Cognition 115 (2010) 54–70

Contents lists available at ScienceDirect

Cognition

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

Two sources of evidence on the non-automaticity of true and false belief ascription

Elisa Back [a,*], Ian A. Apperly [b]

[a] Psychology Research Unit, Kingston University London, Kingston upon Thames, Surrey KT1 2EE, United Kingdom
[b] School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

ARTICLE INFO

Article history:
Received 12 August 2008
Revised 8 September 2009
Accepted 17 November 2009

Keywords:
Theory of mind
True and false belief
Adult

ABSTRACT

A recent study by Apperly et al. (2006) found evidence that adults do not automatically infer false beliefs while watching videos that afford such inferences. This method was extended to examine true beliefs, which are sometimes thought to be ascribed by "default" (e.g., Leslie & Thaiss, 1992). Sequences of pictures were presented in which the location of an object and a character's belief about the location of the object often changed. During the picture sequences participants responded to an unpredictable probe picture about where the character *believed* the object to be located or where the object was located in *reality*. In Experiment 1 participants were not directly instructed to track the character's beliefs about the object. There was a significant reaction time cost for belief probes compared with matched reality probes, whether the character's belief was true or false. In Experiment 2, participants were asked to track where the character thought the object was located, responses to belief probes were faster than responses to reality probes, suggesting that the difference observed in Experiment 1 was not due to intrinsic differences between the probes, but was more likely to be due to participants inferring beliefs ad hoc in response to the probe. In both Experiments 1 and 2, responses to belief and reality probes were faster in the true belief condition than in the false belief condition. In Experiment 3 this difference was largely eliminated when participants had fewer reasons to make belief inferences spontaneously. These two lines of evidence are neatly explained by the proposition that neither true nor false beliefs are ascribed automatically, but that belief ascription may occur spontaneously in response to task demands.

1. Introduction

Theory of mind (ToM) is the ability to reason about mental states such as beliefs, desires, and knowledge. It is widely believed to be central to a range of cognitive activities including our ability to communicate and to explain and predict behaviour (e.g., Baron-Cohen, Tager-Flusberg, & Cohen, 2000; Sperber, 2000). For ToM to serve such functions many authors have supposed that ToM must be computationally efficient and perhaps automatic in its operation (e.g., Leslie & Thaiss, 1992). However, the majority of investigations of ToM have focussed on typically developing children of various ages (e.g., Mitchell & Riggs, 2000), individuals with developmental disorders such as autism (e.g., Baron-Cohen, Leslie, & Frith, 1985), and different species (Tomasello, Call, & Hare, 2003), and these studies provide little direct evidence about the basic cognitive characteristics of ToM. The current research examines adults' true and false belief reasoning to investigate whether these ToM processes occur automatically, spontaneously or deliberately.

Developmental research yields conflicting perspectives on how belief reasoning abilities should be characterised. One view is that, as a child, we learn a wealth of practical knowledge about mental states such as beliefs and desires

[*] Corresponding author. Tel.: +44 (0)208 4172831.
E-mail address: e.back@kingston.ac.uk (E. Back).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.11.008

(e.g., Gopnik & Wellman, 1992; Perner, 1991) that we use to explain and predict behaviour using general reasoning processes. This view would suggest that theory of mind is a relatively ‘high-level’ reasoning process – a “central process” in Fodor’s terminology (Fodor, 1983), likely to depend upon scarce cognitive resources for memory and executive control, and unlikely to be automatic. An opposing view is that theory of mind is the function of one or more specialised innate processing mechanisms (Baron-Cohen, 1989; Carruthers & Smith, 1996; Leslie & Thaiss, 1992) suggesting that theory of mind is a fast, automatic and domain-specific; a “low-level” or “modular” process in Fodor’s terminology (Fodor, 1983, 2000). Indirect evidence in favour of this account arises from studies of infants where 15-month-olds have been found to understand that others can have false beliefs (e.g., Onishi & Baillargeon, 2005). Further evidence also comes from past research which suggests individuals with autism are selectively impaired on theory of mind tasks (e.g., Baron-Cohen, Leslie, & Frith, 1985). However, studies involving children have not yielded any direct evidence on the automaticity of belief ascriptions.

One study of adults provides data that are consistent with automatic inferences about beliefs and desires. Wertz and German (2007) investigated adults’ ability to explain the action of agents according to beliefs and desires. Participants read short stories in which characters acted according to a false belief. For example, Mary places a hairdryer next to her perfume in a drawer and leaves the room. While Mary is absent Gina moves the hairdryer to the cabinet. Mary returns for her hairdryer and goes directly to the drawer. Participants had to endorse or reject sentences explaining the character’s action. It was found that adults were prone to endorse incorrect mental state explanations that referred to a distracter object when the distracter object was approached by an agent (e.g., adults were prone to endorse ‘because she wanted to get her perfume from the drawer’, despite the fact that Mary approached the drawer on the false belief that it contained her hairdryer). Wertz and German (2007) suggest that a “theory of mind module” automatically produces explanations when the participant sees an agent approach an object, which involve a desire for the object and a true belief about the object’s location. The authors argue that participants are prone to make errors by endorsing a statement that fits with one of these automatically generated explanations for the agent’s behaviour.

Importantly however, these findings do not guarantee that inferences about the agent’s mental states were generated automatically. Since every story presented to participants and every judgement of an explanation for the story character’s behaviour concerned the character’s mental states it is plausible that participants spontaneously made inferences about the mental states of the character that they might not have made if these mental states had not seemed so relevant. Spontaneous inferences are a well-attested phenomenon during text comprehension but this is distinct from the view that these inferences are automatic (see e.g., McKoon & Ratcliff, 1998; Sanford & Garrod, 1998; Zwaan & Radvansky, 1998). These inferences are viewed as spontaneous because participants need not be instructed to make them. They are not viewed as automatic because their processing is contingent on a variety of contextual factors: participants may read and remember the text without necessarily making inferences that go beyond the text. In fact, Wertz and German’s findings do not guarantee that any mental states were inferred in advance of the test question, either automatically or spontaneously. Participants’ tendency to give incorrect endorsements about what the character thought or wanted could be explained if they simply evaluated the test question against their memory for the physical events described in the story. It seems plausible that participants could be distracted by the fact that the sentence gave a plausible explanation for the character’s most recent behaviour, even though this was an incorrect explanation if the rest of the story was taken into account. Thus, Wertz and German’s findings are interesting, and clearly warrant further investigation. But they do not lead to any strong conclusions about the automaticity of belief inferences.

In contrast, several studies suggest that information about another person’s belief or knowledge is not automatically used to interpret their behaviour. Keysar, Lin, and Barr (2003, see also, Keysar, Barr, Balin, & Brauner, 2000) investigated theory of mind use in adult participants. They followed the instructions of a speaker who directed them to move items around a grid. In the first experiment, some of the items in the grid could only be seen from the participant’s position, and participants knew that the speaker was unaware of the existence of these items. However, participants often ignored this, selecting items that the speaker could not see when they were the best referent for the speaker’s instruction. The second experiment showed similar effects when the speaker had a false belief about the identity of an item in the grid. Thus, although adults clearly knew that the speaker’s perspective differed from their own, and did use this information to guide their interpretation on some trials, adults suffered strong interference from their own perspective and so often failed to take the speaker’s knowledge into account.

However, although the findings from these studies suggest that the *use* of information about what someone else thinks may not be automatic, they say nothing about the processes by which that information is generated in the first place. That is to say, there is no contradiction between Wertz and German’s (2007) claim that beliefs are inferred automatically and Keysar et al.’s (2003) claim that beliefs are not used automatically once they have been inferred.

Direct evidence against the automaticity of belief inferences comes from a study by Apperly, Riggs, Simpson, Chiavarino, and Samson (2006). The processing costs of making belief inferences was examined by presenting participants with videos involving a male actor sometimes moving an object from one box to another while a female actor is present or absent and subsequently has a true or false belief about the object’s location. On each trial participants had to respond to an unpredictable probe question that could concern “reality” (the objects and events in each trial) or the beliefs of the male or female actors about objects and events in the scenario. Response times to these probes were the critical measure. In two conditions participants were explicitly instructed to track the female actor’s belief about the object's location, and in these conditions participants found it no harder to respond to a probe about the female character's belief than to a probe about reality. In a further condition – the "incidental false belief condition" – participants viewed exactly the same video stimuli, but were not instructed to keep track of the female character's belief about the object's location. In this case participants responded slower to belief probes than to reality probes, suggesting that when participants had no particular reason to keep track of the female character's beliefs they did not do so, and consequently showed a processing cost when they had to infer the female character's belief ad hoc in response to the probe. These results suggest that false beliefs may not be ascribed automatically. Importantly though, they do not say anything about true beliefs.

Cohen and German (2009) proposed an alternative explanation for why participants might have been slow to respond to belief probes in Apperly et al.'s (2006) incidental false belief condition. They pointed out that the female character's belief needed to be encoded earlier in the event sequence than "reality" information about the object's true location. This raises the possibility that her belief was inferred automatically but whereas this information was retained in the conditions where participants were explicitly instructed to do so it was not retained in the incidental false belief condition. Cohen and German (2009) devised a new condition that shortened the interval over which information about belief might need to be retained, and found that responses to belief probes were now faster than responses to reality probes. This result is consistent with the authors' hypothesis that participants automatically inferred the female character's belief, but it is also consistent with these inferences being made spontaneously. The latter interpretation seems quite plausible because Cohen and German's new condition involved the female character acting in error directly before the belief probe, which may well have prompted participants to infer her false belief spontaneously as an explanation for her behaviour. In the new studies described below we used event sequences based upon Apperly et al.'s original procedure. We find evidence suggesting that information about the female character's belief may indeed be inferred spontaneously and this information is actually retained for long enough to affect participants' later responses to probes.

