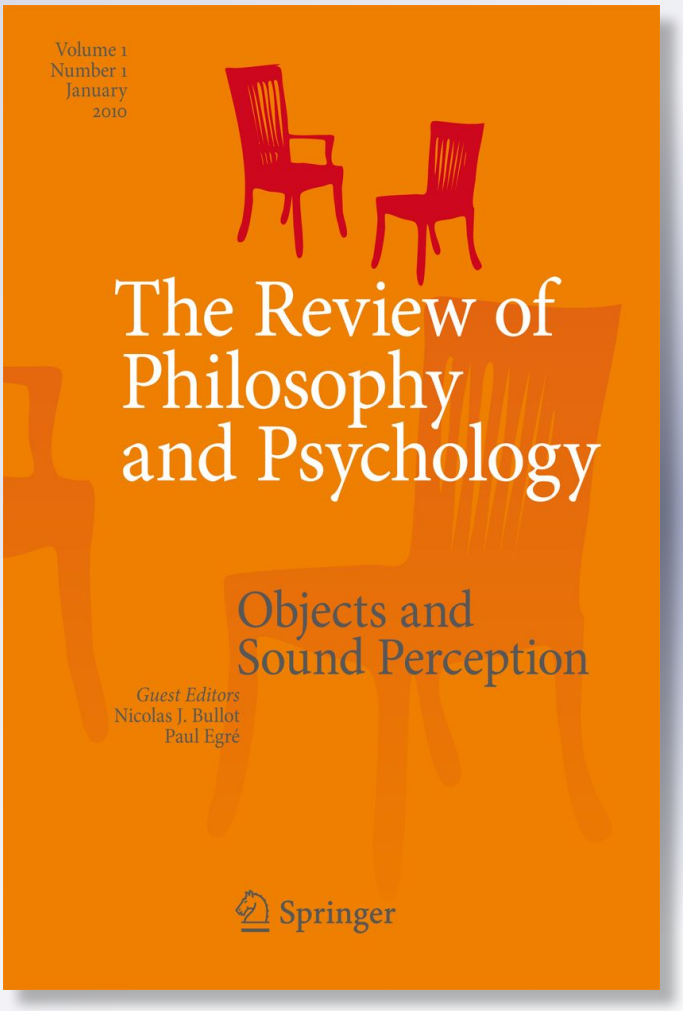
On the Long Road to Mentalism in Children's Spontaneous False-Belief Understanding: Are We There Yet?

Jason Low & Bo Wang

Review of Philosophy and Psychology

ISSN 1878-5158

Rev.Phil.Psych.
DOI 10.1007/s13164-011-0067-y



Springer

Rev.Phil.Psych.
DOI 10.1007/s13164-011-0067-y

On the Long Road to Mentalism in Children's Spontaneous False-Belief Understanding: Are We There Yet?

Jason Low · Bo Wang

Abstract We review recent anticipatory looking and violation-of-expectancy studies suggesting that infants and young preschoolers have spontaneous (implicit) understanding of mind despite their known problems until later in life on elicited (explicit) tests of false-belief reasoning. Straightforwardly differentiating spontaneous and elicited expressions of complex mental state understanding in relation to an implicit-explicit knowledge framework may be challenging; early action predictions may be based on behavior rules that are complementary to the mentalistic attributions under consideration. We discuss that the way forward for diagnosing early mentalism is to analyze whether young candidate mind-readers' visual orienting cohere across different belief-formation by belief-use combinations. Adopting this formal cognitive analysis, we conclude that whilst some studies come tantalizingly close to sign-posting mentalism in infants and young children's spontaneous responses, the bulk of evidence for early mentalism grades into behaviorism.

Human beings can interpret others' actions, motivations and thoughts in terms of underlying mental states—this everyday reasoning is termed as 'theory of mind' (ToM). Thus, when watching the film 'The Sixth Sense', we understand the plot twist wherein the child psychologist character *wants* to treat the boy for hallucinating seeing dead people but—*not realizing* that ghosts do not always know that they are dead—*falsely believes* that he himself is alive. Decades of research, using elicited response tasks that ask direct verbal questions, concluded that children only start to understand false-belief from around their 4th birthday onwards (Wellman et al. 2001). Intriguingly, recent research findings suggest that false-belief understanding may be present much earlier in life—when children are tested on spontaneous response tasks such as anticipatory looking (AL—Southgate et al. 2007) or violation-of-expectancy (VOE—Onishi and Baillargeon 2005).

J. Low (✉) · B. Wang
School of Psychology, Victoria University of Wellington, PO Box 600, Wellington, New Zealand 6140
e-mail: Jason.Low@vuw.ac.nz

Published online: 06 August 2011

In this review, we discuss whether extant evidence on children's spontaneous looking responses in false-belief tasks warrants ascription of early mentalism. We explain that extant research has not yet managed to arbitrate between behavioral and mental state accounts that explain the same input–output function, and thus suggest that it is not yet clear whether mentalism can even be signposted in the spontaneous looking responses displayed by infants and young preschoolers. We then go on to discuss possible approaches that may do the job in determining when and how early in life children may be mentalists.

1 Dissociation Between Spontaneous and Elicited False-Belief Responding

Despite the quantitative strengths of Wellman et al.'s (2001) seminal meta-analysis of research on children's ToM development, only children's correct versus incorrect judgments to direct false-belief questions were treated as the essential dependent variable. Data on children's non-judgmental eye gazing to two candidate locations in the false-belief task were entirely excluded (Moses 2001). Indeed, studies emphasizing young preschoolers possessing a kernel of non-judgmental knowledge can have substantial theoretical stake in illuminating how false-belief understanding develops (e.g., Clements and Perner 1994; Ruffman et al. 2001). Clements and Perner were the first to introduce a spontaneous response task that incorporated an AL modality to test preschoolers' false-belief sensitivity. The task was based on the classical unexpected transfer false-belief scenario. However, before asking children a direct false-belief question, participants heard the story narrator wonder aloud: "I wonder where [the agent] will look for the object". Clements and Perner used the "I wonder" prompt to trigger children's spontaneous visual anticipation of where an agent with a false-belief about the location of an object will search for the object. The results showed that 3-year-olds, but not younger, spontaneously looked towards the correct location where the agent thought the object was, even though they made wrong explicit verbal judgments of the agent's future searching behavior. Clements and Perner suggested that accuracy in 3-year-olds' AL might reflect an implicit understanding of false-belief.

