by older children (Burnside, Ruel, Azar, & Poulin-Dubois, 2017; Burnside, Wright, & Poulin-Dubois, 2018; Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2017; Oktay-Gür, Schulz, & Rakoczy, 2018; Wang & Leslie, 2016), thus we excluded them. Out of them, 2 additional articles were eliminated because of insufficient reported data to calculate the effect sizes (Song & Baillargeon, 2008; Träuble, Marinovi, & Pauen, 2010). Finally, the data from 33 articles were included in the meta-analysis.

From these definitive 33 papers, we additionally excluded some conditions in which authors tested older children (Fizke et al., 2017; Grosse Wiesmann, Friederici, Disla, Steinbeis, & Singer, 2018; Kulke, Reiß, Krist, & Rakoczy, 2018; Priewasser, Rafetseder, Gargitter, & Perner, 2018; Schwerk, Priewasser, Sodian, & Perner, 2018) or deaf participants (one condition from Meristo et al., 2012), and other conditions due to insufficient data to calculate the effect sizes (two VOE conditions from Powell, Hobbs, Bardis, Carey, & Saxe, 2018). The final selection of articles amounted to 56 false belief conditions in which a total of 1469 infants were tested. Table 2 reports the descriptive information of each false belief condition included in the meta-analysis like the year of publication, the sample size, infants’ mean age, the paradigm employed, the effect size and the 95% confidence intervals.

Table 3 provides a descriptive summary of the database grouped by the variables coded. The mean age of all the participants of the database was 19.55 months and, from the 56 false belief conditions, 19 employed the VOE paradigm, 15 used the AL procedure and 22 the interactive methodology.

3.2. Meta-analytic models

3.2.1. Model 1

We built a random effects meta-analytic model with maximum likelihood estimation on infants’ performance in spontaneous-response false belief tasks. This first model (model 1) included no covariate. The results showed that the estimated average effect size (calculated in log odds) is equal to \( \beta = 0.57 \), 95% CI [0.33; 0.80] (see Fig. 2). It, therefore, suggests that the average log odds corresponds to an odds of 1.76, meaning that correct performance on spontaneous-response false belief tasks was about 1.76 times more likely than incorrect performance. 9 In terms of probabilities, the probability of correct performance was 64% 10 . The null

---

8 All the data employed in this meta-analysis as well as the R script are available at https://osf.io/re8uj/?view_only=d2605771dd664831a104318db9ff7aa9

9 The calculus of the odds is the exponential function of the log odds. In this case, the exponential function of 0.5663 results in 1.76.

10 Probabilities result from applying the formula \( \text{odds}/(1 + \text{odds}) \).
hypothesis can be rejected ($z = 4.69, p < .0001$).

The total set of studies was heterogeneous as evidenced by significant $Q$-values ($Q[1,55] = 163.81, p < .001$). $T^2$, which reflects the amount of true heterogeneity (i.e., the variance of the true effects), is 0.50 (SE = 0.15). The proportion of observed dispersion in the data that is due to heterogeneity is specified by $I^2 = 71.5\%$, which indicates a medium-high proportion of variance due to heterogeneity rather than chance.

The funnel plot showed that all the effect sizes falling within the grey area would be statistically non-significant in a two-tailed test (Fig. 3). We can see an asymmetrical distribution of the effect sizes, with more studies on the right part of the plot -showing larger effect sizes- that have less precision (smaller samples). This graphical display suggests a publication bias: some studies with non-significant results were not published.

3.2.2. Model 2

Part of the heterogeneity found may be due to the influence of moderators (Viechtbauer, 2010). We examined this possibility by fitting a mixed-effects model including theoretically relevant predictors as the year of publication, infants’ age, the paradigm employed, the type of belief, the type of agent, the motive of the transformation, the design of the study, and the presence of the target object as moderators. This second model 2 includes 8 predictors. We performed a multiple meta-regression on model 2 using the restricted maximum likelihood method. Table 4 shows the results.

In this model, the estimated amount of residual heterogeneity is equal to $T^2 = 0.23$, suggesting that 53.80% ($R^2$) of the total amount of heterogeneity can be accounted for the eight moderators included in the model. The proportion of observed dispersion due to heterogeneity ($I^2$) is reduced to 52.14%. We can reject the null hypothesis based on the omnibus test ($Q_M = 49.67, df = 12, p < .0001$).

Before analyzing the implication of each moderator, a statement of caution is crucial. Given the large number of moderators considered and their confounds with one another, we need to be circumspect in the interpretation of the results about the influence of these variables on children’s performance.

Examining the predictors of model 2, we find that year of publication ($z = -4.54, p < .0001$), infants’ age ($z = 2.26, p = .02$) and the motive of transformation (to trick the agent; $z = 0.32, p = .04$) have a significant influence on performance. The test for residual heterogeneity is significant ($Q_e = 87.20, df = 43, p < .0001$), indicating that other moderators, not considered in the model, are influencing children’s performance. These results indicate that as time passes, the probability of passing the test diminishes -0.15 units (95% CI: -0.22; -0.09) in terms of the average log odds of successful performance. This means that incorrect
performance on spontaneous-response false belief tasks is 0.86 times more likely than correct performance as the year of publication is higher, suggesting that null results or children's incorrect performance on the task are currently being published.