Further evidence consistent with the conclusion that belief inferences are not made automatically comes from a neuroimaging study by Saxe, Schultz, and Jiang (2006). These authors examined brain activity while participants viewed cartoon stimuli in which a character either did or did not see an object displaced from its original location to a new location. Importantly, brain activation was compared across two conditions: in one condition participants were instructed to reason about where the character thought the object was located; in a second condition they were instructed to reason according to the character's spatial orientation and the temporal order of events. Thus, in both conditions the stimuli afforded a belief inference, but one set of instructions explicitly prompted the inference whereas the other set of instructions did not. Results showed that in brain regions most specifically associated with belief reasoning according to independent tests (right Temporo Parietal Junction in particular, but to a lesser extent left-TPJ and posterior cingulate), significant activation was observed when instructions prompted belief inferences but not when the instructions only prompted inferences about physical objects. That is to say, although the cartoon stimuli always afforded a belief inference, there was no evidence that participants actually reasoned about the beliefs of the cartoon character unless they were directed to do so by the task instructions. Although the data in this study came from trials in which the cartoon character had true beliefs as well as when it had false beliefs, this factor was not analysed, so we cannot be sure that the conclusions from this study extend to both false beliefs and true beliefs. Thus, in the current study we sought for the first time to examine directly whether true beliefs are ascribed automatically.

Why is it important to test true beliefs as well as false beliefs? Several authors have argued on both empirical and theoretical grounds that true beliefs are ascribed by default (e.g., Fodor, 1992; Leslie & Thaiss, 1992). The main empirical motivation for this claim comes from evidence of a systematic bias in the errors made by children on false belief tasks: When children fail to make a correct prediction about what someone with a false belief will think or do children do not guess; they systematically judge that the person in fact has a true belief and will act accordingly (see e.g., Wellman, Cross, & Watson, 2001). Furthermore, this bias may be reflected in response times as well as errors (Kikuno, Mitchell, & Ziegler, 2007). This evidence is reenforced by similar findings of biases in the judgements of adults (see e.g., Birch & Bloom, 2007; Epley, Morewedge, & Keysar, 2004; Keysar et al., 2003; Mitchell, Robinson, Issacs, & Nye, 1996). Indeed, recent evidence also suggests that there is a specific processing cost associated with holding in mind information about false beliefs compared with similar information about beliefs where the truth is unknown (Apperly, Back, Samson, & France, 2008).

The main theoretical motivation for testing true as well as false beliefs comes from the observation that beliefs are *supposed* to be true (i.e., their function is to provide a basis for rational action), and that in everyday matters, most people's beliefs *are* true most of the time (e.g., Fodor, 1992; Leslie, Friedman, & German, 2004). Thus, Fodor (1992) proposes that, all other things being equal, children assume that an agent operates with true beliefs. Similarly, the theory of mind module proposed by Leslie and colleagues is supposed to generate a set of possible beliefs that might explain an agent's behaviour, including a true belief that is ascribed by default. Thus, although a principal motivation for modular accounts is to explain efficient ascription of both true and false beliefs, the strongest claims about automaticity are made for true beliefs, making true beliefs an important test case for empirical investigation. More generally, understanding the processes involved in both true and false belief ascription will be important for explaining the biases observed in the belief reasoning of both children and adults.

The three experiments reported in this article examine whether participants infer and ascribe true beliefs or false beliefs automatically. The method of Apperly et al. (2006) was adapted so that pictures were used instead of videos and picture probes were presented as opposed to text. Response times as well as errors to belief and reality probes in true and false belief contexts were investigated.

2. Experiment 1 – incidental belief task

If belief reasoning is an automatic process then participants should infer beliefs when attending to a stimulus that affords a belief inference, even when there is no reason for doing so. In this experiment, participants were explicitly asked to track the real location of the object and they were not specifically asked to track beliefs (where the female character thinks the object is located). If belief reasoning is automatic then beliefs should be inferred nonetheless, and, all other things being equal, response times to questions about belief should be similar to response times to questions about reality (the object's real location). Alternatively, if beliefs are not inferred automatically then response times to questions about reality will be faster than to questions about beliefs because the belief would have to be inferred ad hoc in response to the probe.

2.1. Method

2.1.1. Participants

Twenty-four undergraduate and postgraduate students from the University of Birmingham participated in this study. Participants were aged between 18 and 34 with a mean age of 22. There were nine males and 15 females. Participants were awarded course credits or paid £10.

2.1.2. Stimuli

Two untrained actors (one male, one female) were asked to act out various theory of mind sequences. The actors sat opposite each other at a table with two boxes and a green object placed in front of them. The actors were instructed how to pose for each photograph (e.g., the male character was asked to remove the object from the left box while the female character was present). A digital camera and tripod were used to capture the images. Pictures were selected that successfully portrayed each theory of mind sequence. Fourteen slides were included in each trial sequence. Each trial began with a blank screen, followed by eight photographs depicting a particular sequence, a question mark slide, a picture probe (see below), a further three photographs depicting movements of the boxes, and an end screen where participants had to locate whether the object was in the left or right box (see Fig. 1 for a summary of the event sequences for experimental trials).

The exact event sequence varied across experimental and filler trials, but what remained consistent is that the male character hid an object in one of two boxes and at some point during the picture sequence the female character indicated where she thought the object was hidden. Participants had to identify the location of the object at the end of each trial by tracking movement of the object/boxes and taking into account whether the female character had a true or false belief when she gave her clue. An example of an event sequence for both true and false belief trials can be seen in Fig. 1. Participants viewed the female character looking in the boxes and then she points to the box that contains the object. The female character's belief is true at this point so participants can infer the object's real location. The female character then leaves the room. In true belief trials she returns immediately to view the male character swapping the object from one box to the other. In false belief trials the male character swaps the object from one box to another before the female character returns to the room. It must be noted that in Experiment 1 the change in the female character's belief was not relevant to the task of locating which box the object was in at the end of the trial. Participants only needed to update their representation of the location of the object. A probe picture would then appear on the screen, either a belief probe (e.g., she thinks the object is in the right box) or a reality probe (e.g., the object is in the right box) or a variety of filler probes that asked about the colour of the object, whether the female or male character thinks the object has been swapped to another box, whether the male character thinks the object is in the left/right box, (see Fig. 2 for examples of experimental picture probes). After the participant responded to the picture probe, the male character either swapped or did not swap the boxes and then a screen appeared where participants were asked to locate the object. The only purpose of this part of the task was to encourage participants to keep track of information relevant for answering reality probes.

2.1.3. Design

There were four blocks of 51 trials and there was an opportunity for participants to have a break at the end of each block. The session included 128 experimental trials and 76 filler trials which consisted of 36 new sequences and 40 sequences that were the same as the experimental trials but with different picture probes. Filler trials were included to reduce the likelihood of participants developing any strategies such as spontaneously inferring the female character's belief about the object location. In each block there were 32 experimental trials (each trial type was presented four times) and 19 fillers that were unique to each block. In the experimental sequences the object began and ended equally often in the left and right hand box, the swap equally often involved moving the object or the boxes, and the female character's beliefs were equally often true or false. The correct answer to picture probes was equally often "yes" and "no" (depending on their correspondence with the sequence), and there were equal numbers of each type of experimental picture probe (e.g., she thinks left/right box; really in left/right box). There were 36 filler trials (new sequences). Sixteen filler items involved the female character indicating where she has seen the object just before the picture probe is presented instead of at the beginning. A further 12 filler items involved the male character swapping the boxes twice where one swap would be seen in the presence of the female character and the other one in her absence. An additional eight filler items involved the male character swapping the object/boxes twice in the presence or absence of the female character.

