Developmental Science

Developmental Science 12:4 (2009), pp 521–535 DOI: 10.1111/j.1467-7687.2008.00793.x

PAPER

A competitive nonverbal false belief task for children and apes

Carla Krachun,[1,2] Malinda Carpenter,[1] Josep Call[1] and Michael Tomasello[1]

1. Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
2. Institute of Cognitive Science, Carleton University, Ottawa, Canada

Abstract

*A nonverbal false belief task was administered to children (mean age 5 years) and two great ape species: chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*). Because apes typically perform poorly in cooperative contexts, our task was competitive. Two versions were run: in both, a human competitor witnessed an experimenter hide a reward in one of two containers. When the competitor then left the room (version A) or turned around (version B), the experimenter switched the locations of the containers. The competitor returned and reached with effort, but unsuccessfully, towards the incorrect container. Children displayed an understanding of the competitor's false belief by correctly choosing the* other *container to find the reward. Apes did not. However, in version A (but not version B), apes looked more often at the unchosen container in false belief trials than in true belief control trials, possibly indicating some implicit or uncertain understanding that needs to be investigated further.*

Introduction

Nonverbal tests of false belief understanding are important for a number of reasons. Besides allowing testing of children or adults whose language skills are inadequate for the verbal tests widely in use, they can be used to test nonhuman animals. The great apes in particular, because of their close evolutionary relatedness to humans, can provide valuable insights into the nature of human social understanding. This is especially so when children's and apes' performance on the same tasks can be directly compared. Yet only a handful of studies have made such comparisons, especially in the domain of false belief understanding.

The main stumbling block has been that the standard tests involve considerable verbal interaction. For example, in the 'Sally-Anne' task (Baron-Cohen, Leslie & Frith, 1985; Wimmer & Perner, 1983), children watch as a protagonist (usually a doll or storybook character) places an object in one location and leaves the scene, whereupon another character arrives and moves the object to another location. She then leaves, the protagonist returns, and the experimenter asks children where the protagonist will look for the object. The other most commonly used false belief test, the 'Smarties' task (Perner, Leekam & Wimmer, 1987), also demands linguistic skills, as participants must verbally specify the true and believed contents of a candy box filled with crayons, for example. More recent tasks designed to elicit nonverbal responses (e.g. Carlson, Wong, Lemke & Cosser, 2005; Carpenter, Call & Tomasello, 2002; Sapp, Lee & Muir, 2000) still involve the processing of linguistic information such as instructions or requests, making them impossible for use with pre- or non-linguistic participants. Onishi and Baillargeon (2005) developed a truly nonverbal false belief test for infants (see also Southgate, Senju & Csibra, 2007; Surian, Caldi & Sperber, 2007), but the looking measure they use cannot say anything about participants' ability to translate understanding into adaptive action. Further nonverbal tests suitable for both children and nonhuman animals are needed if we ever hope to know if recognizing and responding adaptively to others' false beliefs crosses species boundaries.

Call and Tomasello (1999) made a first step in this direction with an object-choice task designed to test false belief understanding in children, chimpanzees and orangutans. One experimenter (the hider) hid a reward in one of two identical containers while a second experimenter (the communicator) observed. During training, the communicator then indicated the location of the reward for participants by briefly marking the correct container with a wooden block. Subsequently, in the false belief trials, after the reward was hidden the communicator left the area, whereupon the hider switched the positions of the containers. The communicator then returned and marked the incorrect container. If participants recognized the communicator's false belief about the reward's location,

Address for correspondence: Carla Krachun, Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive, Ottawa, Canada, K1S 5B6; e-mail: ckrachun@connect.carleton.ca or Malinda Carpenter, Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; e-mail: carpenter@eva.mpg.de

they should have chosen the unmarked container. Five-year-old children passed the test but younger children and both ape species failed. Control conditions confirmed that apes did not fail for reasons unrelated to the false belief element of the task. In a task by O'Connell and Dunbar (2003) using a somewhat similar logic, chimpanzees passed the false belief test but failed an important true belief control condition, calling the false belief results into serious question.