On the other hand, when older infants are tested, the log odds of successful performance increase 0.08 units (95% CI: 0.01; 0.15), meaning that correct performance is 1.08 times more likely than incorrect performance as infants get older. When the motive of the transformation is to trick the agent, it results in an increase of 0.65 (95% CI: 0.02; 1.29) units in the log odds of correct performance. This indicates that correct performance is 1.91 times more likely than incorrect performance. Anyway, the confidence intervals of the effect sizes of these two predictors almost touch zero. Then, we should cautiously interpret them. The effects of all other predictor

| References | Condition | n  | Age (in months) | Paradigm | Effect size (log.odds) | 95% CI     |
|------------|-----------|----|----------------|----------|------------------------|------------|
| Buttelmann et al. (2009) | 1 | 25 | 18 | Interactive | 0.94 | (0.07; 1.82) |
| Buttelmann et al. (2009) | 2 | 25 | 16.3 | Interactive | 1.39 | (0.41; 2.37) |
| Buttelmann et al. (2014) | 1 | 18 | 18.3 | Interactive | 0.69 | (–0.29; 1.67) |
| Buttelmann et al. (2015) | 1 | 31 | 18.17 | Interactive | 0.06 | (–0.64; 0.77) |
| Crivello and Poulin-Dubois (2018) | 1 | 41 | 18.4 | Interactive | –0.55 | (–1.19; 0.09) |
| Dörrenberg et al. (2018) | 2 | 97 | 18.47 | Interactive | 0.31 | (–0.09; 0.71) |
| Dörrenberg et al. (2018) | 2 | 50 | 24.53 | AL | 0.49 | (–0.08; 1.06) |
| Dörrenberg et al. (2018) | 3 | 49 | 24.53 | AL | –1.13 | (–1.78; –0.47) |
| Dörrenberg et al. (2018) | 4 | 32 | 24.53 | Interactive | 0.51 | (–0.20; 1.23) |
| Fizek et al. (2017) | 1 | 23 | 26 | Interactive | –0.83 | (–1.71; 0.06) |
| Fizek et al. (2017) | 2 | 23 | 26 | Interactive | –0.44 | (–1.28; 0.40) |
| Grosse Wiesmann et al. (2018) | 1 | 25 | 25.2 | AL | 0.08 | (–0.70; 0.86) |
| Grosse Wiesmann et al. (2018) | 2 | 25 | 25.2 | AL | –0.08 | (–0.86; 0.70) |
| Király et al. (2018) | 1 | 15 | 18.1 | Interactive | –1.01 | (–2.16; 0.13) |
| Király et al. (2018) | 2 | 15 | 18 | Interactive | 0.69 | (–0.38; 1.77) |
| Knudsen and Liszkowski (2012a) | 1 | 19 | 18.4 | Interactive | 0.54 | (–0.39; 1.47) |
| Knudsen and Liszkowski (2012a) | 2 | 19 | 18.67 | Interactive | –0.11 | (–1.01; 0.80) |
| Knudsen and Liszkowski (2012b) | 1 | 16 | 18.33 | Interactive | –0.25 | (–1.24; 0.74) |
| Knudsen and Liszkowski (2012b) | 2 | 16 | 24.47 | Interactive | 0.79 | (–0.27; 1.85) |
| Kulke et al. (2018) | 1 | 11 | 24.5 | AL | –0.98 | (–2.31; 0.35) |
| Kulke et al. (2018) | 2 | 9  | 24.5 | AL | 0.22 | (–1.09; 1.54) |
| Luo (2011) | 1 | 12 | 10.3 | VOE | 1.01 | (–0.09; 2.12) |
| Meristot et al. (2012) | 1 | 10 | 23 | AL | 3.04 | (0.21; 5.88) |
| Moriguchi et al. (2018) | 1 | 37 | 18.5 | VOE | 1.19 | (0.55; 1.83) |
| Onishi and Baillargeon (2005) | 1 | 14 | 15.23 | VOE | 2.34 | (0.24; 4.44) |
| Poulin-Dubois and Yott (2018) | 1 | 53 | 18 | Interactive | –0.58 | (–1.14; –0.02) |
| Poulin-Dubois and Yott (2013) | 1 | 51 | 16.13 | VOE | –0.33 | (–1.34; 0.68) |
| Powell et al. (2018) | 1 | 94 | 18.6 | Interactive | 0.52 | (0.10; 0.94) |
| Powell et al. (2018) | 2 | 16 | 25.5 | Interactive | 0.79 | (–0.27; 1.85) |
| Priewasser et al. (2018) | 1 | 14 | 25.63 | Interactive | 2.56 | (0.53; 4.60) |
| Schuwerk et al. (2018) | 1 | 58 | 25.5 | AL | –0.07 | (–0.58; 0.45) |
| Scott and Baillargeon (2009) | 1 | 14 | 17.73 | VOE | 1.13 | (0.09; 2.16) |
| Scott et al. (2010) | 1 | 18 | 18.67 | VOE | 2.32 | (0.47; 4.17) |
| Scott et al. (2015) | 1 | 18 | 17.47 | VOE | 2.54 | (0.66; 4.42) |
| Scott et al. (2015) | 2 | 18 | 17.53 | VOE | 2.12 | (0.31; 3.93) |
| Scott et al. (2015) | 3 | 18 | 17.4 | VOE | 2.12 | (0.31; 3.93) |
| Senju et al. (2011) | 1 | 18 | 18 | AL | 1.25 | (0.14; 2.36) |
| Song et al. (2008) | 1 | 18 | 18.6 | VOE | 2.56 | (0.42; 4.69) |
| Song et al. (2008) | 2 | 16 | 18.57 | VOE | 2.65 | (0.64; 4.66) |
| Southgate et al. (2010) | 1 | 12 | 17.73 | Interactive | 1.10 | (–0.21; 2.41) |
| Southgate et al. (2010) | 2 | 12 | 17.83 | Interactive | 1.61 | (0.09; 3.13) |
| Southgate et al. (2010) | 3 | 12 | 17.97 | Interactive | 1.10 | (–0.21; 2.41) |
| Southgate et al. (2007) | 1 | 10 | 25.13 | AL | 2.20 | (0.13; 4.26) |
| Southgate et al. (2007) | 2 | 10 | 25.2 | AL | 1.39 | (–0.16; 2.94) |
| Surian and Geraci (2012) | 1 | 24 | 17.13 | AL | 0.89 | (0.01; 1.77) |
| Surian and Geraci (2012) | 2 | 16 | 11.5 | AL | –0.25 | (–1.24; 0.74) |
| Surian et al. (2007) | 1 | 28 | 13.13 | VOE | 0.93 | (–0.45; 2.30) |
| Surian et al. (2007) | 2 | 26 | 13.1 | VOE | 0.49 | (–0.93; 1.91) |
| Thoermer et al. (2012) | 1 | 44 | 18.07 | AL | 0.18 | (–0.41; 0.78) |
| Yott and Poulin-Dubois (2012) | 1 | 34 | 18.77 | VOE | 1.58 | (0.30; 2.86) |
| Yott and Poulin-Dubois (2016) | 1 | 72 | 16 | VOE | 0.45 | (0.02; 0.87) |
| Zmij et al. (2015) | 1 | 22 | 18.07 | AL | –0.76 | (–1.66; 0.14) |