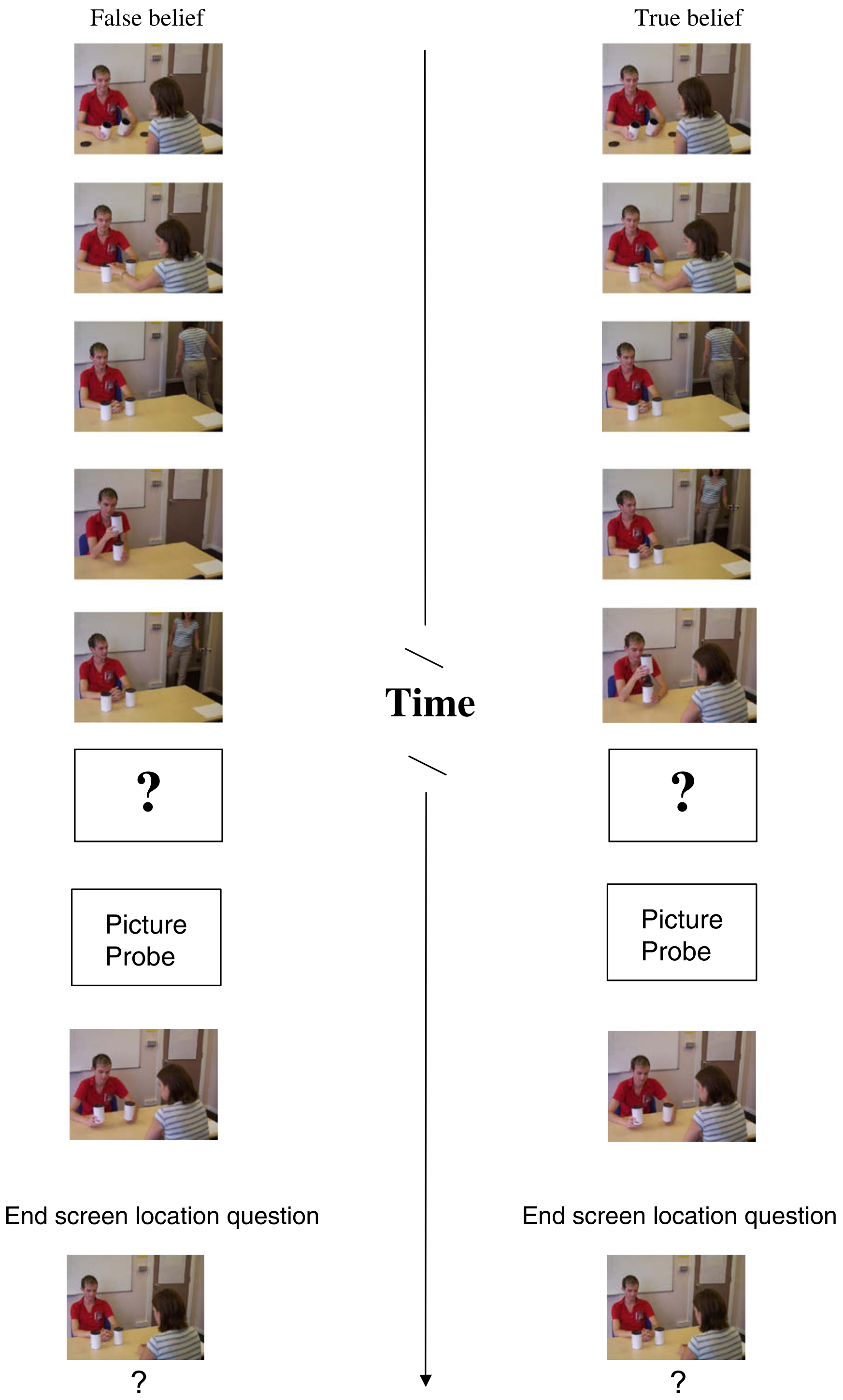

Fig. 1. Examples of a true and false belief experimental sequence. Note that slides that were only used to give additional cues to the movement of the characters/objects have not been included in this figure.

2.1.4. Procedure

Each participant was tested individually in a room. Participants were seated in front of a 15 in. computer screen. The experimenter read out the task instructions and presented examples of trial sequences using a PowerPoint presentation. Each picture in the sequence of slides was displayed for 1500 ms in the experimental task. Participants were informed that during the sequence of slides

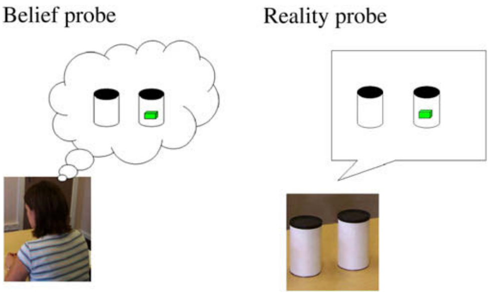

Fig. 2. Picture probes for the experimental belief and reality judgements.

they would be presented with a picture probe and they must decide if it is accurate or inaccurate. Half the participants were asked to press the "N" key on the keyboard if the answer is 'yes' and the "M" key for 'no' and the other half were asked to press the "N" key for 'no' and the "M" key for 'yes'. They were also told that a question mark would appear on the slide before the picture probe. Participants were asked to respond as quickly and accurately as possible. Participants were informed that the picture probe could depict the following: the object's real location (left or right box), whether the female or male character has seen the object or boxes being swapped, where the female or male character thinks the object is (left or right box), what colour she or he thinks the object is. Participants were also asked to locate the object at the end of each trial sequence by left clicking with the mouse if they thought the object was in the left box and to right click if it was in the right box. Again, participants were asked to respond as quickly and accurately as possible.

The experiment was presented on a standard Pentium-based desktop computer using Eprime (http://www.pstnet.com/products/Eprime/). Participants commenced eight practice trials in the presence of the experimenter. These practice trials included an example of each picture probe and a mixture of true and false belief trials. Feedback informing the participant whether his/her response was correct or incorrect was given on the screen after each picture probe and at the end of each trial when participants indicate the object's location. After the participant successfully completed the practice trials and there were no further questions, the experimental session begun. No feedback was given by the Experimenter during the main experiment.

2.2. Results

Participants' response times in Experiment 1 (incidental belief reasoning task) were analysed and Fig. 3 displays the mean response times. For correct responses, response times that were two standard deviations beyond the mean were removed. This criterion resulted in a loss of 105 out of 2872 data points.

A three-way ANOVA on yes and no trials ("yes" trials were when the picture probe was consistent with the event sequence; "no" trials were when the picture probe was inconsistent with the event sequence) × belief type (true or false belief context) × probe type (belief or reality picture probes) was carried out and revealed a main effect of belief type, $F(1,23) = 87.104$, $p < .001$, $\eta p^2 = .791$, participants were faster at true beliefs than false beliefs, as well as a main effect of probe type, $F(1,23) = 11.408$, $p = .003$, $\eta p^2 = .332$, where participants were faster on reality probes than belief probes. Additionally, there was a non-significant trend for faster responses to yes than no trials, $F(1,23) = 3.709$, $p = .067$, $\eta p^2 = .137$. There was a two-way interaction between yes and no trials and probe type, $F(1,23) = 12.326$, $p = .002$, $\eta p^2 = .349$ and a significant three-way interaction of yes and no trials, belief type and probe type, $F(1,23) = 14.124$, $p = .001$, $\eta p^2 = .380$. No other effects were significant (all *F*s < 1.230, all *p*s > .279).

To explore the three-way interaction, separate analyses were undertaken on yes and no trials. Firstly, a two-way ANOVA (belief type × probe type) on yes trials revealed a main effect of belief type, participants were faster at responding to true belief trials than false belief trials, $F(1,23) = 45.736$, $p < .001$, $\eta p^2 = .665$ and a main effect of probe type where participants were faster at responding to reality probes than belief probes, $F(1,23) = 23.070$, $p < .001$, $\eta p^2 = .501$. There was no significant interaction between belief type and probe type, $F(1,23) = 1.741$, $p = .200$, $\eta p^2 = .070$. Since our main interest was in whether a processing cost would be apparent for both false and true beliefs we finally conducted separate comparisons between belief and reality probes in the false belief condition, and between belief and reality probes in the true belief condition. These analyses showed that responses to belief probes were significantly slower than responses to their corresponding reality probes, whether the belief was false or true (both *t*s < 3.973, all *p*s < .002).

A similar two-way ANOVA on "no" trials yielded a main effect of belief type where participants were faster at responding to true belief trials than false belief trials, $F(1,23) = 49.493$, $p < .001$, $\eta p^2 = .683$ and no main effect of probe type, $F(1,23) = 1.113$, $p = .302$, $\eta p^2 = .046$ but there was a trend for faster response times to reality than belief probes. Moreover, there was a significant interaction between belief type and probe type, $F(1,23) = 11.776$, $p = .002$, $\eta p^2 = .339$. This interaction was investigated with *t*-tests, which showed no significant difference in response times to belief and reality probes when in a false belief context, $t(23) = .944$, $p > .05$ but faster responses to reality probes than belief probes when in a true belief context, $t(23) = 3.179$, $p = .004$.

With respect to accuracy scores a three-way ANOVA revealed a main effect of belief type, $F(1,23) = 25.929$, $p < .001$, $\eta p^2 = .530$, where participants were more accurate on true belief trials than false belief trials. No other effects were significant (all *F*s < 2.036, all *p*s > .167). Fig. 4 displays accuracy scores and it can be seen that the significant differences in response times reported in this experiment are not due to a trade off between speed and accuracy.

2.3. Discussion

In most cases, whether probes were presented at a time when the female character's belief was true or false, participants were slower to respond to questions about belief than reality. This is consistent with the hypothesis that participants were not automatically inferring beliefs. The exception to this pattern was on false belief trials that re-

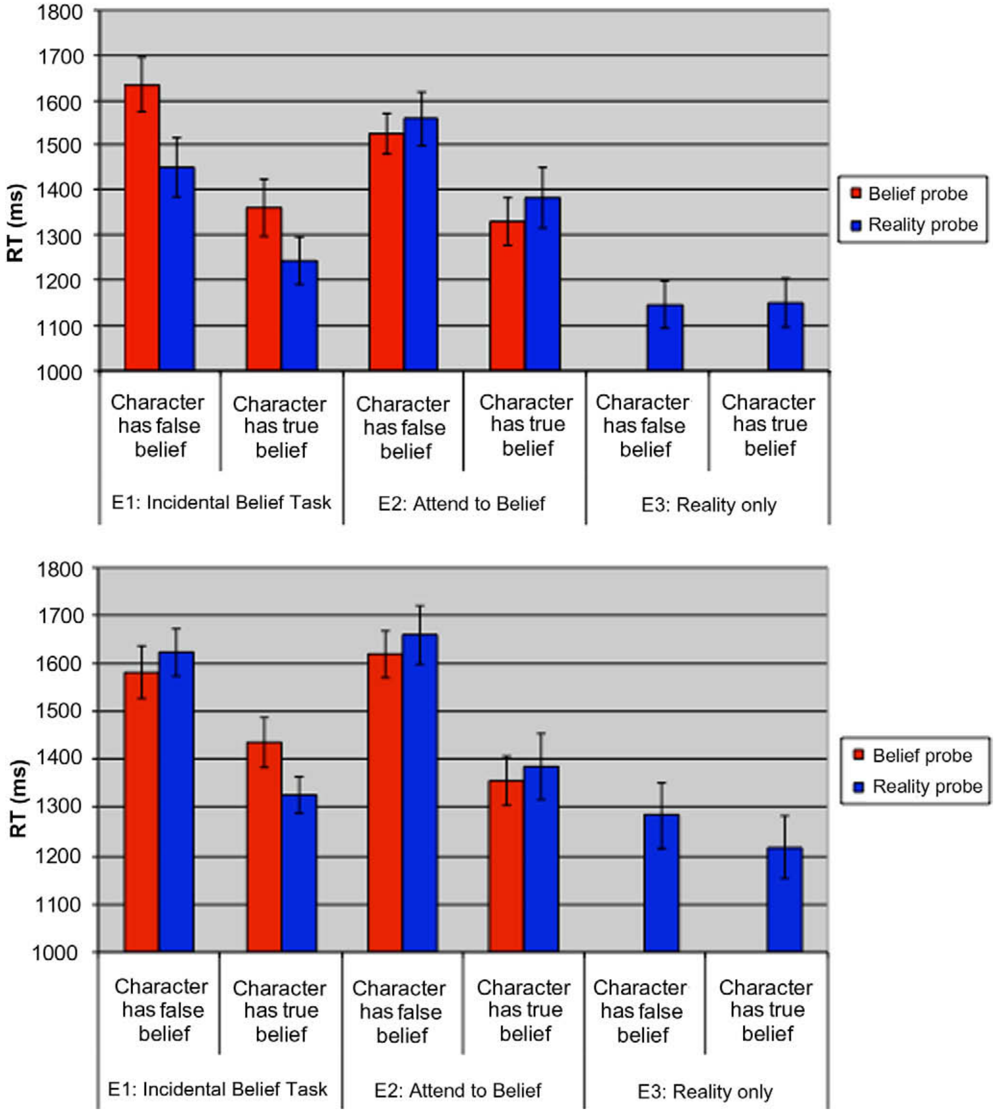

Fig. 3. Mean response times (error bars depict one standard error) for the experimental conditions of Experiments 1, 2, and 3. The upper panel charts trials that required a "yes" response and the lower panel charts trials that required a "no" response.