Thus far, then, apes have not shown any convincing signs of understanding false beliefs, whereas human children have. But recent findings suggest that apes may have been unfairly disadvantaged in previous studies. Namely, chimpanzees perform better in some experimental tasks when the social context is competitive rather than cooperative; and cooperative procedures that involve communicative elements are particularly problematic for them. For example, Hare and Tomasello (2004) demonstrated that chimpanzees in an object-choice task could not use a cooperative experimenter's pointing gesture to locate a food reward hidden in one of two containers. They could, however, locate the food with above-chance accuracy when the experimenter reached for the container in an apparent attempt to take the food. Support for the importance of competition also comes from a paradigm used to test chimpanzees' understanding of what others can or cannot see (Hare, Call, Agnetta & Tomasello, 2000) and know or do not know (Hare, Call & Tomasello, 2001). Dominant and subordinate apes were pitted against one another in a contest for hidden food. Subordinates witnessed all the hiding events and therefore knew where all food items were located. Dominants knew about only some of the food, either because it was clearly visible or because they were allowed to witness the hiding of that food. Subordinates avoided food dominants could see (or had seen hidden) in favor of food only they themselves knew about, providing evidence that they recognized states of seeing and knowing in others. (Note that Karin-D'Arcy & Povinelli, 2002, failed to replicate this result, but Bräuer, Call & Tomasello, 2007, suggest that important methodological differences could have played a role in this.) These findings contrast with earlier cooperative–communicative tests of mental-state attribution in apes, which yielded consistently negative results (e.g. Call, Agnetta & Tomasello, 2000; Povinelli & Eddy, 1996; Povinelli, Rulf & Bierschwale, 1994; Reaux, Theall & Povinelli, 1999).

Note that Hare and colleagues' (2001, Experiment 1) study described above included a change-of-location condition. Dominants saw the experimenter hide food behind one of two occluders, but the experimenter then moved the food to behind the second occluder when dominants were not looking. In a control condition dominants witnessed the switch. Subordinates were more likely to approach the food when dominants had not seen it moved than when they had. On the surface, this manipulation appears to be a false belief task, testing whether subordinates understood that dominants falsely believed the food to still be in its original location. However, all that subordinates needed to understand was that the dominant, not having seen the experimenter place the food behind the second occluder, had no idea it was there. Thus, chimpanzees could have solved the task by attributing lack of perceptual access to a hiding event rather than by false belief attribution (see also Gómez, 2004, for a similar interpretation of a study by Gómez & Teixidor, 1992).

If competition is critical, as recent research suggests, the following question arises: did apes fail Call and Tomasello's (1999) nonverbal false belief task because they had no understanding of false belief, or because a cooperative experimenter attempted to communicate the location of the food? Given the latter possibility, our goal was to create a nonverbal false belief test that was competitive, did not require participants to understand communicative intentions, and could be administered to both children and apes. Our task combined elements of Call and Tomasello's (1999) nonverbal false belief test and Hare and Tomasello's (2004) pointing-versus-reaching test. An experimenter hid a reward in one of two identical containers while a competitor observed (and participants, who could not see the hiding themselves, saw that she observed). Unbeknownst to the competitor, the experimenter then switched the locations of the containers, leading the competitor to falsely believe the reward to be in its original location. As the competitor reached with effort for the *incorrect* container, participants were given the opportunity to choose a container. For comparison, we included a true belief test in which the competitor witnessed the switch and thus reached for the correct container.

Our main question of interest was whether participants would show evidence of tracking the competitor's belief states by choosing the container she reached for in the true belief test but the container she did not reach for in the false belief test. We also considered that participants' choices might not tell the whole story. Participants could possess some recognition of the competitor's false belief that might not be reflected in their active choice responses but could nevertheless show up in other behaviors. Researchers have, for example, noted 'ancillary' behaviors suggestive of uncertainty or indecision in nonhuman animals faced with difficult decisions, such as looking or moving back and forth between options (see Smith, Shields & Washburn, 2003). Regarding apes in particular, Suda and Call (2006) observed that apes who performed at moderate levels in a liquid-conservation task acted more indecisively (e.g. by changing their choice of container) than apes who were either highly successful or unsuccessful in the task. The authors argued that this behavior indicated cognitive conflict in apes who could not reconcile the facts that one cup looked like it contained more liquid than the other cup, but should contain less. Likewise, if participants in our study showed more signs of indecisiveness

in false belief trials than in true belief trials, this could also be indicative of uncertainty, because the competitor reached in a way that looked like she knew where the food was even though she should not know its true location.