Total: 1469

Note: n = sample size; CI: confidence interval.
3.2.3. Model 3
Lastly, we examined a simpler model by fitting a mixed-effects model only including the year of publication and the paradigm as moderators. The results of the multiple meta-regression for model 3, that includes two predictors, are presented in Table 5.

The estimated amount of residual heterogeneity is equal to $I^2 = 0.19$, suggesting that 61.30% ($R^2$) of the total amount of heterogeneity can be accounted for the inclusion of these two moderators in the model. The proportion of observed dispersion due to heterogeneity ($I^2$) is 48.19%. We can reject the null hypothesis based on the omnibus test ($QM = 40.12$, df = 3, $p < .0001$). A closer inspection of the predictors shows that year of publication appears to have a significant influence on performance ($z = -4.32$, $p < .0001$) as well as the VOE paradigm ($z = 2.92$, $p = .003$).

Results indicate that as the year of publication is higher, there is a reduction of -0.11 (95% CI: -0.16; -0.06) in the average log odds of successful performance. This result indicates that incorrect performance on spontaneous-response false belief tasks is 0.90 times more likely than correct performance as time passes, in the same vein as the result of model 2. On the other hand, performing the task in the VOE paradigm implies a change of 0.75 (95% CI: 0.25; 1.26) in increasing the log odds of successful performance in the task. This means that correct performance is 2.12 times more likely than incorrect performance when the test is run with the VOE

Table 3
Overall characteristics of the database.

| Variables                        | Total conditions |
|----------------------------------|------------------|
| Paradigm                         |                  |
| AL                               | 15               |
| VOE                              | 19               |
| Interactive                      | 22               |
| Familiarization                  |                  |
| 0                                | 22               |
| 1                                | 1                |
| 2                                | 17               |
| 3                                | 6                |
| 4                                | 5                |
| 5                                | 2                |
| 6                                | 3                |
| Belief                           |                  |
| (Un)Expected contents            | 1                |
| Identity                         | 7                |
| Location                         | 46               |
| Non-obvious property             | 2                |
| Agent                            |                  |
| real present person              | 38               |
| videotaped person                | 13               |
| virtual agent                    | 5                |
| Target object Presence           |                  |
| absent                           | 17               |
| real, present                    | 39               |
| Motive transformation            |                  |
| other reason                     | 38               |
| trick                            | 18               |
| Salience of agent's MS           |                  |
| absence                          | 41               |
| back turned                      | 11               |
| first-person experience          | 2                |
| first-person experience (belief revision) | 1            |
| used a blindfold                 | 1                |
| Design                           |                  |
| BS                               | 42               |
| WS                               | 14               |
| Test trials                      |                  |
| 1                                | 40               |
| 2                                | 11               |
| 3                                | 1                |
| 4                                | 3                |
| Interaction                      |                  |
| no                               | 34               |
| yes                              | 22               |

| Variables                        | Mean            | Range          |
|----------------------------------|-----------------|----------------|
| Age                              | 19.55           | [10.1 - 26]    |
| n                                | 26.23           | [9 - 97]       |
| Dropped participants             | 11.79           | [0 - 60]       |
Fig. 2. Forest plot of all the false belief conditions grouped by the paradigm employed. Each study is represented on a different row. The effect size and the 95% confidence interval for each of the conditions are reported on the right. The black square in the middle of each line represents the effect size of that condition and its size varies according to its weight in the meta-analysis. The mean effect size of all conditions and its 95% confidence interval appears at the bottom. The meta-analytic information of each paradigm appears below each subgroup of conditions.
The models were compared using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Table 6 contains the values of each criterion for the three models. Comparing model 2 with model 1, both AIC and BIC decrease: AIC is reduced from 160.26 to 123.74, and BIC decreases from 164.28 to 148.40. This shows that the fit of the second model is improving. If we then consider model 3, in comparison with model 2, AIC increases (AIC = 129.09) while BIC considerably decreases (BIC = 138.84). BIC penalizes for each additional variable included in the model, thus, the fit of model 3—with only two parameters—

Table 4
Results of the analysis of a multiple meta-regression for model 2, assuming a mixed effects model.