quired a "no" response to the probe. It is not clear why this condition differed from the pattern in the three other conditions. A second, and unexpected finding was that participants were generally faster at responding to probes when in a true belief rather than a false belief context. These two findings are investigated further in Experiments 2 and 3. Experiment 2 explores the response time difference to belief and reality probes to see whether this effect can be eliminated when participants are asked to monitor the female character's belief and Experiment 3 investigates the difference in response times to reality probes when in a true or false belief context.

3. Experiment 2 – attend to belief

This experiment investigates further the effect of probe type that was obtained in Experiment 1 where response times were slower to belief probes than reality probes. Our tentative interpretation of this finding is that participants were not always encoding beliefs and the observed processing cost for belief probes reflects the need to infer this information ad hoc in response to the probes. Following the rationale of Apperly et al. (2006), if this is true then when participants are asked to explicitly monitor and consequently encode beliefs, then this should reduce or eliminate the response time difference between belief and reality probes. Alternatively, if beliefs are inferred automatically, or if participants are always inferring beliefs spontaneously for the stimuli in Experiment 1 then we would need an alternative explanation for the observed difference in response times to belief and reality probes. It might be, for instance, that belief probes were intrinsically more difficult than reality probes, or formulating responses about beliefs was intrinsically more difficult than formulating responses about reality. If this explanation is

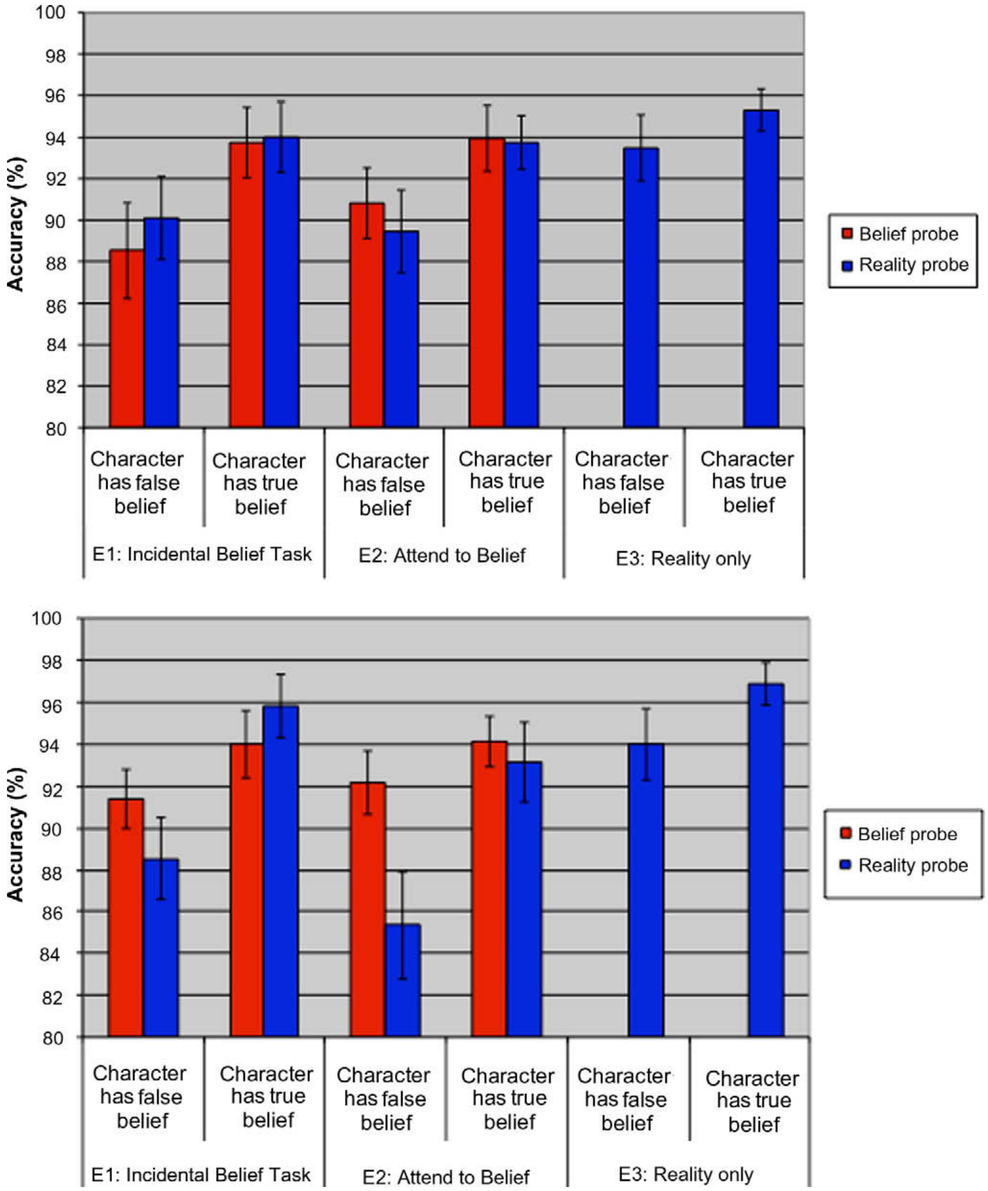

Fig. 4. Mean accuracy (error bars depict one standard error) for the experimental conditions of Experiments 1, 2, and 3. The upper panel charts trials that required a "yes" response and the lower panel charts trials that required a "no" response.

correct then we would not expect such difficulties to be eliminated by instructing participants to keep track of the female character's beliefs in Experiment 2.

3.1. Method

3.1.1. Participants

Thirty-two undergraduate and postgraduate students from the University of Birmingham participated in this study. None had participated in Experiment 1. Participants were aged between 18 and 23 with a mean age of 20 (12 males and 20 females). Participants were awarded with course credits or paid £10.

3.1.2. Stimuli

The same stimuli were presented as in Experiment 1.

3.1.3. Design

Experiment 2 followed the same design as Experiment 1.

3.1.4. Procedure

The procedure was the same as in Experiment 1 except that participants were explicitly asked to keep track of where the female character thinks the object is and participants were asked to locate whether the female character thinks the object is in the left or right box by clicking the left or right mouse button at the end of each sequence of slides. The only purpose of this part of the task was to encourage participants to keep track of information relevant for answering belief probes.

3.2. Results and discussion

Participants' response times in Experiment 2 (attend to belief) were analysed and Fig. 3 displays the mean response times. For correct responses, response times that were two standard deviations beyond the mean were removed. This criterion resulted in a loss of 172 out of 3753 data points.

A three-way ANOVA (yes and no trials × belief type × probe type) on response times revealed a main effect of belief type, $F(1,31) = 99.558$, $p < .001$, $\eta p^2 = .763$ where participants were faster at responding to true beliefs than false beliefs. Participants were also faster on belief probes than reality probes, $F(1,31) = 5.100$, $p = .031$, $\eta p^2 = .141$. Furthermore, participants were faster at responding to yes than no trials, $F(1,31) = 5.000$, $p = .033$, $\eta p^2 = .139$. There was also a non-significant trend for an interaction between yes and no trials and belief type, $F(1,31) = 3.473$, $p = .072$, $\eta p^2 = .101$. No other effects were significant (all Fs < .025, all ps > .875).

With respect to accuracy scores, a three-way ANOVA showed a main effect of belief type, $F(1,31) = 15.373$, $p < .001$, $\eta p^2 = .332$, participants were more accurate on true belief trials than false belief trials and participants were more accurate on belief than reality probes, $F(1,31) = 6.932$, $p = .013$, $\eta p^2 = .183$. There was a significant interaction between belief type and probe type, $F(1,31) = 4.938$, $p = .044$, $\eta p^2 = .124$). Post-hoc analyses revealed higher accuracy scores on belief than reality probes when in a false belief context, $t(31) = 2.732$, $p = .010$, whereas there were no differences in accuracy scores between belief and reality probes when in a true belief context, $t(31) = .682$, $p = .500$. Furthermore there was an interaction approaching significance between yes and no trials and probe type, $F(1,31) = 4.099$, $p = .052$, $\eta p^2 = .117$, where there was a trend for a larger difference in accuracy scores between belief and reality probes on no trials than on yes trials. No other effects were significant (all Fs < 2.632, all ps > .115). Figs. 3 and 4 display the relative mean response times and accuracy scores for this Experiment compared to Experiment 1.