Subsequent research replicated and extended Clements and Perner's (1994) findings. Ruffman et al. (2001) confirmed that 3-year-olds who gave incorrect verbal false-belief answers did not even want to bet a modest number of counters on the location as indicated by their correct eye gaze. Thus, children do not appear to be conscious of the knowledge conveyed in their AL responses. Low (2010) attempted to give functional meaning to the implicit-explicit knowledge framework by speaking to the different patterns of correlations involving spontaneous and elicited false-belief responding. Studying individual differences amongst 3- to 4-year-olds, Low found that AL was not correlated with language ability or executive functioning performance but complex grammatical understanding and conflict control ability were associated with direct false-belief answering. Furthermore, whilst there was a robust dissociation between AL and verbal answering, individual differences in children's looking time towards the correct belief-based location uniquely correlated with individual differences in explicit verbal false-belief answering. Hence, there is some degree of plausibility in the representational redescription of false-belief

understanding such that implicit awareness is transformed for participation in direct judgments and, further, the development of fully explicit understanding is partly advanced by executive functioning and language abilities.

Whilst accurate AL amongst 3-year-olds were securely replicated across Clements and Perner (1994), Low (2010) and Ruffman et al. (2001), these studies also shared the same limitation of visual orienting being triggered by a verbalized "I wonder" prompt. It is possible that early false-belief understanding does not undergo implicit-explicit representational format changes per se. For example, young children may be unable to express their latent understanding due to task specific factors (cf. Csibra and Southgate 2006; Southgate et al. 2007). It is possible that very young preschoolers might interpret "where" in the verbal prompt as referring to the location of the hidden object, rather than the location of the actor's subsequent actions (however, see Clements 1995, for counter arguments).

2 Spontaneous Non-Verbal False-Belief Performance

Southgate et al. (2007) developed an elegant non-verbal AL paradigm to test whether 25-month-olds were able to show signs of false-belief understanding. In their task, children were first familiarized with two events in which a puppet hid a toy in one of two boxes whilst a female agent looked on. Then the agent opened one of two windows (which was above the boxes) to retrieve the toy when both windows were illuminated. The illumination of the windows was paired with a simultaneous bell sound. In doing so, Southgate et al. ingeniously created the light and sound pairing as a non-verbal prompt to elicit AL toward one of the windows later in the test trial. In the test trial, the agent saw the puppet hide the toy in the left or the right box. A phone then rang behind the agent, leading her to turn toward the sound. Whilst the agent was facing away, the puppet retrieved the toy and left with it. Another important modification compared to previous spontaneous-response tasks was that the target toy was always removed from the scene in the test trial, avoiding the possibility that knowledge of the object's location may mask children's sensitivity to false-belief. By making the task non-verbal and removing the target object from the scene, Southgate et al. uncovered that 2-year-olds correctly looked in anticipation of where an agent with a false-belief would search for an object—indicating that toddlers might be able to attribute false-beliefs to others.

Even earlier sensitivity to false-belief has been claimed by another group of studies using the VOE paradigm in which looking fixations to unexpected as compared to expected outcomes are contrasted. In Onishi and Baillargeon's (2005) ground-breaking study, they found that 15-month-olds looked much longer when an agent's behavior was inconsistent, as opposed to consistent, with her false-belief, indicating that infants might implicitly attribute to the agent a false-belief about the location of an object. Onishi and Baillargeon's findings have been replicated (Träuble et al. 2010) and even extended to 13-month-olds (Surian et al. 2007). Recently, using the VOE paradigm, Kovács et al. (2010) reported that infants as young as 7-months could compute an agent's belief and maintain it even in the absence of the agent.

The above results, however, are intensely debated. On one hand, the findings appear to be commensurate with the possibility that false-belief understanding may be present very early in life, and may have an innate basis (Leslie 2005). On the other hand, Perner and Ruffman (2005) offered a number of alternative explanations. For instance, with respect to Onishi and Baillargeon's data, they suggested that looking times may be due to infants grappling with new agent-object-location association links in the test event that appear different from the ones formed during familiarization. It is also not clear whether infants attributed false-belief or ignorance to the agent (Southgate et al. 2007).

Senju et al. (2011) ruled out the possibility that infants' success is based only on associations in the stimuli. In their AL study, 18-month-olds first experienced either a visually identical opaque or trick blindfold (looked opaque but could be seen through), and subsequently watched a video sequence (similar to Southgate et al. 2007) in which an actor wore the respective blindfold while an object was displaced in front of her. Infants who experienced the opaque blindfold showed AL that the actor's search action would accord with her having a false-belief about the object's location, but infants who experienced the trick blindfold did not. It is important to emphasize that infants could not observe either themselves or others wearing the blindfold and, further, the same video sequences was presented in both conditions. The results, then, are unlikely due to infants acquiring and grappling with possible associations between blindfold-wearing and observable behaviors. It is an important discovery that 18-month-olds are at least able to represent what an agent did and did not see at a given point in time, and are able to use such information for anticipating an agent's future action.

Ignorance-based interpretations are readily contained by Southgate et al.'s (2007) AL study. Since the target item was removed completely from the scene in Southgate et al.'s design, an 'ignorance leads to error' interpretation would have no basis in predicting preschoolers' correct visual orienting: with the object removed, both locations are incorrect. And yet, 2-year-olds looked first at the empty belief-based location when anticipating where the agent would search for the toy. Recent VOE data also qualifies against ignorance interpretations of looking fixations. Scott and Baillargeon (2009) tested 18-month-olds in an unexpected identity belief-inducing event and reported that infants' looking patterns were different depending upon whether the agent had a mistaken belief or was ignorant. Their results indicated that infants in the 2nd year of life distinguish that mistaken agents should act in accord with their false-belief whilst ignorant agents should act randomly.

3 Efforts in Meeting Perner and Ruffman's (2005) Challenge: Ruling out Behavior-Rules

Whilst new generation AL and VOE evidence speak against association and ignorance explanations, behavior rule interpretations may be the final frontier to confidently ascribing mentalism as underwriting infants and young preschoolers' spontaneous false-belief responding. For example, extent AL research has only managed to test young preschoolers' spontaneous false-belief understanding via a single-shot unexpected transfer belief-inducing scenario (e.g., Clements and Perner

1994; Low 2010; Ruffman et al. 2001; Senju et al. 2011; Southgate et al. 2007). According to Perner and Ruffman (2005), young children's AL can be inevitably open to a rule-based explanation (e.g., children attribute that the agent will look for the object where he or she last saw it). Restricted documentations of accurate responses being coherent across diverse belief-inducing scenarios have implications for whether we can straightforwardly classify responses stripped to the level of eye gaze as reflecting implicit mental-state understanding (as opposed to implicitly implemented in procedural rules). For Perner and Ruffman, "the [mentalistic] conclusions from the standard false-belief task are warranted only because understanding of false-belief around 4 years can be demonstrated in a variety of belief-inducing situations" (p. 216).