| Estimate | SE   | z     | p      | 95% CI        |
|----------|------|-------|--------|---------------|
| intercept| 309.30| 67.83 | 4.56   | < .001        | (176.36; 442.24) |
| Year     | −0.15 | 0.03  | −4.54  | < .001 ***    | (−0.22; −0.09)   |
| Age      | 0.08  | 0.04  | 2.26   | 0.02 *        | (0.01; 0.15)     |
| Interactive Paradigm | −1.03 | 0.64  | −1.60  | 0.11          | (−2.29; 0.23)    |
| VOE Paradigm | 0.24  | 0.54  | 0.44   | 0.66          | (−0.82; 1.29)    |
| Belief I | −1.03 | 0.83  | −1.24  | 0.21          | (−2.66; 0.59)    |
| Belief L | −0.87 | 0.75  | −1.16  | 0.25          | (−2.35; 0.61)    |
| Belief NP| −0.54 | 1.09  | −0.50  | 0.62          | (−2.69; 1.50)    |
| Videotaped person | −1.14 | 0.60  | −1.90  | 0.06          | (−2.32; 0.04)    |
| Virtual agent | −0.64 | 0.57  | −1.12  | 0.27          | (−1.77; 0.49)    |
| Motive of transformation trick | 0.65  | 0.32  | 2.03   | 0.04 *        | (0.02; 1.29)     |
| WS Design| 0.11  | 0.27  | 0.40   | 0.69          | (−0.42; 0.64)    |
| Target object real, present | −0.23 | 0.40  | −0.58  | 0.56          | (−1.02; 0.56)    |

Note: SE = standard error; CI = confidence interval; I = identity; L = location; NP = non-obvious property; WS = within-subject.

Table 5
Results of the analysis of a multiple meta-regression for model 3, assuming a mixed effects model.

| Estimate | SE   | z     | p      | 95% CI        |
|----------|------|-------|--------|---------------|
| intercept| 223.42| 51.65 | 4.33   | < .001        | (122.19; 324.65) |
| Year     | −0.11 | 0.03  | −4.32  | < .001 ***    | (−0.16; −0.06)   |
| Interactive Paradigm | 0.16  | 0.22  | 0.72   | 0.47          | (−0.27; 0.58)    |
| VOE Paradigm | 0.75  | 0.26  | 2.92   | 0.00 *        | (0.25; 1.26)     |

Note: SE = standard error; CI = confidence interval.

3.2.4. Models’ comparison
The models were compared using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Table 6 contains the values of each criterion for the three models. Comparing model 2 with model 1, both AIC and BIC decrease: AIC is reduced from 160.26 to 123.74, and BIC decreases from 164.28 to 148.40. This shows that the fit of the second model is improving. If we then consider model 3, in comparison with model 2, AIC increases (AIC = 129.09) while BIC considerably decreases (BIC = 138.84). BIC penalizes for each additional variable included in the model, thus, the fit of model 3—with only two parameters—
improves in comparison to model 2 that possesses eight predictors. As the AIC for model 2 and model 3 are not substantially different, the fit of the third model is better than the fit of the first one and therefore it is preferable over the other two models.

### 3.3. A meta-analysis per paradigm

We also run separate meta-analysis per each of the paradigms employed (see Fig. 2). Results show that the VOE paradigm had a significant mean effect size $\beta = 1.28$, 95% CI [0.86; 1.69], implying that correct performance on the task was about 3.58 times more likely than incorrect performance (see Table 7). The results from the interactive paradigms also indicate a significant mean effect size $\beta = 0.31$, 95% CI [0.01; 0.61], suggesting that correct performance on the task was about 1.36 times more likely than incorrect performance. However, this is a lower probability and the confidence interval almost reach zero. On the other hand, the meta-analysis for the AL paradigm provided no significant mean effect size ($\beta = 0.16$, 95% CI: -0.24; 0.57).

Results from the VOE and the interactive paradigm suggest that the average effect size found when all paradigms were included as moderators come from these methodologies.

### 3.4. Main effects analyses

We also inspected the main effects of all the variables we have coded. We will first consider the effect of the year of publication and we will then analyze the other methodological moderators separately. Year of publication might have a special status, likely indexing publication bias rather than methodological differences\(^\text{11}\).

#### 3.4.1. Year of publication

The year of publication, on its own, has a significant influence on performance ($z = -5.32$, $p < .0001$). Insofar as the year of publication is higher, the probability of passing the test falls -0.14 units (95% CI: -0.19; -0.09) in terms of the average log odds of successful performance (see Fig. 4). This result suggests that children’s incorrect performance on spontaneous-response false belief tasks is 0.87 times more likely than correct performance as the year of publication increases, in the same direction as we found in models 2 and 3.

#### 3.4.2. Methodological moderators

We found that the following variables produced a significant main effect: the paradigm, the type of agent, the object’s movements, the salience of the agent’s mental state and the sample size. Each of them has, on its own, a significant influence on young children’s performance. All the remaining methodological moderators produced non-significant main effects. Table 8 presents the estimates, the $p$ values and the confidence intervals for all the significant predictors.

The effect of the paradigm was already analyzed in model 3. On the other hand, when the agent was a videotaped person, there was a change of -0.62 (95% CI: -1.15; -0.08) in reducing the average log odds of successful performance.\(^\text{12}\) Regarding the object’s movements, when it was moved twice outside the view of the agent, there was a change of -1.16 (95% CI: -1.90; -0.42) in reducing the average log odds of successful performance. Then, when the agent’s false belief was induced because she was turned back during

---

\(^{11}\) Thanks to an anonymous reviewer for this suggestion.

\(^{12}\) We may say that, in the anticipatory looking paradigm, agents are mainly videotaped persons. However, this category does not completely overlap with paradigm: a videotaped person is also used in one VOE study and there are some AL tasks that employ a virtual agent (not a person).
the key events, there was a change of -0.77 (95% CI: -1.31; -0.24). The sample size also seemed to influence performance: when bigger the sample size, lower the probability of finding a correct performance in the task ($\beta = -0.01$, $p = .02$, 95% CI: -0.02; 0.00). However, the confidence interval for this effect size includes zero.