Findings from Experiment 2, where participants were asked to keep track of where the female character thinks the object is located, reversed the pattern of response times to belief and reality probes observed in Experiment 1. From the results of Apperly et al. (2006) our hypothesis was that participants were slow to respond to belief probes in Experiment 1 because they had not inferred the female character's belief in advance of the probe, but this cost would be eliminated in Experiment 2 when participants were instructed to infer the female character's belief in advance of the probe. The cost for belief probes was indeed eliminated in Experiment 2, but our hypothesis cannot, on its own, account for the finding that responses to belief probes actually became faster than responses to reality probes.

An alternative hypothesis is that the different instructions in Experiments 1 and 2 led participants to attach different priorities to information about belief and reality. On this hypothesis, participants in Experiment 1 responded more quickly to reality probes than belief probes because the instruction to identify the object's location at the end of each trial led them to prioritise reality information, whereas participants in Experiment 2 responded more quickly to belief probes because they were explicitly instructed to keep track of the female character's beliefs. We might expect the extent of such prioritisation to depend on the force of the experimental instructions, and on this basis, we should expect the effect observed in Experiment 2 to be at least as large as the effect observed in Experiment 1, since the instructions to track belief in Experiment 2 were at least as forceful as the instructions to identify the object's real location in Experiment 1. To test this we conducted a direct comparison of the size of any prioritisation effect between Experiments 1 and 2.

3.3. Comparison of Experiments 1 and 2

We re-coded probe types according to whether the prioritisation hypothesis predicted that they probed for information that was prioritised by the experimental instructions. Thus, for Experiment 1, reality probes were coded as "prioritised" and belief probes were coded as "not-prioritised", whereas in Experiment 2, belief probes were coded as "prioritised" whereas reality probes were coded as "not-prioritised". We then conducted an analysis to test whether the degree of prioritisation varied between Experiments 1 and 2. (Thus, in effect we examined whether the differences between belief and reality probes were of the same magnitude in the two experiments, despite their different directions.)

For response time data we conducted an ANOVA with three within-subject factors, probe type (prioritised, non-prioritised), yes and no trials and belief type (true versus false) and one between-subject factor (Experiment). There were significant main effects of probe type, $F(1,54) = 16.824$, $p < .001$, $\eta p^2 = .238$, belief type, $F(1,54) = 182.819$, $p < .001$, $\eta p^2 = .772$, and no versus yes responses, $F(1,54) = 8.668$, $p = .005$, $\eta p^2 = .138$. There was a significant two-way interaction between no versus yes responses and probe type, $F(1,54) = 6.029$, $p = .017$, $\eta p^2 = .1$, two significant three-way interactions between no versus yes responses, probe type and Experiment, $F(1,54) = 4.982$, $p = .03$, $\eta p^2 = .084$, and no versus yes responses, probe type and belief type, $F(1,54) = 6.002$, $p = .018$, $\eta p^2 = .1$, and the four-way interaction between all factors was also significant, $F(1,54) = 5.766$, $p = .02$, $\eta p^2 = .096$.

We decomposed the four-way interaction into separate three-way ANOVAs for yes and no responses. The ANOVA for yes responses showed main effects of belief type, $F(1,54) = 107.284$, $p < .001$, $\eta p^2 = .665$, corresponding to faster responses on true belief trials than false belief trials. There was also a main effect of probe type, $F(1,54) = 19.515$, $p < .001$, $\eta p^2 = .265$, with faster responses when prioritised information was probed. Finally, there was a significant interaction between probe type and experiment, $F(1, 54) = 4.646$, $p = .036$, $\eta p^2 = .079$. No other effects approached significance. Paired *t*-tests revealed that in Experiment 1 participants were faster at responding to prioritised probes (i.e., reality probes) than non-prioritised probes (i.e., belief probes), $t(23) = 4.803$, $p < .001$, whereas in Experiment 2 the numerical trend for faster responses on prioritised probes (belief probes) than non-prioritised probes (reality probes) was not significant, $t(31) = 1.622$, $p = .115$.

We conducted a second three-way ANOVA (belief type × probe type × Experiment) on response time data from no responses. This analysis showed main effects of belief type, $F(1, 54) = 108.77$, $p < .001$, $\eta p^2 = .668$, corresponding to faster responses on true belief trials than false belief trials. There was also a main effect of probe type, $F(1, 54) = 4.312$, $p = .043$, $\eta p^2 = .074$, with faster responses when prioritised information was probed. There were significant interactions between probe type and belief type, $F(1, 54) = 5.073$, $p = .028$, $\eta p^2 = .086$, and between probe type, belief type and Experiment, $F(1, 54) = 4.288$, $p = .043$, $\eta p^2 = .074$.

To decompose this three-way interaction we conducted separate two-way ANOVAs for false belief and true belief trials. Firstly, a two-way ANOVA (Experiment × probe type) on data from false belief trials revealed no significant effects, all $Fs(1, 54) < 2.014$, all $ps > .162$, all $\eta p^2 < .036$. Secondly, a similar two-way ANOVA on data from true belief trials revealed a main effect of probe type, $F(1, 54) = 13.097$, $p = .001$, $\eta p^2 = .195$, with faster responses on prioritised probes than non-prioritised probes. No other effects were significant, all $Fs(1, 54) < 1.88$, all $ps > .175$, all $\eta p^2 < .034$.

For accuracy data we conducted a four-way ANOVA with three within-subject factors, probe type (prioritised, non-prioritised), yes and no trials and belief type (true versus false) and one between-subject factor (Experiment). There was a significant main effect of belief type, $F(1, 54) = 36.267$, $p < .001$, $\eta p^2 = .402$, and a trend for a significant effect of probe type, $F(1, 54) = 3.513$, $p < .066$, $\eta p^2 = .061$. There was a trend for a significant three-way interaction between no versus yes responses, probe type and Experiment, $F(1, 54) = 3.613$, $p = .063$, $\eta p^2 = .063$, and the four-way interaction between all factors was also significant, $F(1, 54) = 4.689$, $p = .035$, $\eta p^2 = .080$.

We decomposed the four-way interaction into separate three-way ANOVAs for yes and no responses. The three-way ANOVA for yes responses showed main effects of belief type, $F(1, 54) = 19.002$, $p < .001$, $\eta p^2 = .260$, corresponding to more accurate responses on true belief trials than false belief trials. No other effects approached significance.

The three-way ANOVA on no responses showed main effects of belief type, $F(1, 54) = 29.117$, $p < .001$, $\eta p^2 = .350$, and probe type, $F(1, 54) = 5.194$, $p = .027$, $\eta p^2 = .088$, and significant interactions between probe type and Experiment, $F(1, 54) = 8.882$, $p = .004$, $\eta p^2 = .141$, and between probe type, belief type and Experiment, $F(1, 54) = 8.099$, $p = .006$, $\eta p^2 = .13$.

To decompose this three-way interaction we conducted separate two-way ANOVAs for false belief and true belief trials. Firstly, a two-way ANOVA (Experiment × probe type) on data from true belief trials revealed no significant effects, all $Fs(1, 54) < 1.83$, all $ps > .182$, all $\eta p^2 < .033$. Secondly, a similar two-way ANOVA on data from false belief trials revealed a significant interaction between Experiment and probe type, $F(1, 54) = 13.446$, $p = .001$, $\eta p^2 = .199$. No other effects were significant, all $Fs(1, 54) < 2.254$, all $ps > .139$, all $\eta p^2 < .04$. Paired *t*-tests revealed that in Experiment 2 participants were more accurate at responding to belief probes (prioritised) than reality probes (non-prioritised), $t(31) = 3.933$, $p < .001$. In Experiment 1 the difference was non-significant, $t(23) = 1.44$, $p = .161$, but the numerical trend was actually in an anomalous direction, with a tendency for more accurate responses on belief probes (non-prioritised) than reality probes (prioritised).

In summary, the omnibus analysis of response times showed that participants responded more quickly to probes about prioritised information than probes about non-prioritised information, and this was consistent with a non-significant trend for more accurate responses to probes about prioritised information. These findings lend support to the general hypothesis that participants varied how they prioritised information according to the experimental instructions. However, this effect was moderated by interactions between probe type and experiment, suggesting that the prioritisation hypothesis may be insufficient to explain the pattern of findings completely.

In error data the effect of probe type varied between Experiments 1 and 2 in one cell of the design. For no responses in the false belief condition participants in Experiment 2 responded more accurately to prioritised probes than non-prioritised probes, whereas there was a tendency for the opposite pattern in Experiment 1. This pattern is not easily explained in terms of either of our hypotheses: neither differences in prioritisation according to experimental instructions nor differences in encoding of information about belief and reality across the two experiments would have predicted this pattern. However, we are reluctant to propose further hypotheses on the basis of this effect at this point because the analyses of Experiment 1 presented earlier already demonstrated that the overall pattern observed in this cell of the design in Experiment 1 was clearly anomalous in comparison with the other three cells (i.e., yes responses in the false belief condition, and both yes and no responses in the true belief condition).

In contrast, the interaction between probe type and experiment in response time data can be interpreted with more confidence. For trials on which participants gave yes responses (50% of the total data), the effect of probe type was significantly greater in Experiment 1 than in Experiment 2, and this effect was not off-set by the pattern of errors for yes responses, or contradicted by any effects in response time data from no trials (the other 50% of total data). From the point of view of the prioritisation hypothesis it is not clear why this interaction should occur. As mentioned above, the instruction to track the female character's belief in Experiment 2 is at least as forceful as the instruction to identify the object's real location in Experiment 1, leading to the expectation that prioritisation should be at least as large in Experiment 2 as in Experiment 1. We suggest that the observed pattern may instead be explained by our original hypothesis: that is to say, in addition to an effect of prioritisation, participants incurred processing costs for belief probes in Experiment 1 because they had not automatically inferred the female character's belief in advance of the probe.