To meet Perner and Ruffman's (2005) challenge, different VOE studies have tested and confirmed infants' reasoning in different belief-inducing scenarios, such as false-beliefs about location, contents, identity, and properties. Infants show accurate looking across these contexts (e.g., Scott and Baillargeon 2009; Scott et al. 2010; Song and Baillargeon 2008; Song et al. 2008; Surian et al. 2007). VOE research has been cutting-edge in documenting the breadth of infants' looking fixations as being potentially fitting with an early implicit mind-reading hypothesis. That said, the combined evidence of success amongst infants in diverse belief-inducing contexts is largely based on data between age groups and between children. Researchers are only beginning to uncover whether children of a given age who show accurate looking responses in one false-belief inducing scenario will necessarily show accurate looking responses in other false-belief inducing scenarios (He et al. 2011). Comprehensive data bearing on coherence in spontaneous responses across diverse false-belief inducing tasks in the very same age cohort (weaker evidence) and the very same infants/children (stronger evidence) are still lacking.

Furthermore, behavior rule analyses of individual VOE studies are possible. Consider, for example, Song and Baillargeon's (2008) study. In their study, 14.5-month-olds watched events where an agent faced a stuffed skunk and a doll with blue pigtails; the agent consistently reached for the doll indicating that she preferred it over the skunk. In the false-belief trial, whilst the agent was absent, the experimenter hid the doll in a plain box and hid the skunk in a hair box (the hair box had a tuft of blue hair protruding from under its lid). The results indicated that infants who saw the agent grasp the plain box looked reliably longer than those who saw the hair box outcome event. For Song and Baillargeon, the findings suggested that infants mentalistically reasoned that the agent mistakenly perceived the tuft of hair as part of the doll and hence falsely believed that the doll was hidden inside the hair box and so were surprised when she reached for the hair box**.** However, it is equally possible for infants to base their looking responses upon the general behavior rule that a person will reach for the box where attributes of the object to be retrieved are visible. Song and Baillargeon also found that when the agent had a preference for the stuffed skunk over the doll, infants looked reliably longer when the agent grasped the hair box than those who saw the plain box test event. Even in this situation, a complementary general behavior rule could also explain the pattern of looking fixations: a person will avoid reaching to a location where attributes of the object not to be retrieved are partly visible.

This challenge in arbitrating between mind-reading and behavior-reading explanations does not mean that VOE research has failed to control for *any* behavior rule interpretations. Scott et al. (2010), for example, reasoned that if 18-month-olds possessed a robust implicit understanding of false-belief and are not constrained by behavior search rules for predicting agents' actions, they should be able to show appropriate visual expectations on a *non*-search task. Their reasoning parallels research indicating that 4-year-olds who verbally pass search based false-belief tasks (e.g., unexpected transfer) also verbally pass non-search false-belief tasks (e.g., appearance-reality) (e.g., Gopnik and Astington 1988). In Scott et al.'s study, as an agent watched, an experimenter demonstrated that her object had a non-obvious property (it made a rattling sound when shaken). Next, the experimenter asked the agent to perform the action by choosing between two test objects: one was identical to the experimenter's object and the other differed in color and pattern. However, infants (but not the agent) were shown in a prior trial that the different test object rattled when shaken, but the identical test object did not. Thus, the agent will have a false-belief about similar objects having similar non-obvious properties. Infants looked longer when the agent (upon being asked by the experimenter, "Can you do it?") reached for the non-identical test object compared to the identical test object. Infants appear to be able to attribute others' false-beliefs about non-obvious properties. Scott et al. anticipated that such findings could still be open to infants following a particular behavior rule that agents will select an object that looks most like a prior object when reproducing an interesting sound. Thus, they conducted a clever follow-up experiment whereby the experimenter first demonstrated to the agent the properties of her object and the two test objects by shaking each object in turn. The agent had knowledge that the different test object rattled whilst the identical test object did not make any sound. Then, whilst the agent was absent, the experimenter removed the lids from the two test objects and poured the marbles from the different test object into the identical test object. In this reversed false-belief scenario, infants flexibly accommodated to the reversal of the agent's awareness—they looked reliably longer when the agent reached for the identical test object compared to the different test object.

Träuble et al. (2010) also introduced variations in agents' information access in order to disentangle between rule-based and mind-reading accounts of infants' looking fixations. In their study, a box was placed on either arms of a balance beam. The agent could operate the balance beam by lifting one side manually leading the ball to roll from one box to another without making any noise. Infants assigned to the false-belief condition watched as the agent observed the ball roll from one location to the other. Then, whilst the agent's back was turned, infants saw the ball change its location. For infants allocated to the manual-control condition, the only difference was that whilst the agent's back was turned, the agent herself used her hand to lift one side of the balance beam causing the ball to change its location. Infants in the false-belief condition looked longer when the agent reached inside the ball box compared to the empty box (replicating Onishi and Baillargeon 2005). Infants in the manual-control condition, however, looked longer when the agent reached inside the empty box compared to the ball box. Träuble et al. suggest that differences in reactions between the two conditions indicate that infants' looking responses are not pinned to the behavior rule that agents will look where they last saw. If this rule was dominating, then infants should look longer when the agent

reached for the ball box in the manual control condition, regardless of whether she might know the object's whereabouts during the test event even if she did not see the location reversal. Whilst Scott et al. (2010) and Träuble et al.'s (2010) attempts to contain non-mentalistic interpretations of children's looking responses are forward thinking, we will go on to discuss how each of these studies has not yet managed to control for behavior-rule explanations that are complementary to the mind-reading explanations under consideration.

4 Plausibility Arguments are Relevant but Not Decisive

To date, no experiment has yet been designed whose positive results can only be explained and predicted by a mental-state account but not by a complementary behavior-rule account. To differentiate between the two accounts, researchers have raised plausibility arguments, especially about parsimony (Perner 2010, 2011). According to the principle of parsimony, the account that makes the fewest assumptions will be preferred. However, as Perner (2011) has pointed out, depending on how one "counts" assumptions, each account may claim to have parsimony on their side.