Finally, the following Figs. 5–7 show how the variables are distributed in the conditions recruited for the meta-analysis. Fig. 5a shows how all the studies that found big effect sizes were run with small samples. When bigger sample sizes were employed, the effect sizes tended to decrease. Fig. 5b, on the other hand, shows the relation between dropped participants and the effect sizes found in each condition: only a few studies excluded more than 20 participants. In addition, when all infants were included in the sample, the resulting log odds of correct performance was low (it only happened in two studies).

Effect size as a function of infants’ age is shown in Fig. 6. We can see that, as infants become older, correct performance tended to decrease although it was a non-significant trend ($p = .06$). Interestingly, most of the tests were carried out with 18-month-old infants. This age group, however, showed an uneven performance on the task: they correctly passed the VOE tests but showed all possible results in the other two paradigms considered (above, at chance and below chance performance). Interestingly, there were no studies carried out in the specific age range from 19 to 23 months old. Also, there was only one study employing the VOE paradigm with older infants (24 months old) and interactive tests were consistently conducted from 18-month-old.

Fig. 7 shows the dispersion of the data in each paradigm according to (a) the design employed, (b) the interaction between the

![Meta-regression scatterplot](image)

**Fig. 4.** Meta-regression scatterplot of the log odds of correct performance as a function of the year of publication, separating the studies according to the paradigm employed (anticipatory looking, interactive or violation-of-expectation paradigms). Each circle represents one false belief condition and its size varies depending on its weight on the meta-analysis. Each meta-regression line takes into account the effect size predicted by both moderators (year of publication and paradigm).

| Variables | Estimate | Standard Error | z    | p       | 95% CI        |
|-----------|----------|----------------|------|---------|---------------|
| Paradigm VOE | 1.10 | 0.29 | 3.83 | < .001 ** | (0.54; 1.66) |
| Agent videotaped person | -0.62 | 0.27 | -2.26 | 0.024 * | (-1.15; -0.08) |
| Object movements 2 | -1.16 | 0.38 | -3.06 | 0.002 ** | (-1.90; -0.42) |
| Salience of agents MS back turned | -0.77 | 0.27 | -2.83 | 0.005 ** | (-1.31; -0.24) |
| Sample size | -0.01 | 0.01 | -2.29 | 0.022 * | (-0.02; 0.00) |

**Table 8**

Significant Main Effects of methodological moderators.

Note: SE = standard error; CI = confidence interval; MS = mental state.

Similarly, the agent is turned back during the belief induction phase only in the AL tasks. However, this category does not overlap with all the studies using the AL paradigm since it does not occur in four tasks.

This result contradicts with the effect of infants’ age found in model 2: when age was included as a predictor, a positive estimate was obtained. However, the negative influence that age has on its own was explained in model 2 by another predictor (i.e., the year of publication).
child and the experiment, and (c) the motive of the transformation. Interactive paradigms mostly employed between-subjects designs (by testing the child only in a FB condition), while within-subjects designs were used in the other paradigms (although not as often as between-subjects ones; Fig. 7a). Fig. 7b shows how interaction overlaps with the paradigm: AL and VOE paradigms are non-interactive measures and interaction appears in those tasks in which infants should help or point to the agent. Finally, Fig. 7c shows that the transformation of the scene (i.e., change of location) was mainly done to trick the agent in the interactive tasks. All the other changes that provoked a false belief in the agent were done for other (including no) reason.
4. Discussion

This meta-analysis reveals that, on average, correct performance on spontaneous-response false belief tasks is more likely than incorrect performance, suggesting that the tasks are tapping a real phenomenon. However, the meta-analysis simultaneously uncovers important issues: (i) an increase in year of publication implies a decrease in correct performance; (ii) a considerable heterogeneity is present among studies; (iii) the asymmetrical distribution of effects sizes in the funnel plot suggests publication bias; and (iv) correct performance depends on the type of paradigm used. Taken together, these results indicate that the developmental puzzle is still unsolved.

First of all, we saw that large effect sizes were found at the onset of the research in implicit false belief attribution, but this outcome could not be replicated in more recent experiments. Relatedly, those large effect sizes were found with small samples, whereas later studies used bigger samples.

Secondly, the high percentage of heterogeneity found across studies suggests that, despite the significant effect, we still do not know which factors really explain the variance across studies. Since two of the parameters (i.e., the year of publication and the paradigm) considerably reduced such variance among studies -from a high ratio to a medium ratio of heterogeneity-, there was still much variance in the dataset that the model could not explain.

Thirdly, the skewed distribution of the effect sizes hints at the presence of publication bias in this literature. Current initiatives to look for unpublished papers and to replicate previous studies imply that many researchers become aware of this bias. For example, some journals have invited to submit replications of the tasks (for instance, the special issue Understanding theory of mind in infancy and toddlerhood published by the journal Cognitive Development) and the scientific community is doing an effort to conduct direct as well as conceptual replications (Kulke & Rakoczy, 2018). Especially relevant for the future of the field is the ManyBabies framework (Frank et al., 2017) which is currently running a large size replication on infants’ theory of mind. The project ManyBabies 2 aims to run three waves of multi-lab studies developing tasks to measure VOE, AL and interactive behaviors. In this effort, researchers from
different theoretical point of views will collaborate to devise the experiments and run them in their own labs. The results of the present meta-analysis, then, constitute a first step on that direction insofar as it provides an overall picture of the field that would be comparable with a future overview of the area as new results are accumulated.

Fourth, the type of task employed seems to be relevant. When the year of publication and the paradigm were included in the model, only the VOE paradigm –among the other two– produced a significant effect. This consequence implies that infants reliably performed above chance only within this methodology. But if false belief attribution were achieved at these ages, then it should take place regardless of the task used, just as it was found for elicited-response FBT. Studies employing within-subject design are critical to shed light on this lack of convergence across paradigms (Poulin-Dubois & Yott, 2018; Powell et al., 2018).