An additional and unexpected finding from Experiments 1 and 2 was that participants were consistently faster at responding in true belief conditions than in false belief conditions. It is critical to emphasise here that participants were not only faster at responding to probes about true beliefs than probes about false beliefs, but also were faster at responding to reality probes presented in a true belief scenario than identical reality probes presented in a false belief scenario. Because the probes were identical in true belief and false belief conditions this difference cannot be accounted for by variation in the complexity of the probes. True belief and false belief conditions consisted of the very same set of events (e.g., box swapping, pointing, female character leaving and re-entering room), and only minor variation in their combination (whether or not the female character was in the room when the boxes were swapped). It seems unlikely that the large differences in response time and accuracy between false belief and true belief conditions could be accounted for by these minor variations in event sequence. Instead we believe these results may be due to participants suffering greater interference between information about belief and reality in the false belief condition than in the true belief condition.

The possibility of interference when holding in mind information about belief and reality is illustrated by Apperly et al. (2008). In this study adults read short sentences describing reality and a character's belief, and then verified whether a picture probe correctly corresponded to this information. Processing costs (reaction times and/or error rates) were higher when the information in the sentences described a false belief than when belief and reality were unrelated, and these processing costs were apparent whether participants were probed for information about belief or reality. These findings provide a ready explanation for the results of Experiment 2 in the current study, where participants were instructed to track the female character's belief and in doing so presumably also kept track of reality. If participants held information about belief and reality in mind then the findings of Apperly et al. (2008) would clearly predict greater processing costs for probes in the false belief condition than in the true belief condition, whether the probes concerned belief or reality.

However, care is needed in applying this explanation to the results of Experiment 1. One possible interpretation of the results of Experiment 1 and of Apperly et al. (2006) is that in these "incidental belief reasoning tasks" participants never encode information about the character's beliefs and so always do so ad hoc in response to test probes. If this were correct then there would be no information about belief to generate interference with information about reality, and so participants' responses to reality probes should not differ in true belief and false belief conditions. However, there are at least two alternative ways in which the observed interference might be explained.

One possibility is that our interpretation of Experiment 1 is wholly incorrect: beliefs are in fact inferred automatically, and so generate interference with information about reality, particularly when beliefs are false. Of course, this would leave the observed differences in reaction time between belief and reality probes in need of an explanation, though perhaps, as Cohen and German (2009) suggest, this could be due to the need to retain information about belief from the time of encoding to the time when this information is probed. Another possibility is that beliefs are *not* inferred automatically but may be inferred spontaneously, without explicit instruction. We suppose that many belief inferences in everyday life will in fact be spontaneous. The phenomenon of spontaneous (rather than automatic) inferences is familiar from work on discourse processing, where readers frequently make inferences to form a coherent interpretation of the text (see e.g., McKoon & Ratcliff, 1998; Sanford & Garrod, 1998; Zwaan & Radvansky, 1998), and indeed there is evidence that many other social inferences are spontaneous (e.g., Uleman, Saribay, & Gonzalez, 2008). The probability that readers will make a spontaneous inference is determined by a variety of contextual and motivational factors (e.g., McKoon & Ratcliff, 1998), and it is this that distinguishes spontaneous inferences from inferences that are made automatically whenever participants attend to a stimulus that affords the inference. Factors that might have led participants in Experiment 1 (and potentially in Apperly et al., 2006, Experiment 1) to make spontaneous belief inferences are the frequency with which information about beliefs was probed and the fact that keeping track of the female character's beliefs was actually necessary in some filler trials in order to interpret her signal and infer the location of the object. Could it be that in Experiment 1 participants were not automatically inferring beliefs (resulting in slower responses to belief probes than to reality probes), but were spontaneously inferring beliefs on at least some trials (resulting in greater interference between belief and reality information in false belief conditions than in true belief conditions)?

In Experiment 3 we sought to distinguish between these possibilities by seeking to reduce even further any cues that might prompt participants to make spontaneous inferences about the female character's beliefs. Thus, in Experiment 3 participants were not instructed to track the female character's beliefs, probe questions never concerned the female character's beliefs and filler trials were eliminated if they required participants to track the female character's beliefs in order to locate the object. Aside from these changes, participants viewed exactly the same stimuli as in Experiments 1 and 2 that clearly afforded belief inferences about the depicted characters. If participants track beliefs automatically then we should still observe a difference in response times on reality probes between true and false belief contexts because participants will still be suffering interference from belief. If they are not tracking beliefs automatically and if we are successful in reducing the likelihood of spontaneous belief inferences, then participants should suffer less overall interference between belief and reality when responding to reality probes and the difference in response times to true and false belief contexts should be reduced or eliminated.

4. Experiment 3 – reality only

4.1. Method

4.1.1. Participants

Twenty-four undergraduate and postgraduate students from the University of Birmingham participated in this study. None had participated in the previous experiments. Participants were aged between 18 and 31 with a mean age of 22. There were 11 males and 13 females. Participants were awarded with course credits or paid £10.

4.1.2. Stimuli

The same stimuli were used as in Experiments 1 and 2 except some of the picture probes differed. No picture probes were presented that would lead to participants thinking about where the female character would think the object is (i.e., 64 trials of she thinks left and right were replaced with 32 trials of whether the object/boxes were swapped and 32 trials whether the objects/boxes were not swapped). Filler trials were also excluded that may instigate a participant thinking about the female character's perspective (i.e., late indication trials) and replaced with 16 other existing fillers.

4.1.3. Design

The design was the same as in the previous experiments.

4.1.4. Procedure

The procedure was the same as in Experiment 1 as the task was to locate which box the object was in. However the picture probes differed in that belief probes were not presented in the explanation to participants.

4.2. Results

Here and throughout the following analyses, note that only data from reality probes are being analysed. Participants' response times in Experiment 3 (reality only) were analysed and Fig. 3 displays the mean response times. For correct responses, response times that were two standard deviations beyond the mean were removed. This criterion resulted in a loss of 139 out of 2934 data points. In order to investigate whether response times to reality probes differ according to whether a true or false belief context is presented, a two-way ANOVA (yes and no trials × belief type) was carried out. Importantly, there was *no* main effect of belief type, $F(1,23) = 2.168$, $p = .154$, $\eta p^2 = .086$, therefore participants were just as fast at responding to reality probes when presented in a true or false belief context. There was a main effect of yes and no trials, $F(1,23) = 11.114$, $p = .003$, $\eta p^2 = .326$, where participants were faster on yes trials than no trials. There was a non-significant trend for an interaction between yes and no trials and belief type, $F(1,23) = 3.169$, $p = .088$, $\eta p^2 = .121$, which showed a trend for faster responses in true belief contexts than false belief contexts on no trials but not on yes trials.

With respect to accuracy scores, a two-way ANOVA (yes and no trials × belief type) revealed a main effect of belief type, $F(1,23) = 4.404$, $p = .047$, $\eta p^2 = .161$, participants were marginally more accurate at true belief trials (96.1) than false belief trials (93.8). No other effects were significant (all *F*s < .434, all *p*s > .517).

4.3. Comparing Experiments 1 and 3

Similar analyses were undertaken to compare response times and accuracy for reality probes between Experiments 1 (belief and reality probes presented) and 3 (only reality probes presented). A three-way ANOVA (yes and no trials × belief type × Experiment) on response times revealed main effects of yes and no trials, $F(1,46) = 20.233$, $p < .001$, $\eta p^2 = .305$, responses were faster on yes trials than no trials, belief type, $F(1,46) = 69.084$, $p < .001$, $\eta p^2 = .600$, responses were faster on true belief than false belief trials, and there was an effect of Experiment, $F(1,46) = 9.946$, $p = .003$, $\eta p^2 = .178$, responses were faster in Experiment 3 than Experiment 1. There were two 2-way interactions, the first was between belief type and Experiment, $F(1,46) = 43.888$, $p < .001$, $\eta p^2 = .488$, in Experiment 1 participants were faster at true belief trials than false belief trials, $t(23) = 9.862$, $p < .001$ but in Experiment 3 there was only a non-significant trend for faster responses on true belief trials, $t(23) = 2.013$, $p = .056$. The second interaction was between yes and no trials and belief type, $F(1,46) = 7.586$, $p = .008$, $\eta p^2 = .142$, on no trials responses were faster to true belief trials than false belief trials, $t(47) = 6.088$, $p < .001$ and this was also the case for yes trials, $t(47) = 4.101$, $p < .001$ (but the difference between means were not as large). None of the other effects were significant (all *F*s < .470, all *p*s > .496).

With respect to accuracy, a three-way ANOVA revealed a main effect of belief type where participants were more accurate on true belief trials, $F(1,46) = 16.628$, $p < .001$, $\eta p^2 = .266$ and there was a non-significant trend for better accuracy on Experiment 3 than Experiment 1, $F(1,46) = 3.273$, $p = .077$, $\eta p^2 = .066$. Although there was a trend for the difference between true and false belief trials to be smaller in Experiment 3 than in Experiment 1, this interaction was not significant, $F(1,46) = 2.793$, $p = .101$, $\eta p^2 = .057$. None of the other effects were significant (all *F*s < 1.906, all *p*s > .174). Figs. 3 and 4 display the mean response times and accuracy rates for Experiment 3 compared to Experiments 1 and 2.

4.4. Comparing Experiments 2 and 3

Further analyses were undertaken to compare response times and accuracy for reality probes between Experiments 2 and 3. A three-way ANOVA (yes and no trials × belief × Experiment) revealed that responses were faster on yes than no trials, $F(1,54) = 13.760$, $p < .001$, $\eta p^2 = .203$, responses were faster on true than false belief trials, $F(1,54) = 36.675$, $p < .001$, $\eta p^2 = .404$ and response times were faster in Experiment 3 than Experiment 2, $F(1,54) = 18.001$, $p < .001$, $\eta p^2 = .250$. In addition to these effects, there was a significant interaction between belief type and Experiment, $F(1,54) = 21.175$, $p < .001$,

$\eta p^2 = .282$, where participants in Experiment 2 were faster at responding on true belief trials than on false belief trials $t(31) = 10.264$, $p < .001$, whereas in Experiment 3, there was only a trend for faster responses on true belief trials, $t(23) = 2.013$, $p = .056$. Furthermore, there was a non-significant trend for an interaction between yes and no trials and belief type, $F(1, 54) = 3.671$, $p = .061$, $\eta p^2 = .064$. Responses were faster on true belief trials than false belief trials on both "no" trials $t(55) = 5.454$, $p < .001$ and "yes" trials, $t(55) = 3.920$, $p < .001$. None of the other effects were significant (all Fs < 1.126, all ps > .293).