For researchers who favor a rich interpretation, a mental-state account is argued to involve just "one rule" that governs infants' expectations about agents' behavior: people act based on the information available to them (He et al. 2011; Scott et al. 2010). Whilst it is possible to spin out in specific situations which specific false-belief infants might attribute (e.g., about the location, identity, properties, or contents of an object), the bottom line is that, if children can determine that the agent holds false information about an aspect of a scene, they will expect the agent to act on it. For example, He et al. sought to uncover whether VOE fixations amongst children in same age group (2.5-year-olds) were consistent between the unexpected transfer and unexpected contents belief-inducing scenarios. The results of their study showed that—across the two belief-formation scenarios—toddlers consistently looked longer at unexpected outcomes wherein agents' search actions conflicted with their false-beliefs about desired items being in a particular location or were incongruent with their false-beliefs about which packages contained items they wanted. A recent study by Scott and Baillargeon (2009) also suggested that infants may be capable of attributing two false-beliefs simultaneously: the agent's false-belief about the identity of the penguin toy under the transparent cover being a single piece leads her to hold the additional false-belief that the 2-piece penguin toy must be under the opaque cover. If we considered that the mental-state account only involved one rule that people act on information available to them, then no new rule would be needed for reasoning in new false-belief inducing scenarios: the rule is always the same and makes the same predictions. In contrast, when more and more belief-inducing situations are examined, a behavior-rule account must come up with more and more local heuristics to explain positive results. On this analysis, there appears to be theoretical parsimony in favoring a ToM explanation over a behavior-rule explanation of infants and young preschoolers' spontaneous false-belief responding.

The analysis changes dramatically, however, depending on how else researchers define and "count" behavior and ToM rules. Perner (2010) has argued that when one

explicates the computational abstractions infants and young preschoolers must make in order to anticipate how agents' minds interface with their environment and their actions, a mentalism account also requires a number of rules (see also Whiten 1996). Importantly, whilst Perner views the input–output functions to be the same in both accounts, their internal structures differ. A behavior rule captures a direct and specific link between a situation and an action (S→A). If we combine different situations with different actions, then the number of the behavior-rules needed for correct predictions in all cases will be the number of different situations multiplied by the number of different actions. On the other hand, in a mentalism account, a mental state (e.g., false-belief) is considered as an intervening variable (S→M→A) (Whiten). If we combine different situations with different actions, then the number of mental-state rules for correct predictions in all cases will be the number of different situations plus the number of different actions. Put formally, increases in rule complexity for determining coherent false-belief attribution are determined by computational procedures needed for combining belief-formation (*f*) by belief-use (*u*) situations wherein the number of rules required for claiming ToM-rules over behavior-rules for action prediction are ($f+u$) and ($f\times u$), respectively. Applied to He et al.'s (2011) study, toddlers' attributions were diagnosed across 2 belief-formation scenarios (unexpected transfer and unexpected contents) by 1 belief-use scenario (predicting where agents would *look* for a target object). A ToM-rules account would require a total of 3 (2+1) mentalistic abstractions to predict coherent looking responses. The first mentalistic abstraction is that if the agent did not see the object transferred from location 1 to location 2, then the agent will have a false-belief that the object is in location 1. The second abstraction is that if the agent did not see the contents of containers at locations 1 and 2 being swapped, then the agent will have a false-belief that the container at location 1 will contain the desired item. In order to correctly predict or anticipate the agent's desires to look for the target object across both belief-inducing situations, the final abstraction in the web is that if the agent has a false-belief, then the agent will look for the target item at location 1. By comparison, a behavior-rules account would only require a total of 2 (2×1) abstractions to accommodate coherent looking responses. The first behavior-rule is that agents will look for an item in a location where they last saw it. The second behavior-rule is that agents will look for an item in containers that usually contain it. Of course, a hidden assumption here is that the unit cost for a given rule is the same, regardless of whether that rule is behavioral or mentalistic. Overall, based on this alternative treatment, a behavior-rules account turns out to be more economic of toddlers' representational resources compared to a ToM-rules account.

Other VOE studies that have attempted to contain behavior rule interpretations are amenable such an alternative cognitive analysis. With respect to Träuble et al.'s (2010) study, for example, the complexity of a ToM-rules account for accommodating where agents will look for unexpectedly and manually transferred objects could take the following form. First, in the false-belief condition, if the agent did not see the ball roll from locations 1 to 2, she would have a false-belief of its whereabouts in location 1. Second, in the manual-change condition, if the agent did not see but intentionally caused the ball to roll from locations 1 to 2, she would have a true-belief of its whereabouts in location 2. Third, if the agent has a false- or true-belief, she will look in locations 1 or 2, respectively. Nonetheless, compared to the total of 3 ToM-rules

required to deploy correct looking across Träuble et al.'s tasks, a behavior-rules analysis is also possible, and more economic of infants' representational resources. The first behavior-rule is that agents will look for objects where they last saw them roll to. The second rule is that agents will look for objects where they last felt them roll to.

Similarly, Scott et al.'s (2010) data indicating toddlers' coherent VOE responses across two belief-formation scenarios (standard and reversed false-belief property induction) for anticipating what agents will *do* to reproduce interesting sounds (1 belief-use situation) is also qualified by behavior-rules when subjected to such a formal cognitive analysis. A ToM-rules account would require 3 (2+1) abstractions to cover all correct predictions for what agents' actions might be. First, if agents are ignorant of a dissimilar test item sharing non-obvious properties with the pivot item, they will have a false-belief of the hidden property of the perceptually similar test item. Second, if agents are ignorant of the insides of the similar and dissimilar test items being reversed, they will have a false-belief of the hidden property of the dissimilar test item. Combining across scenarios, the third ToM-rule stipulates that if agents have a false-belief of the similar or dissimilar items they will pick up those respective items to reproduce an interesting sound. However, only two (2×1) complementary behavior-rules are needed to account for coherence in looking times: agents will reach for matching items to reproduce an interesting sound; and agents will reach for items that they last saw/heard made an interesting sound to reproduce an interesting sound. We believe that He et al. (2011), Scott et al. and Träuble et al. (2010) are on the right track; but they only combined multiple belief-inducing situations with a *single* use of belief (e.g., predicting where an agent will *look* for an object or predicting what an agent will *do* to produce an effect). In such instances, a ToM-rules account would have an explanation deficit whilst a behavior-rules account would have a natural explanation by being more economic of representational resources ($f+1$ rules$>f \times 1$ rules, respectively).

Based on the extant findings, it is premature to conclude that theoretical parsimony favors an early mentalism explanation per se. Advocates of either a behavior-rule account or a ToM account can claim to have theoretical parsimony on their side, depending on how one treats stimulus-action and mental-state rules. We also cannot straightforwardly evaluate between the relative parsimony of behavior-rule compared to mentalistic models when children's information processing systems are not the outcomes of a rational design but a messy evolutionary process (Apperly and Butterfill 2009; Perner 2010, 2011). Plausibility arguments in and of themselves are not conclusive; they are only relevant considerations for guiding research. We need more decisive empirical data to differentiate between mentalistic and behavior-rule accounts.