In addition, many of the VOE effects are due to a single lab; further, analytic decision-making varies substantially from paper to paper in this body of work. These studies combine three looking time measures in different ways. Disruptive looking and continuous-looking measures (used to end a trial), and a cumulative minimum looking time (before which a trial could not end) are systematically varied across conditions and studies without independent justification (Rubio-Fernández, 2019b). This fact certainly obscures the interpretation of the results in the VOE paradigm.

On the other hand, there is a great variability in the confirmatory looking times across the FB conditions in the VOE paradigm that is surprising in itself, because it could not be predicted a priori (all the looking times for expected and unexpected trials are reported in Table 9). Looking times considerably varied from one experimental condition to another: they stretched from 4.52 to 29.5 s when unexpected outcomes are presented and from 3.27 to 18.8 s in the case of expected events. At least part of this variability could be due to the different materials employed in each task, the setting in which studies were embedded, and the cultural context in which the studies were run. However, the existing variability should be pointed out and make researchers to raise the question of how to set and justify a criterion for expected versus unexpected looking times rather than stipulate them ad hoc as it seems it is done depending on the times found.15

Anyway, what matters for the VOE methodology is that longer looking times, in fact, indicate a surprise reaction rather than duration per se. As it was well established long ago, longer looking may also indicate a preference for novelty (Fantz, 1967): without surprise, longer looks might indicate a preference for an event that is seen as new rather than unexpected. This point brings us to a second consideration, common in the debate about the use of this methodology to assess conceptual development in first year-olds: the familiarization phase involved in the VOE paradigm might just induce perceptual expectations instead of the assumed one in terms of belief and intention attribution. The fact that a long and constant familiarization phase is required reinforces this possibility, providing the participants with some anticipation of where the elements involved in the scene are likely to move around. In other words, the familiarization phase might induce the infant to attend to a particular location (Falck, Brinck, & Lindgren, 2014) or to a particular perceptual configuration (Heyes, 2014)16.

---

### Table 9

Looking times for expected and unexpected events in the VOE paradigm.

| Article                        | Condition | Mean looking time (in seconds) | Effect size (log.odds) |
|-------------------------------|-----------|--------------------------------|------------------------|
| Dörrenberg et al. (2018)      | 1         | 18.63                          | 17.31                  | 0.32                    |
| Luo (2011)                    | 1         | 20                             | 15.5                   | 1.01                    |
| Moriguchi et al. (2018)       | 1         | 4.52                           | 3.27                   | 1.19                    |
| Onishi and Baillargeon (2005) | 1         | 26.8                           | 17.47                  | 2.34                    |
|                              | 2         | 19.89                          | 10.81                  | 3.08                    |
| Poulin-Dubois et al. (2013)   | 1         | 7.43                           | 8.59                   | -0.33                   |
| Scott and Baillargeon (2009)  | 1         | 25.5                           | 18.8                   | 1.13                    |
|                              | 2         | 28.8                           | 18.4                   | 1.82                    |
| Scott et al. (2010)           | 1         | 23.3                           | 13.6                   | 2.32                    |
|                              | 2         | 23.1                           | 16                     | 1.80                    |
| Scott et al. (2015)           | 1         | 19.6                           | 11.3                   | 2.54                    |
|                              | 2         | 18.3                           | 10.5                   | 2.12                    |
|                              | 3         | 19                             | 8.5                    | 2.12                    |
| Song et al. (2008)            | 1         | 29.5                           | 17.1                   | 2.56                    |
|                              | 2         | 29.1                           | 14.6                   | 2.65                    |
| Surian et al. (2007)          | 1         | 14.48                          | 10.56                  | 0.93                    |
|                              | 2         | 17.5                           | 13.3                   | 0.49                    |
| Yott and Poulin-Dubois (2012) | 1         | 11.44                          | 7.22                   | 1.58                    |
| Yott and Poulin-Dubois (2016) | 1         | 7.72                           | 6.19                   | 0.45                    |

---

15 Because of this variability, similar looking times are found in apparently opposite event trials. For instance, in Onishi and Baillargeon (2005) first false belief condition, a mean looking time of 17.47 seconds is taken to indicate an expected look; while in the second false belief condition, a mean looking time of 19.89 seconds was taken to indicate an unexpected look (just two seconds of difference).

16 As a corollary, accumulated looking time is a very rough measure of VOE –social looking has been proposed as an alternative, more reliable measure of VOE than looking time (Dunn & Bremner, 2017). While the VOE paradigm relies on a measure that only focuses on the macrostructure of looking, looking times are in fact composed of individual fixations of variable duration and a mix of active information processing and blank stares.
However, the studies did not include a measure of familiarization or habituation. An independent way to assess the generated expectations, if any, during the familiarization phase would be welcome. The design should also involve greater perceptual variability in terms of colors, distances, objects and all the perceptual details of the situation. It could easily be tested in (i) a control condition in which perceptual changes were introduced in the familiarization phase to check whether the VOE response also appeared, thereby indicating conceptual understanding; and (ii) a control condition in which the familiarization phase involved perceptually different instances of the same intentional event. Additionally, if infants were already able to ascribe beliefs in the familiarization phase, a very brief familiarization period would suffice.

Main effect analyses showed how multiple aspects, on their own and in combination, decreased the probability of correct performance in these tasks. We may be unaware of many other variables that moderate infants’ performance on these tasks. Although we attempted to codify more parameters, like the separation between locations and the delay between hiding the object and testing, most authors did not report such information. For instance, the distance between the two locations (i.e., boxes) was reported in only 13 out of 56 conditions. In those studies, such distance considerably varied from 5 cm to 100 cm. On the other hand, the delay between hiding the object on its last position and testing was not specified. In some cases, it could be estimated but it was a difficult task since infants controlled the end of the trials in the VOE studies. Any estimation, then, would be imprecise. The analysis of both parameters could give us a hint for the explanation of the variability in the data.