With respect to accuracy scores, a three-way ANOVA (yes and no trials × belief × Experiment) revealed higher accuracy scores on true belief trials than false belief trials, $F(1, 54) = 14.761$, $p = .001$, $\eta p^2 = .215$ and there were higher accuracy scores in Experiment 3 than Experiment 2, $F(1, 54) = 9.136$, $p = .004$, $\eta p^2 = .145$. None of the other effects were significant (all Fs < 3.068, all ps > .086).

4.5. Discussion

Removing participants' motivation to track the female character's belief about the object location in Experiment 3 had two effects. First, in both false belief and true belief conditions, participants responded more quickly and at least as accurately to reality probes than they did in Experiments 1 and 2. It is possible that this was due to participants' task in Experiment 3 being generally simpler than it was in Experiments 1 and 2 because they were never asked about where the female character thought the object was located. However, we do not favour this interpretation because in Experiment 3 participants were still probed for a variety of information about the man, the female character and reality, and the frequency of the critical reality probes, concerning the location of the object, was identical in all three experiments. Instead, we suggest that participants responded to reality probes more quickly in Experiment 3 because they were much less likely to have inferred where the female character thought the object was located, so had less potentially confusable information in mind when they responded to reality probes. Importantly, this effect held in both true belief and false belief conditions suggesting that neither true nor false belief information was available to generate a processing cost when selecting responses to the reality probes.

The second effect of removing participants' motivation to track the female character's belief about the object's location was that the large difference in response times to reality probes in false belief and true belief conditions observed in Experiments 1 and 2 was significantly reduced. This finding can be understood if we suppose that participants in Experiment 3 were much less likely to infer the female character's belief. Without this information in mind participants did not suffer greater interference in their responses to reality probes when the female character's belief happened to be false than when it happened to be true.

However, although the difference in response patterns to reality probes in true belief and false belief contexts was reduced in Experiment 3 it was not eliminated entirely. In reaction times there remained a non-significant trend for faster responses to true belief probes, and true belief probes were answered more accurately to a small (2.3%) but significant degree. It is not clear why these differences remain. It is possible that although the event sequences for false belief and true belief conditions were very similar, the small differences that existed made true belief conditions marginally easier to process. Another possibility is that Experiment 3 reduced the frequency with which participants spontaneously inferred the female character's belief about the object's location but participants continued to make such inferences on some trials, and therefore continued to suffer some interference from this information. A final possibility that cannot be entirely ruled out is that beliefs are, in fact, inferred automatically and so are still inferred in Experiment 3. It could be that because this information was never probed it was less salient and so interfered less when participants responded to reality probes. The current evidence does not allow us to distinguish definitively between these explanations for the small residual difference between true belief and false belief trials in Experiment 3.

5. General discussion

5.1. Are belief inferences automatic?

The current paper provides two sources of evidence bearing on the automaticity of belief inferences. First, Apperly et al. (2006) found evidence that human adults do not automatically infer false beliefs when faced with a stimulus that affords a belief inference. In Apperly et al.'s study, when participants had no particular reason to be monitoring beliefs they were slower to respond to probe questions about where a character in the video stimuli thought an object was located than to probe questions about the object's real location. This difference was eliminated in other conditions where participants were explicitly instructed to track the character's beliefs. The first aim of the current paper was to test whether the same effects would be observed for both false beliefs and true beliefs.

In Experiment 1, where participants had no particular reason to be monitoring beliefs, responses were indeed usually faster when making judgements about reality than belief, whether the beliefs were false or true. This was consistently the case in true belief conditions. In false belief conditions participants responded more quickly to belief probes than to reality probes when the correct answer was "yes" (consistent with Apperly et al., 2006) but not when the correct answer was "no" (the design of Apperly et al., 2006 study meant that "no" responses could not be interpreted). The opposite pattern was observed in Experiment 2 where participants were explicitly told to track beliefs. This pattern is clearly consistent with the idea that participants were not automatically tracking either true or false beliefs in Experiment 1, but could track beliefs for the same stimuli when explicitly instructed to do so. However, as pointed out in the discussion of Experiment 2, this hypothesis alone cannot explain why responses to belief probes were actually faster than responses to reality probes in Experiment 2.

One alternative hypothesis is that the differing instructions of Experiments 1 and 2 led participants to prioritise information about belief and reality differently. Could it be that participants were encoding information about both reality *and* belief in both experiments but prioritised reality information in Experiment 1 (because of the requirement to locate the object at the end of each trial) and belief information in Experiment 2 (because of the instruction to track beliefs)? The current data, and the findings from Apperly et al. (2006) give some reasons for doubting a strong version of this hypothesis, which assumes that automatic processing guarantees that belief and reality are always tracked. In this case, the effects observed in Experiments 1 and 2 were exclusively due to different task instructions leading participants to attach opposite priority to information about belief and reality, and so we might have expected to observe equivalent and opposite differences in response times to belief and reality probes in the two experiments. In fact, although responses to reality probes were clearly faster than responses to belief probes in most comparisons of Experiment 1, the opposite pattern in Experiment 2, though significant, was weaker. Moreover, in a condition equivalent to Experiment 2, Apperly et al. (2006) actually found no difference in response times to belief and reality probes, whereas Apperly et al. did find the effect of slower responses to belief probes observed in Experiment 1. This is consistent with the possibility that the difference between belief and reality probes observed in Experiment 2 of the current paper is less reliable, or reflects a smaller effect, than the opposite effect observed in Experiment 1. There is no reason to expect this to be the case if the only difference between the experiments was the priority participants gave to belief and reality information that they had automatically inferred and encoded.

However, the current findings are entirely consistent with a weaker version of the strategic priority hypothesis, whereby information about belief and reality may be given different priorities if this information has actually been inferred. On this account, faster responses to belief probes in Experiment 2 were indeed due to participants prioritising beliefs over reality, and the opposite prioritisation may indeed have contributed to the opposite effect observed in Experiment 1 because, as described below, there are reasons for thinking that participants sometimes made spontaneous belief inferences in Experiment 1. But because this hypothesis does not assume that information about belief is automatically inferred, it leaves the way open for a further factor to help explain why the effect of probe type observed in Experiment 1 was larger than the effect observed in Experiment 2, and why Apperly et al. (2006) observed no effect of probe type in a condition equivalent to Experiment 2. We suggest that this further factor was that participants in Experiment 1 had not always inferred information about the woman's belief in advance of the probe, because neither true nor false belief inferences are automatic.

The second source of evidence on the automaticity of belief inferences came from the patterns of interference observed between information about belief and reality. Experiments 1 and 2 provided consistent evidence suggesting that both belief and reality probes were easier to process when the target character had a true belief than when she had a false belief. We suggest that this effect is primarily due to participants suffering interference when they hold in mind information about belief and reality, with more interference when the belief is false than when it is not (Apperly et al., 2008). This interpretation is compatible with the idea that beliefs are not inferred automatically because in Experiment 2 participants were explicitly told to track where the female character thought the object was located. Likewise in Experiment 1 there were multiple cues (frequent belief probes, the necessity to track the female character's belief on some filler trials) that might have led participants spontaneously to track where the female character thought the object was located on at least some trials. Importantly, this interpretation is inconsistent with Cohen and German's (2009) suggestion that participants do not retain information about belief without overt instruction: if our interpretation is correct then participants in Experiment 1 sometimes spontaneously inferred the female character's beliefs (without overt instruction to do so) and retained this information until the probe point later in the sequence, at which point it caused interference with judgements about reality probes.

Our interpretation is further supported by the results of Experiment 3, where we took additional steps to eliminate cues that might lead participants to track the female character's belief spontaneously. In this case, response times to reality probes were significantly faster, the difference in response times to reality probes in true belief and false belief conditions was much reduced, and this was not at the expense of changes in the accuracy of responding to these probes. We suggest that this is because participants no longer suffered more interference from the female character's beliefs in the false belief condition than in the true belief condition. Importantly, we suggest that this is because participants did not infer the female character's beliefs in Experiment 3.

In sum our view is that the best account of the current findings, and those in the existing literature, entails the conclusion that participants do not infer beliefs automatically, and that this effect operates over and above any effects of experimental instructions on the priority that participants give to different information. However, people clearly can infer beliefs in the course of a task when told to do so. And, more importantly, people undoubtedly do infer beliefs spontaneously – without explicit instruction – in many circumstances, and the likelihood that they do so will be determined by the relevance of a belief inference for making sense of the stimulus that is currently being processed, and the availability of resources for making the inference. Nonetheless, as already indicated, we cannot entirely rule out the possibility that beliefs are inferred automatically. Of course, it would be wrong to mistake the current absence of evidence for evidence of absence. However, we do suggest that the current studies help shift the burden of proof further onto those who believe that belief inferences are automatic. Since there are no published experimental findings that give positive evidence for automatic belief inferences, we favour the conclusion that neither true nor false belief inferences are automatic.