5 Potential Approaches for Differentiating Mentalism and Behavior-rules

Depending on whether researchers focus on spontaneous looking responses or elicited verbal responses, existence of the ability to attribute false-belief mental states ranges from 7-months to 4-years of age (Kovács et al. 2010; Wellman et al. 2001). Presently, it is difficult to rule out behavior-rule explanations of children's early

spontaneous looking responses. In fact, an argument in favor of behavior-rules is also possible for 4-year-olds. Although 4-year-olds start to give correct verbal predictions on standard false-belief tasks, their verbal justifications still smack of behavior-rule following (e.g., "because that's where he put the chocolate"; "because he couldn't see mother putting it there") (Perner 2011). Wimmer and Mayringer (1998) reported that mentalistic justifications only became prominent from 6-years of age (e.g., "because he thinks the ice cream man stays there"). Nonetheless, in the current literature, it is generally agreed that children from their 4th birthday do show ToM understanding. Scientists' agreement over such a conclusion stems from decades of research showing that 4-year-olds are able to flexibly master different belief-formation situations (e.g., unexpected transfer, unexpected contents, misinformation, appearance-reality) and, further, are able to deploy their attributions for different uses or demands (e.g., judging where agents will look for, where they will think, what they will say, where they will point to) (Wellman et al.). The fact that there is no great variation occurring at this onset age for solving different tasks with different demands is important in terms of constraining a behavior-rule account. As Apperly (2011) pointed out, "it would be truly surprising if children [at this age] just happened to learn simultaneously a whole set of behavioral rules suitable for predicting agents' actions on the basis of false-beliefs" (p. 42). Evidence of 4-year-olds' verbal predictions cohering across a variety of tasks combining different belief-formation scenarios with belief-use situations around the same time partly provides strong evidence that children at this age *are* reasoning about mental content.

By the same token, if we wish to conclude that children younger than 4-years possess an understanding of ToM at some level, we also need to provide evidence of infants and young preschoolers passing tests of coherence. Specifically, we will need to provide evidence that infants and young preschoolers' spontaneous looking responses cohere across different false-belief inducing situations combined with different response demands at the same time. Let us canvas one such scenario. Consider the idea of testing whether 3-year-olds' spontaneous false-belief responses are robust across interactions between three different belief-inducing scenarios (e.g., unexpected contents, unexpected transfer, and misinformation) *and* three different desire-inducing situations (e.g., anticipating through the "I wonder" prompt where agents would look for an item, say where an item is, or point to where an item is) (see Fig. 1).

If a sample of 3-year-olds showed robust and correct AL or VOE responding across all of these combinations at the same time, a mentalistic account would have a natural explanation by needing only to posit a total of 6 (3+3) ToM-rules (see Fig. 1, bottom panel). In contrast, a behavior-rules account would have an explanatory deficit since a total of 9 (3 x 3) such rules would be needed to cover the range of evidence (see Fig. 1, top panel). If researchers were to discover that young preschoolers' pass in their spontaneous looking on all of the combinations at the same time, it would suggest that the cognitive system can compute how different epistemic and connative states interact and generalize to an invariant and intervening mental state (false-belief) which, in turn, supports correct action prediction in all cases. If such evidence were obtained, we could say that young children possess implicit mentalistic understanding because it is unlikely that children acquire all of the 9 behavior-rules at the same time.

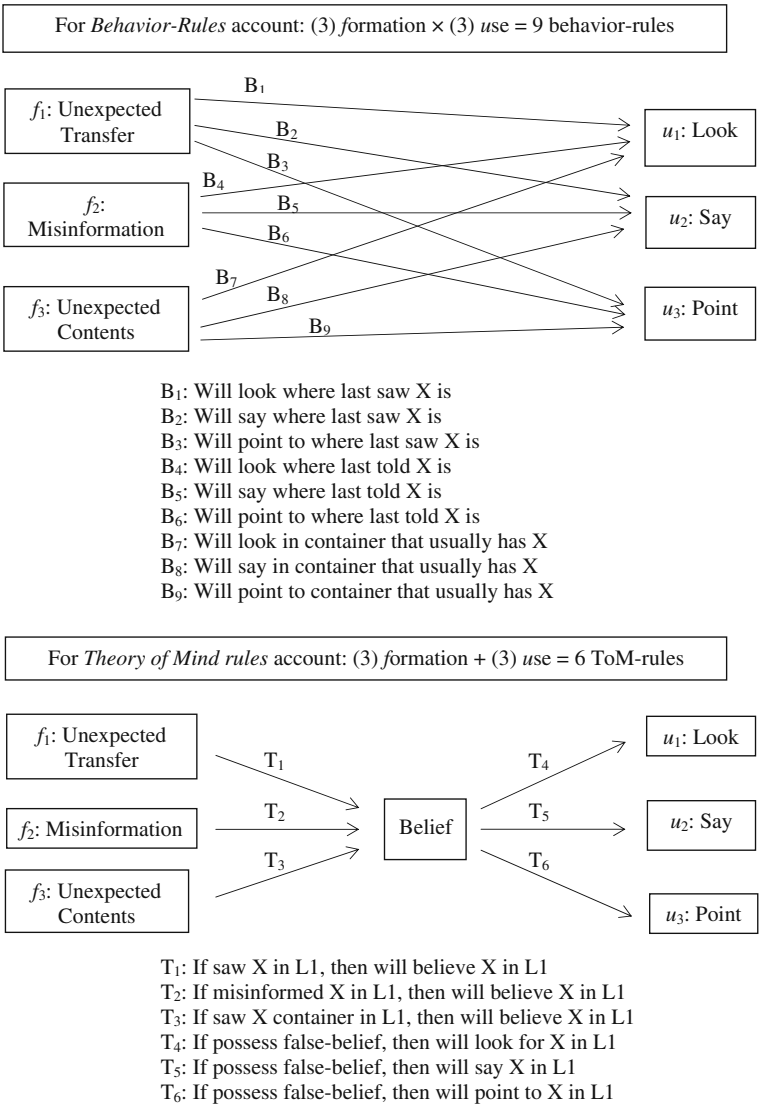


Fig. 1 Differences in complexity of behaviour-rules (B) and ToM-rules (T) in a 3 (belief-*f*ormation) by 3 (belief-*u*se) combination

It should be clear, then, studying visual orienting or visual fixation across multiple belief-formation scenarios combined with multiple belief-use scenarios is at least one critical step for diagnosing when young candidate mind-readers show implicit recognition of states of mind (see Fig 2).