4.1. Study limitations

We remark some potential limitations to the current meta-analysis that simultaneously constitute venues for new research. First, we only focused on peer-reviewed journal publications; hence we did not have access to reports that failed to be published. The present meta-analysis, as we claimed before, is a necessary first step to have a panoramic view of the evidence published so far. It is also an open invitation to submit non-published results in order to increase the number of false belief conditions considered and rerun the analyses. Second, other moderators that we did not code could have influenced the phenomenon (like nationality, the distance between locations, the delay between hiding and testing, etc.). Third, although we chose to work with the more prolific paradigms in the field, new ways to measure implicit false belief ascription are being developed, like the use of electroencephalography (EEG) to record brain wave patterns during the task (Southgate & Vernetti, 2014).

Four, we selected sample sizes in which infants were younger than 26 months old. In future work, it would also be compelling to study the developmental trend of the capacity to attribute beliefs and see what happens in the transition from two to three years old. But this decision would also imply broadening the inclusion criteria in order to select verbal spontaneous-response paradigms, like the verbal AL and the verbal VOE tasks (He, Bolz, & Baillargeon, 2012; Scott, He, Baillargeon, & Cummins, 2012). Finally, we only included conditions in which infants were required to attribute false beliefs to another. However, almost all studies include additional conditions, like true belief as well as ignorance and other control conditions. Then, infants’ above chance performance on implicit FBT would be doubtful if they, for example, performed at chance or below chance level in the true belief condition. As Rubio-Fernández (2019a) correctly pointed out, different kind of criteria have been proposed for passing the FBT in infants and older children. While a differential performance between true and false conditions is expected for infants, older children also have to perform above chance level in both conditions in the traditional task. So, a careful analysis of infants’ performance in spontaneous-response true belief tasks is also needed to get a more comprehensive view of infants’ mentalistic abilities.

5. Conclusion

In conclusion, this meta-analysis is the first one in proposing an integrative view of all the published studies on spontaneous-response false belief tasks in children younger than 26 months old. Infancy is a key developmental period in which implicit methodologies are being used to test the origins of many psychological constructs, such as theory of mind and the specific capacity to attribute false beliefs. The emerging picture from this systematic review, however, is puzzling. Although correct performance on spontaneous-response FBT was on average more likely than incorrect performance, the analyses revealed publication bias and high heterogeneity across the data. While several authors have shown that preschoolers display a robust developmental pattern from a below-chance to an above-chance performance in the standard FBT (Wellman et al., 2001), infants’ performance on the nonverbal FBT is equivocal and does not show any clear developmental pattern.

Understanding the behavior of another agent according to her false belief would minimally require robust attributions. We mean that the attributer, who ascribes a false belief to an agent, should correctly help or inform her when necessary, look in anticipation to the location where she will go and look longer when she acts incongruently with her epistemic state. Experimental settings should move forward in that direction and test the same children across a battery of implicit tasks that employ different paradigms. In this way, we could examine individual differences in performance (e.g., infants that show understanding across all the implicit measures and infants that fail all of them) and convergent validity of the tests (like Poulin-Dubois & Yott, 2018 who failed to find convergent

(footnote continued)

(Astle, 2007, p. 50). In the VOE paradigm researchers manually collect information about infants’ looking time to the scene but which aspects of it infants are exactly looking at remain unknown or uncontrolled. The use of more sophisticated technology, like eye-tracker devices, is needed to examine what infants are doing during that period of time. An example of that endeavor is Dörrenberg and colleagues’ (2018) paper in which they also measure pupil size with an eye-tracker.
validity). Similarly, tasks that concurrently combine two implicit measures are welcomed, like measuring infants’ anticipatory look and (un)expected looks (e.g., Dörrenberg, Rakoczy, & Liszkowski, 2018) or participant’s anticipatory look and interactive behavior (work in progress).

Declaration of Competing Interest

None.

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness (research project FFI2017-86351-R). Pamela Barone was supported by a Ph.D. scholarship from the Spanish Ministry of Economy and Competitiveness (Grant Number BES-2014-067640). We would like to thank Laura Danón, Henry Wellman, Hannes Rakoczy, the members of EvoCog group, assistants to the Spring School 2017 on Social Cognition, Emotion and Joint Action, and four anonymous reviewers for their very valuable comments on previous drafts of this manuscript.

References

Apperley, I. A. (2011). Theory of mind. Hove, UK: Psychology Press.
Apperley, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? Psychological Review, 116(4), 953. https://doi.org/10.1037/a0016923.

Aslin, R. N. (2007). What's in a look? Developmental Science, 10(1), 48–53. https://doi.org/10.1111/j.1467-6884.2007.00563.x.

Avina, J. B., & Meca, J. S. (2015). Meta-análisis en ciencias sociales y de la salud. Madrid, Spain: Editorial Síntesis.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? Cognition, 21(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8.

Bennett, J. (1978). Some remarks about concepts. Behavioral and Brain Sciences, 1(4), 557–560. https://doi.org/10.1017/S0140525X00006573.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. Chichester, UK: John Wiley & Sons.

Burnside, K., Ruel, A., Kear, N., & Poulin-Dubois, D. (2017). 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.

Burnside, K., Wright, K., & Poulin-Dubois, D. (2018). Social orienting predicts implicit false belief understanding in preschoolers. Journal of Experimental Child Psychology, 175, 67–79. https://doi.org/10.1016/j.jecp.2018.05.015.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. Cognition, 112(2), 337–342. https://doi.org/10.1016/j.cognition.2009.05.006.