5.2. Are true beliefs special?

Yes, and no. On the one hand, the current study provided clear evidence that true beliefs are easier to ascribe than false beliefs, with consistently faster and more accurate responses to true belief probes compared with false belief probes in both Experiments 1 and 2. On the other hand, participants treated true belief probes differently from reality probes. In Experiment 1 there remained a processing cost for responding to true belief probes compared with matched reality probes, and indeed, this cost was no different in the true belief condition than in the false belief condition. In both true and false belief conditions, judging the female character's belief carried a processing cost. What varied was the overall difficulty of the conditions, which affected judgements about both belief and reality. We propose that false belief conditions were more difficult because of the conflict between belief and reality (consistent with Apperly et al., 2008), which affected *both* belief and reality probes. It was the absence of this conflict that made true beliefs easier to ascribe, not the default ascription of true belief, or a strategy of judging reality in place of belief. Consistent with this conclusion that true beliefs are not ascribed by default, when participants were no longer probed for information about the female character's belief (Experiment 3) participants' responses to reality probes were faster than in Experiments 1 or 2 whether they were presented in a false belief scenario or a true belief scenario. That is to say, Experiment 3 not only minimised the additional cost of responding to reality probes in a false belief context compared with a true belief context, it also resulted in an overall decrease in processing cost for reality probes, whether they were presented in a false or a true belief context. We suggest that this is because participants had not inferred either false or true beliefs in Experiment 3, and that the absence of this information made it easier for participants to respond to reality probes in either context.

What, then, are we to make of the frequent claims in the literature that true beliefs are a default ascription, resulting in judgement biases in both children and adults (Birch & Bloom, 2007; Epley et al., 2004; Keysar et al., 2003; Mitchell et al., 1996; Wellman et al., 2001)? First, there need be no disagreement over the empirical phenomena. Even if true beliefs are not ascribed by default, Experiments 1 and 2 did clearly show that it is easier to answer questions about a situation involving someone's true belief about reality than to answer questions about a situation involving someone's false belief about reality. Thus, at least some of the judgment biases that have been attributed to default true belief ascription may be due instead to participants finding it generally easier to process situations involving true beliefs.

Second, we see no reason why true belief ascription is actually required to explain many of the defaults and biases observed in children and adults. Both Fodor (1992) and Leslie et al. (2004) have suggested that it would make sense for there to be a default assumption that beliefs are true because beliefs are supposed to be true, so a system with limited processing resources (such as a young child, or an adult who is distracted or under time pressure) might do well to work on this assumption. Leslie and colleagues (1992) go further in proposing that this default actually involves ascribing a true belief. However, true belief ascription as such does not seem necessary for an agent to reap the rewards of happening to share the same perspective as the target. Put another way, the agent's default could be to ignore belief entirely and simply make predictions about the target on the basis of the agent's own beliefs and knowledge. This possibility is explicitly suggested by some developmental psychologists (e.g., Wellman, 1991) and is clearly consistent with the suggestion that adults' judgements are egocentric by default, and only accommodate to the perspective of others under effortful cognitive control (e.g., Epley et al., 2004). Thus, the possibility of a tendency to default to one's own point of view would give the cognitive economy sought by theorists who propose default true belief ascription, without entailing that true beliefs are necessarily ascribed. Of course, this ''no-belief'' default would mean that actually judging what someone else thinks (as opposed to ignoring their perspective entirely) comes with a processing cost. This is what Experiment 1 of the current study suggests.

6. Conclusion

In the introduction we noted that developmental research leads to conflicting perspectives on whether belief reasoning abilities should be characterised as a ''low-level'', potentially modular process or a ''high-level'' central process. Our current findings suggest that adults' belief reasoning is not automatic, but is sensitive to the relevance of belief reasoning for the on-going task and to explicit instructions to infer beliefs. These findings are clearly more consistent with the view that belief reasoning – as assessed in the current experiments – is a ''high-level'' central process, not an automatic modular process. This fits with evidence that the ability to reason about beliefs takes children several years to acquire (e.g., Apperly & Robinson, 2003; Carpendale & Chandler, 1996; Flavell, 1999; Wellman et al., 2001), that it is related to developments in executive function (e.g., Hughes, 1998; Perner, 1998; Sabbagh, 2006; Zelazo, Jacques, Burack, & Frye, 2002), and that the developmental dependency between belief reasoning and executive function is due, at least in part, to the fact that executive function makes critical contributions to adult-like abilities that children are developing (Apperly, Samson, & Humphreys, 2009).

However, this raises an interesting question about the role of belief reasoning in cognition. To the degree that belief reasoning is a central process that makes relatively high demands on scarce cognitive resources for working memory and executive control it may not be well-suited to the very rapid guidance of online social interaction and communication, or to explaining the appearance of theory of mind abilities in infants (e.g., Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007) or non-human animals (e.g., Emery & Clayton, 2009; Tomasello et al., 2003) who have limited working memory and executive control. In part, it is recognition of this problem that motivates theorists such as Leslie et al. (2004) and Sperber and Wilson (2002) to argue for a fast, automatic theory of mind module, which would be capable of keeping up with rapidly evolving online processes. In this sense we agree with these authors' analysis of the problem: there is a clear need to explain how social interaction is guided rapidly and efficiently in a way that is sensitive to the beliefs, desires and intentions of the participants in the interaction. But the current data add to the reasons for doubting that this guidance is always achieved via the inference of beliefs, desires and intentions as normally understood, and as normally assessed in tests such as the false belief task.

This problem has recently been discussed by Apperly and Butterfill (2009), who argue that human adults have two types of cognitive system for theory of mind that make complementary trade-offs between flexibility and cognitive efficiency. On this two-systems account, adults share with infants (and perhaps some non-human species) a capacity for ascribing belief-like states. This capacity makes low demands on working memory and executive function, and may be relatively automatic in operation, but its efficiency comes at the cost of limitations on the cues to which the system is responsive and complexity of the belief-like states that can be ascribed. These limitations are only overcome when children develop the ability to reason about beliefs *as such*. This ability is flexible, but is not automatic and makes substantial demands on working memory and executive function, even in adults. The current findings clearly fit with the characteristics of the latter system for reasoning about beliefs as such, and we speculate that this system is recruited in the current study by the requirement for participants to make explicit judgements about beliefs.

Acknowledgement

This work was supported by a grant from the ESRC: RES-000-23-1419.

References

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adult performance on a non-inferential theory of mind task. *Cognition, 106*, 1093–1108.

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

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C., & Samson, D. (2006). Is belief reasoning automatic? *Psychological Science, 17*, 841–844.

Apperly, I. A., & Robinson, E. J. (2003). When can children handle referential opacity? Evidence for a systematic variation in 5- and 6-year-old children's reasoning about belief and belief reports. *Journal of Experimental Child Psychology, 85*, 297–311.

Apperly, I. A., Samson, D., & Humphreys, G. W. (2009). Studies of adults can inform account of theory of mind development. *Developmental Psychology, 45*(1), 190–201.

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case specific developmental delay. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 30*, 285–297.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a ''theory of mind''? *Cognition, 21*, 37–46.

Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D. J. (2000). *Understanding other minds: Perspectives from developmental cognitive neuroscience* (2nd ed.). New York: Oxford University Press.

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*, 382–386.

Carpendale, J. I. M., & Chandler, M. J. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Development, 67*, 1686–1706.

Carruthers, P., & Smith, P. K. (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.

Cohen, A. S., & German, T. C. (2009). Encoding of others' beliefs without overt instruction. *Cognition, 111*, 356–363.

Emery, N. J., & Clayton, N. S. (2009). Comparative social cognition. *Annual Review of Psychology, 60*, 87–113.

Epley, N., Morewedge, C., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology, 40*, 760–768.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology, 50*, 21–45.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1992). A theory of the child's theory of mind. *Cognition, 44*, 283–296.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. (1992). Why the child's theory of mind is really a theory. *Mind & Language, 7*, 145–171.

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16*, 233–253.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32–38.

Keysar, B., Lin, S. H., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 98*, 25–41.

Kikuno, H., Mitchell, P., & Ziegler, F. (2007). How do young children process beliefs about beliefs? Evidence from response latency. *Mind & Language, 22*, 297–316.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in ''theory of mind''. *Trends in Cognitive Sciences, 8*, 528–533.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development neuropsychological evidence from autism. *Cognition, 43*, 225–251.

McKoon, G., & Ratcliff, R. (1998). Memory-based language processing: Psycholinguistic research in the 1990s. *Annual Review of Psychology, 49*, 25–42.

Mitchell, P., & Riggs, K. J. (Eds.). (2000). *Children's reasoning and the mind*. Hove: Psychology Press.

Mitchell, P., Robinson, E. J., Issacs, J. E., & Nye, R. M. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition, 59*, 1–21.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*, 255–258.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: Bradford Books/MIT-Press.

Perner, J. (1998). The meta-intentional nature of executive functions and theory of mind. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes*. Cambridge: Cambridge University Press.

Sabbagh, M. (2006). Executive functioning and preschoolers' understanding of false beliefs, false photographs, and false signs. *Child Development, 77*(4), 1034–1049.

Sanford, A. J., & Garrod, S. C. (1998). The role of scenario mapping in text comprehension. *Discourse Processes, 26*, 159–190.

Saxe, R., Schultz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience, 1*(3–4), 284–298.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by two-year-olds. *Psychological Science, 18*, 587–592.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: An interdisciplinary perspective*. New York: Oxford University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language, 17*, 3–23.

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states – The question is which ones and to what extent. *Trends in Cognitive Sciences, 7*, 153–156.

Uleman, J. S., Saribay, A. S., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impression and implicit theories. *Annual Review of Psychology, 59*, 329–360.

Wellman, H. M. (1991). From desires to beliefs: Acquisition of a theory of mind. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading* (pp. 19–38). Oxford: Basil Blackwell.

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

Wertz, A., & German, T. (2007). Belief-desire reasoning in the explanation of behaviour: Do actions speak louder than words? *Cognition, 105*, 184–194.

Zelazo, P., Jacques, S., Burack, J. A., & Frye, D. (2002). The relation between theory of mind and rule use: Evidence from persons with autism-spectrum disorders. *Infant and Child Development, 11*, 171–195.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.