Whilst researchers have investigated spontaneous responding across different belief-formation scenarios, they have only required children to put their belief attribution to a single use (e.g., anticipating where agents will search/look for a desired object) (e.g., He et al. 2011). Further, whilst infants succeed at VOE, helping, and referential communication tasks wherein each of the 3 tasks required belief

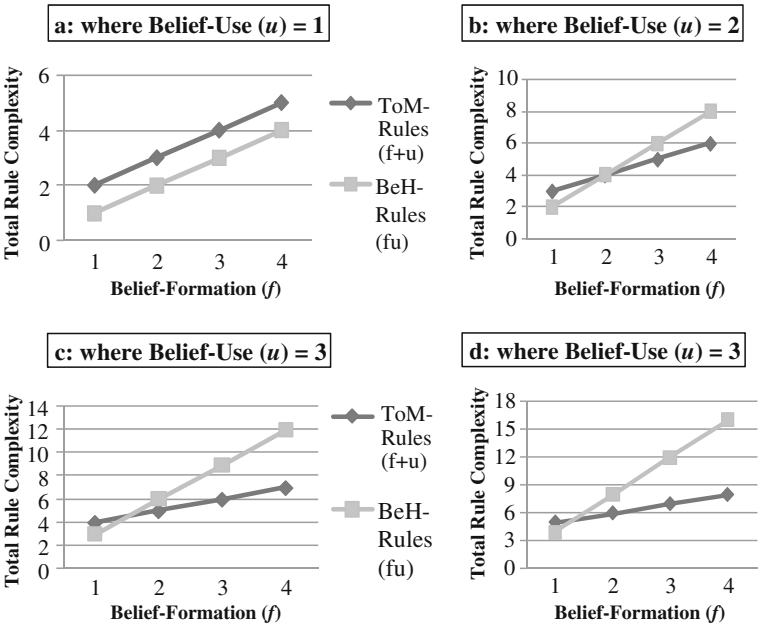


Fig. 2 Changes in total rule complexity for predicting coherent responses across different belief-formation (*f*) by belief-use (*u*) combinations as stipulated in a ToM-rules account (*f*+*u*) compared to a behavior-rules account (*f*×*u*). In (**a**) where looking responses are based on $f \geq 1$ and $u=1$, a behavior-rules (BeH) account is consistently economical of representational resources. In (**b** and **c**) a ToM-rules account is more economical of representational resources, as compared to a BeH-rules account, if $u=2$ and $f \geq 3$, or $u=3$ and $f \geq 2$, respectively. In (**d**) the complexity of computations needed for covering breadth of correct response changes dramatically in favor of a representationally economical ToM-rules account when large batches of belief-use situations ($u \geq 4$) are combined with multiple belief-formation scenarios ($f \geq 2$)

attribution to be put to a different use, all of the tasks involved only 1 belief-formation scenario—unexpected transfer (Buttelmann et al. 2009; Onishi and Baillargeon 2005; Southgate et al. 2010). Even in these cases, if we adopt the specific cognitive analysis introduced here, behavior-rules will always turn out to be more economical of neural resources underpinning the coherence of looking responses (see Fig. 2a). The empirical turning-point favoring ToM-rules over behavior-rules, however, is revealed if and when infants and young preschoolers show accurate spontaneous looking responses simultaneously across at least 2 different belief-inducing situations combined with at least 3 different desire-inducing situations, or vice versa. At these break points, a ToM-rules account would only need to posit a total of 5 abstractive links (2+3 or 3+2) to explain breadth of spontaneous false-belief responding, as compared to the total of 6 links needed under a behavior-rules account (2×3 or 3×2) (see Fig. 2b, c). Of course, a ToM-rules account would have significant edge over a behavior-rules account by being substantially economical of representational resources if individual infants and young preschoolers turn out to be even capable of spontaneously classing large batches of belief-formation scenarios as leading to the same mental state and, further, putting this classification to an armamentarium of uses (e.g., where belief-formation is ≥2 and belief-use is ≥4) (see Fig. 2d).

At the time of writing, we are aware that Wang et al. (2011a, b) adopted the specific cognitive analysis outlined here and used a within-subjects design to test whether Chinese preschoolers (3- and 4-year-olds) showed accurate spontaneous false-belief understanding across two belief-formation scenarios (misinformation and unexpected contents) combined with three belief-use situations (anticipating where an agent will look, say or point to where a desired item was located). Wang et al. used a neutral Mandarin translation of the verbal "I wonder" prompt to trigger children's AL. As far as we know, they are also the first to document spontaneous AL-based false-belief understanding amongst children growing up in a non-Western culture learning to speak a non-Indo-European language. Wang et al. found that, despite extant indications of Chinese children possessing a surfeit of executive functioning abilities (see Sabbagh et al. 2006) that could enhance their expression of ToM understanding, 3-year-old Chinese preschoolers consistently showed correct first looks across the full range of belief-formation by belief-use combinations just as they consistently erred on elicited judgments. The consistency of 4-year-olds' accurate AL responses dovetailed with their correct verbal answering. It is promising that Wang et al.'s Chinese data confirmed accurate AL simultaneously spreading across a 2 (belief-formation) by 3 (belief-use) combination—thus suggesting mentalism (5 ToM-rules), as compared to behaviorism (6 stimulus-action rules), as being more economical of young preschoolers' representational resources. It remains unknown as to whether 2-year-olds (in Mainland China or elsewhere) would pass such stringent tests of coherence. Extrapolating from current directions in the work of Baillargeon and colleagues (e.g., He et al. 2011; Scott et al. 2010) and Wang et al., we are hopeful that it is only a matter of time when VOE and AL researchers boldly go on to discover that the very same infants and toddlers are also capable of making accurate spontaneous responses across multiple belief-formation by multiple belief-use combinations. The field can make significant advances in exploring these issues if we adopt an explicit cognitive analysis that arbitrates in a principled manner when early spontaneous responses to false-belief tasks are mentalistic and when they are not. As argued here, we cannot just yet confidently ascribe mentalism over smart behavior rule following to the full range of evidence on infants, toddlers and young preschoolers' spontaneous responses in false-belief tasks. We can get there if we start to gauge children's spontaneous responses against the empirical yardstick of minimally showing understanding across different belief-situation by desire-purpose combinations required to rule out application of behavior-rules.

We have canvassed the promising solution of designing experiments to test the coherence of infants and toddlers' spontaneous responses so as to differentiate which account is more economical of neural resources, thereby being the more efficient account. But as acknowledged earlier, evolution and development do not always yield the most rational or economical solutions to problems. Whilst tests of coherence for differentiating which account requires the fewest rules (and thus the most parsimonious) are relevant, they are not decisive. Other kinds of converging empirical evidence will also be needed to indicate that behavior-rules may be insufficient to explain all instances of efficient mentalistic reasoning. Other approaches include documenting: whether dual task interference has an impact on explicit verbal responses but not AL responses (Newton and de Villiers 2007); whether there are positive transfer effects from trained to untrained false-belief tasks

(for transfer tests to be potent, novel belief-formation by belief-use scenarios will be needed so that researchers can gain some degree of control over the acquisition of children's knowledge) (Perner 2010); and whether there are real signature upper limits to the flexibility by which infants and young preschoolers can apply and generalize their early tacit ToM understanding (Apperly and Butterfill 2009). We turn to spotlight the issue of limits in spontaneous knowledge because it is analogous to our discussion of the issue of coherence in spontaneous understanding across combinations of belief-formation by belief-use scenarios.