In false belief studies with human children, spontaneous looking behavior has become an increasingly popular measure because of its potential to tap into understanding not evident in elicited responses. Clements and Perner (1994) and Garnham and Ruffman (2001) presented children with a standard verbal change-of-location task, in which an object is moved in a story character's absence. In addition to asking children to indicate (by stating or pointing) where the story character would look for the object, the experimenters also coded anticipatory looking as the character was about to return. Children who indicated the incorrect location nevertheless often looked at the correct location. The authors concluded that these looks betrayed the presence of knowledge that was not available to children in making their active choices (they labeled it 'implicit'). Using a procedure in which children were asked to place bets indicating their confidence in their choices, Ruffman, Garnham, Import and Connolly (2001) demonstrated that correct anticipatory looking despite incorrect responding may alternatively indicate uncertain understanding, especially in older children. Southgate *et al.* (2007) argued that their positive anticipatory looking results show that even children as young as 2 years old attribute false beliefs to others, with performance limitations leading to failures in elicited-response tasks. Finally, violation-of-expectancy looking-time methods have been used to argue that infants barely older than a year are sensitive to a character's true and false belief states, as evidenced by their looking at a scene longer when a character's actions and beliefs are incongruent than when they are congruent (Onishi & Baillargeon, 2005; Surian *et al.*, 2007).

Looking responses have been used successfully as a measure with nonhuman primates (e.g. Myowa-Yamakoshi, Yamaguchi, Tomonaga, Tanaka & Matsuzawa, 2005; Santos, Seelig & Hauser, 2006; Uller, 2004), but we found no published reports of them being used in false belief studies with apes. In the current study, we therefore measured looking behavior across experimental conditions in both apes and children. Differences in looking across the true and false belief trials would suggest that participants, even if they failed our test, might nevertheless possess some degree of understanding not reflected in their active choice responses.[1] Children and apes were run concurrently. There were two slight variations on the experimental procedure, the main difference being whether or not the competitor was out of the room (version A) or had her back turned (version B) during the switching of the containers. Version B was added because the children seemed distracted by the competitor leaving the room in version A, and their performance in that version was weaker than expected. A different group of children participated in each of the two versions and all apes participated in both versions.

[1] We also considered analyzing two other ancillary behaviors: how often participants changed their choice and how often they attempted to choose both containers. However, both of these behaviors happened too rarely to be informative.

Study 1: Children

Method

Participants

Forty children recruited from kindergartens in Leipzig, Germany were tested, 20 with version A of the procedure and 20 with version B.

Version A. Participants were eight males and 12 females, 59–62 months old (mean = 61 months). Seven further children were tested but excluded from analyses, five because of procedural error and two for cheating (e.g. standing to see over the screen blocking their view, or trying to look inside the containers). Four additional children failed to meet the pretest criterion (see below).

Version B. Participants were 11 males and nine females, 54–61 months old (mean = 58 months). Four more children were excluded from analyses because of procedural error, and two children's test sessions were abandoned due to uncooperativeness.

Experimental set-up

Children sat facing an adult competitor across a table (100 × 60 cm) with a sliding platform that could be moved back and forth between the child and competitor (see Figure 1). Two small opaque containers were used to conceal a sticker reward. An opaque screen (32 cm high, 59 cm wide) obstructed the child's view of the containers when necessary. Plexiglas panels (52 cm high, 59 cm wide) blocked the child's and the competitor's access to the containers except through two armholes (39 cm apart from the center of each hole).