Buttelmann, D., Over, H., Carpenter, M., & Tomasello, M. (2014). Eighteen-month-olds understand false beliefs in an unexpected-contents task. Journal of Experimental Child Psychology, 119, 120–126. https://doi.org/10.1016/j.jecp.2013.10.002.

Buttellmann, F., Suhrie, J., & Buttellmann, D. (2015). What you get is what you believe: Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task. Journal of Experimental Child Psychology, 131, 94–103. https://doi.org/10.1016/j.jecp.2014.11.009.

Butterfill, S. A., & Apperley, I. A. (2013). How to construct a minimal theory of mind. Mind and Language, 28(5), 606–637. https://doi.org/10.1111/mila.12036.

Carruthers, P. (2013). Mindreading in infancy. Mind & Language, 28(2), 141–172. https://doi.org/10.1111/mila.12014.

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.

Crivello, C., & Poulin-Dubois, D. (2018). Infants’ false belief understanding: A non-replication of the helping task. Cognitive Development, 46, 51–57. https://doi.org/10.1016/j.cogdev.2017.10.003.

Dennett, D. (1978). Beliefs about beliefs. Behavioral and Brain Sciences, 1(4), 568–570. https://doi.org/10.1017/S0140525X00006664.

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.

Dunn, K., & Bremner, J. G. (2017). Investigating looking and social looking measures as an index of infant violation of expectation. Developmental Science, 20(6), e12452. https://doi.org/10.1111/desc.12452.

Falk, A., Bröckl, I., & Lindgren, M. (2014). Interest contagion in violation-of-expectation-based false-belief tasks. Frontiers in Psychology, 5(January), 1–5. https://doi.org/10.3389/fpsyg.2014.00023.

Fanz, R. L. (1967). Visual perception and experience in early infancy: A look at the hidden side of behavior development. In H. W. Stevenson, E. H. Hess, & H. L. Rheingold (Eds.). Early behavior: Comparative and developmental approaches (pp. 181–224). New York: Wiley.

Fiske, E., Butterfill, S., van de Loo, L., Reindl, E., & Rakoczy, H. (2017). Are there signature limits in early theory of mind? Journal of Experimental Child Psychology, 162, 209–224. https://doi.org/10.1016/j.jecp.2017.05.005.

Fodor, J. A. (1992). A theory of the child’s theory of mind. Cognition, 43(3), 283–296. https://doi.org/10.1016/0010-0277(92)90004-2.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Flocchia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Infancy, 22(4), 421–435. https://doi.org/10.1111/inf.12182.

Gopnik, A., & Astington, J. W. (2018). Children’s understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. Child Development, 59(1), 28–37. https://doi.org/10.1111/cdev.11038.

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.2017.07.007.

Grose Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. Developmental Science, 20(5), e12454. https://doi.org/10.1111/desc.12454.

Harman, G. (1978). Studying the chimpanzee’s theory of mind. Behavioral and Brain Sciences, 1(4), 576–577. https://doi.org/10.1017/S0140525X000076743.

He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-year-olds succeed at a verbal anticipatory-looking false-belief task. British Journal of Developmental Psychology, 30(1), 14–29. https://doi.org/10.1342/0140525X.2011.02070.x.

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

Higgins, G. (2003). Measuring inconsistency in meta-analyses. BMJ, 327(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557.

Hogrebe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of Epistemic States. Cognitive Development, 1(4), 567–582. https://doi.org/10.1016/0885-2014(86)90037-6.

Jacob, P. (2013). A puzzle about belief-ascertainment. In B. Kaldis (Ed.). Mind and society: Cognitive science meets the philosophy of the social sciences. Synthese Library, Berlin: Springer.

Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in 3-year-old children. Proceedings of the National Academy of Sciences, 2018, 03505. https://doi.org/10.1073/pnas.1803505115.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586. [https://doi.org/10.1111/j.1467-9280.2007.01943.x](https://doi.org/10.1111/j.1467-9280.2007.01943.x).

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30*(1), 30–44. [https://doi.org/10.1111/j.2044-835X.2011.02046.x](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](https://doi.org/10.1111/j.2044-835X.2011.02067.x).

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences, 115*(34), 8491–8498. [https://doi.org/10.1073/pnas.1804761115](https://doi.org/10.1073/pnas.1804761115).

Trauble, B., Marinovi, V., & Pauen, S. (2010). Early Theory of Mind Competencies: Do Infants Understand Others? Beliefs? *Infancy, 15*(4), 434–444. [https://doi.org/10.1111/j.1532-7078.2009.00025.x](https://doi.org/10.1111/j.1532-7078.2009.00025.x).

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. [https://doi.org/10.18637/jss.v036.i03](https://doi.org/10.18637/jss.v036.i03).

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](https://doi.org/10.1111/mila.12099).

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta analysis of theory of mind development: The truth about false belief. *Child Development, 72*(3), 655–684. [https://doi.org/10.1111/0010-0277.00004-5](https://doi.org/10.1111/0010-0277.00004-5).

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

Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology, 30*(1), 156–171. [https://doi.org/10.1111/j.2044-835X.2011.02060.x](https://doi.org/10.1111/j.2044-835X.2011.02060.x).

Yott, J., & Poulin-Dubois, D. (2016). Are infants’ theory of mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development, 17*(5), 683–698. [https://doi.org/10.1080/15248372.2015.1086771](https://doi.org/10.1080/15248372.2015.1086771).

Zmyj, N., Prinz, W., & Daum, M. M. (2015). Eighteen-month-olds’ memory interference and distraction in a modified A-not-B task is not associated with their anticipatory looking in a false-belief task. *Frontiers in Psychology, 6*, 857. [https://doi.org/10.3389/fpsyg.2015.00857](https://doi.org/10.3389/fpsyg.2015.00857).