If one is right to interpret that there exists two types of false-belief reasoning systems—an early developing implicit representational system that has a modular basis and a slowly emerging explicit representational system that is partly underpinned by important growths in executive functioning and language ability (Low 2010; Low and Simpson 2011)—then modular understandings must exhibit efficiency at the expense of flexibility (Apperly and Butterfill 2009). Apperly and Butterfill suggested that the economy/efficiency inherent in children's implicit understanding of mental states may be achieved by representing only links between agents, objects and properties and *avoiding* complexities entailed in processing links between agents and propositions. An implicit representational system would then allow for the processing of links between visual access, knowledge or false-belief and consequent behavior (e.g., Baillargeon et al. 2010; Kovács et al. 2010; Senju et al. 2011). At the same time, however, implicit mindreading should not be expected to support processing of Level-2 types of mental representation problems which require recognition that an object may present different visual appearances to two people if they view it from different positions (Apperly 2011). Indeed, Samson et al. (2010) found that in a Level-1 visual perspective taking task that required tracking what an agent can or cannot see, adults exhibited egocentric interference when judging how many dots an agent could see (participants were slower at judging the agent's perspective when they themselves saw more dots). Participants also showed altercentric interference when judging how the number of dots they themselves could see (adults were slower when the agent saw fewer dots than when the agent saw the same number). In this case, altercentric interference is an important finding because it shows that adults are implicitly and automatically computing the agent's perspective even when there was no need to do so. Nonetheless, there appears to be limits to the mileage in implicit automatic mental-state processing. In a Level-2 type of task that required representation of the particular way in which an agent thinks about something (e.g., a digit is read as "9" from the participant's perspective but read as "6" from the agent's point of view), however, Surtees et al. (2010, cited in Apperly 2011) found no evidence of altercentric interference: adults only showed egocentric interference when judging the agent's perspective. The presence of altercentric interference in the Level-1 but not Level-2 tasks suggests that whilst we do implicitly and automatically process an agent's perspective, there are limits to the deployment of implicit mentalistic understanding. The study of limits in implicit mentalistic understanding is still in its early stage. It will be important to suitably replicate the null results for automatic Level-2 perspective processing (and in more than one paradigm) with infants and young preschoolers. As such, it will be critical to discover whether the implicit mental-state reasoning system available to infants and young preschoolers will efficiently cohere across a range of spontaneous false-

belief tasks but will distinctively not support the spontaneous processing of propositional content for a range of Level-2 ToM tasks. To echo Apperly and Butterfill, the discovery of signature limits in children's spontaneous responses to ToM tasks is another powerful technique that allows us to tell when an implicit mentalistic reasoning system is and is not being deployed.

6 Going Beyond Spontaneous Visual Responses

Almost all current evidence on children's implicit insight about false-belief rests on eye movements or looking time (Baillargeon et al. 2010; Low 2010). If this were the only way in which children's implicit understanding was manifest, then we would not expect this early conceptual insight to have much impact on their everyday behavior. To assess what pre-linguistic children understand of others and their worlds, some researchers have begun using interactive tasks, including helping (e.g., Buttelmann et al. 2009) and referential-communication paradigms (e.g., Southgate et al. 2010). An important characteristic of such interactive paradigms is that infants and children are not required to predict what an agent will do by looking to the correct location, but they are required to initiate a more active behavioral response. For example, in the false-belief condition of Buttelmann et al., an experimenter showed infants (16- and 18-month-olds) two lidded boxes and demonstrated how to lock and unlock them, and thereupon the boxes were left unlocked. Next, a male agent entered the room, hid a toy in one of the boxes, and then left. While the agent was absent, the experimenter moved the toy to the other box and locked both boxes. When the agent returned, he tried to open the box where he had hidden the toy, without success, and then sat centered behind the boxes. The rationale was that if participants understood the agent's false-belief and wanted to help, they should infer that he wanted the toy he thought was in there. In this case, children should not go and help him open the first box but, instead, go to the other box and extract the desired item for him. Buttelmann et al. found that when prompted to help the agent, the majority of 18-month-olds approached the other box containing the toy which the agent did not approach.

Southgate et al. (2010) provided convergent evidence using a referential-communication task to test whether 17-month-olds were able to understand a communicator's false-belief and, further, use this understanding in a pragmatic context. The infants watched as a female agent hid two different toys in two lidded boxes and then left. While she was gone, an experimenter switched the toys. The agent then returned, pointed to one of the boxes, and announced that the toy inside it was a "sefo". Southgate et al. reported that when prompted to get the sefo, most infants approached the other box (i.e., the one that the agent did not point to). There is narrow age range tested across these interactive paradigms (17- and 18-months) to promisingly suggest that infants might be able to coherently deploy their false-belief understanding across both helping and communication contexts.

That said, it will still be important to test whether the very same infants (i.e., using a within-subjects design) are able to use their false-belief understanding actively across both task contexts. Indeed, in Buttelmann et al.'s (2009) study, 16-month-olds performed randomly in the true-belief condition as compared to their 18-month-old

counterparts. Since 16-months-olds failed to attribute the agent's mental state in the true-belief condition, their accurate responses in the false-belief condition may not be fully indicative of understanding false-belief per se. This complication raises the possibility that 16-month-olds might not necessarily transfer an apparent false-belief understanding from an inactive looking context (as in the VOE scenarios) to an active helping context or vice versa. Demonstrating that infants' spontaneous responses cohere across inactive and active false-belief inducing situations at the same time will be partly important for ascribing mentalism per se. Buttelmann et al. suggested that 16-month-olds might possess a genuine understanding of others' mental state, but had difficulty inhibiting going for the target toy in the true-belief condition. Scott et al. (2010) have also raised this possibility, suggesting that interactive situations might require more inhibitory control compared to spontaneous visual responding. Scott et al. argued that in the spontaneous visual response paradigms such as VOE and AL, children are passive observers, who can focus solely on the agent's perspective on the scene, with their own perspectives being less salient. On the other hand, interactive tasks may impose higher levels of inhibitory control because children need to actively hold in mind both their own and the agent's perspectives. Future studies investigating whether children's inhibitory skills predict their performance across non-interactive and interactive response tasks can provide insight into the processing load underlying these tasks. Nonetheless, helping and referent retrieval goals can be the basis of generating different belief-use situations in spontaneous false-belief tasks so that researchers are not focused on measuring only children's eye gazing responses. To that aim, documenting whether looking responses can simultaneously spread to accurate helping responses across diverse belief-inducing and desire-inducing tasks will help establish when young children grasp false-belief and how such understanding is implemented.

7 Conclusion

Overall, whilst extant work on early false-belief understanding is suggestive, the evidence is not yet compelling. There is currently insufficient evidence of young children's visual fixations or anticipations cohering across different situation-purpose combinations so as to rule out learning of behavior-rules. Although this kind of evidence is hard to come by, researchers are now starting to analyze whether spontaneous understanding—in children of a given age cohort and amongst the very same children—coheres across different false-belief inducing situations combined with different response demands. An explicit cognitive analysis of how early understanding is structured and spread can partly help us take a step in the right direction on the long road to answering how early in development children mentalistically and implicitly grasp false-belief.

Acknowledgments We are grateful to Josef Perner for helpful theoretical encouragements. We also thank Victoria Southgate and two anonymous reviewers for their generous comments.

References

Apperly, I. 2011. *Mindreaders: The cognitive basis of "theory of mind"*. Hove: Psychology.

Apperly, I.A., and S.A. Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological Review* 116: 953–970.

Baillargeon, R., S.M. Scott, and Z. He. 2010. False-belief understanding in infants. *Trends in Cognitive Science* 14: 110–118.

Buttelmann, D., M. Carpenter, and M. Tomasello. 2009. Eighteen-month-olds show false beliefs understanding in an active helping paradigm. *Cognition* 112: 337–342.

Clements, W.A. 1995. *Implicit theories of mind*. Unpublished doctoral dissertation, University of Sussex.

Clements, W.A., and J. Perner. 1994. Implicit understanding of belief. *Cognitive Development* 9: 377–395.

Csibra, G., and V. Southgate. 2006. Evidence for infants' understanding of false beliefs should not be dismissed. *Trends in Cognitive Science* 10: 4–5.

Gopnik, A., and J.W. Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development* 59: 26–37.

He, Z., M. Bolz, and R. Baillargeon. 2011. False-belief understanding in 2.5-year-olds: Evidence from change-of-location and unexpected-contents violation-of-expectation tasks. *Developmental Science* 14: 292–305.

Kovács, A.M., E. Téglás, and A.D. Endress. 2010. The social sense: Susceptibility to others' beliefs in human infants and adults. *Science* 330: 1830–1834.

Leslie, A.M. 2005. Developmental parallels in understanding minds and bodies. *Trends in Cognitive Sciences* 9: 459–462.

Low, J. 2010. Preschoolers' implicit and explicit false-belief understanding: Relations with complex syntactical mastery. *Child Development* 81: 579–615.

Low, J., and S. Simpson. 2011. Effects of labeling on preschoolers' explicit false-belief performance: Outcomes of cognitive flexibility or inhibitory control? *Child Development*.

Moses, L.J. 2001. Executive accounts of theory of mind development. *Child Development* 72: 688–690.

Newton, A.M., and J.G. de Villiers. 2007. Thinking while talking: Adults fail nonverbal false belief reasoning. *Psychological Science* 18: 574–579.

Onishi, K.H., and R. Baillargeon. 2005. Do 15-month-old infants understand false beliefs? *Science* 308: 214–216.

Perner, J. 2010. Who took the cog out of cognitive science? Mentalism in an era of anti-cognitivism. In *Perception, attention, and action: International perspectives on psychological science (Volume 1)*, ed. P.A. Frensch & R. Schwarzer, 241–261. Psychology Press.

Perner, J. 2011. Theory of mind—an unintelligent design: From behaviour to teleology and perspective. In *Handbook of theory of mind*, ed. A.M. Leslie and T.C. German. NJ: Erlbaum.

Perner, J., and T. Ruffman. 2005. Infants' insight into the mind: How deep? *Science* 308: 214–216.

Ruffman, T., W. Garnham, A. Import, and D. Connolly. 2001. Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology* 80: 201–224.

Sabbagh, M.A., F. Xu, S.M. Carlson, L.J. Moses, and K. Lee. 2006. The development of executive functioning and theory of mind. *Psychological Science* 17: 74–81.

Samson, D., I.A. Apperly, J.J. Braithwaite, B.J. Andrews, and S.E. Bodley Scott. 2010. Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology. Human Perception and Performance* 36: 1255–1266.

Scott, R.M., and R. Baillargeon. 2009. Which penguin is this? Attributing false beliefs about identity at 18 months. *Child Development* 80: 1172–1196.

Scott, R.M., R. Baillargeon, H. Song, and A.M. Leslie. 2010. Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology* 61: 366–395.

Senju, A., V. Southgate, C. Snape, M. Leonard, and G. Csibra. 2011. Do 18-month-olds really attribute mental states to others? A critical test. *Psychological Science*.

Song, H., and R. Baillargeon. 2008. Infants' reasoning about others' false perceptions. *Developmental Psychology* 44: 1789–1795.

Song, H., K.H. Onishi, R. Baillargeon, and C. Fisher. 2008. Can an agent's false belief be corrected through an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition* 109: 295–315.

Southgate, V., A. Senju, and G. Csibra. 2007. Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science* 18: 587–592.

Southgate, V., C. Chevallier, and G. Csibra. 2010. Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science* 13: 907–912.

Surian, L., S. Caldi, and D. Sperber. 2007. Attribution of beliefs by 13-month-old infants. *Psychological Science* 18: 580–586.

Träuble, B., V. Marinović, and S. Pauen. 2010. Early theory of mind competencies: Do infants understand others' belief? *Infancy* 15: 434–444.

Wang, B., J. Low, J. Zhang, and Q. Qinghua. 2011a. Chinese preschoolers' implicit and explicit false-belief understanding. *British Journal of Developmental Psychology: Special Issue on Implicit-Explicit False-Belief Understanding.*

Wang, B., J. Low, J. Zhang, and Q. Qinghua. 2011b. Chinese 3-year-olds show coherent false-belief anticipatory looking across different belief-formation and belief-use task combinations. *Manuscript in preparation.*

Wellman, H.M., D. Cross, and J. Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development* 72: 655–684.

Whiten, A. 1996. When does smart behavior-reading become mind-reading? In *Theories of theories of mind*, ed. P. Carruthers and P.K. Smith, 277–292. Cambridge: Cambridge University Press.

Wimmer, H., and H. Mayringer. 1998. False belief understanding in young children: Explanations do not develop before predictions. *International Journal of Behavioral Development* 22: 403–422.