Design

All children were first given a warm-up and pretest. Half of participants then received four true belief trials followed by four false belief trials and half received the opposite order. A standard verbal false belief test (the Sally-Anne test; Baron-Cohen *et al.*, 1985) was also administered (in version A all children received it at the end of the session; in version B half the children received it at the beginning and half at the end). Version B also included four control trials administered immediately

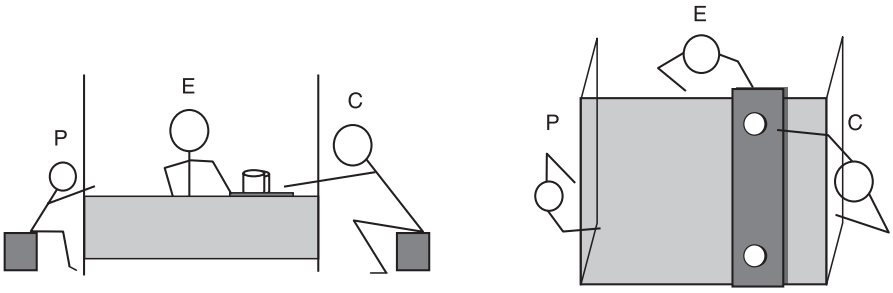

Figure 1 *Experimental set-up: front view on left and top view on right. P = Participant, E = Experimenter, C = Competitor.*

before the test trials. Location of the sticker was randomly determined with the constraints that it had to be on the left and right an equal number of times and could not be in the same container for more than two consecutive trials. All trials were administered in one session.

Procedure

Warm-up. Children were tested in a quiet room in their kindergartens by two female experimenters, one acting as the competitor. They were first given two warm-up trials to introduce them to the general procedure and establish the competitive context. An experimenter (E) put two containers onto the sliding platform, one on the left and one on the right. She explained that they were going to play a game in which the child had to try to get a sticker hidden in one of the containers, and that the competitor would also try to get the sticker. E then placed the sticker in one container while both the child and competitor watched. E began sliding the platform first in the direction of the competitor, who reached with effort but unsuccessfully for the container with the sticker inside (hereafter, the 'correct' container). E then slid the platform over to the child, who could then reach for and choose a container. To establish the competitive context, in the second trial the competitor managed to reach the container and take the sticker before children could do so. The competitor acted pleased when she got the sticker and disappointed when the child got it, and she often made competitive comments such as 'Okay, this time *I'm* going to get the sticker!'

Pretest. Following the warm-up, a pretest was given to verify that participants could use the competitor's reach as a cue to the reward's location. In these trials, before hiding the sticker E positioned an opaque screen to block the child's (but not the competitor's) view of the containers. Children could see the competitor's face over the screen, and they observed as the competitor watched E hide the sticker. The competitor witnessed the hiding with clear interest, leaning forward, nodding her head, and saying things such as 'ah hah' to show she was paying attention. So that participants could not use the competitor's gaze to infer where the reward was being hidden, the competitor directed her gaze straight ahead during the hiding, occasionally looking rapidly back and forth between the containers. She did not track the reward and never focused her gaze on one container. When E removed the screen the competitor reached unsuccessfully for the correct container. E then slid the containers towards children so they could choose a container. If children understood that the competitor knew the location of the sticker because she had witnessed the hiding, they should choose the same container as she did.

In version A, children received 12 pretest trials and needed at least nine correct to pass. Because the pretest proved to be unnecessarily long (most children earned perfect scores), for version B we reduced the criterion to three trials in a row correct.

False belief test. As in the pretest, E blocked the child's view of the containers with a screen and hid the sticker as the competitor observed (and the child saw that she observed). E removed the screen and the competitor immediately either left the room (version A) or turned around in her seat (version B) before she had a chance to reach. She gave some excuse for doing so (e.g. to make a phone call or blow her nose). E then got the child's attention and switched the positions of the containers without revealing the location of the sticker, smiling mischievously and glancing occasionally at the door (version A) or the competitor's back (version B) to make sure she did not see the switch. In version B, the competitor muttered to herself and became highly absorbed in her task while her back was turned, clearly not attending to E's actions. E secretly signaled the competitor when she was finished (by coughing inconspicuously), and the competitor returned to her position facing the child. E then slid the platform first in the direction of the competitor, who reached with effort, but unsuccessfully, for the incorrect container. E next slid the platform over to the child while the competitor continued to reach. Children who recognized the competitor's false belief should choose the container the competitor was *not* reaching for. Children who chose the correct container were allowed to keep the sticker; otherwise E slid the platform back over to the competitor, who took the sticker.

The competitor showed obvious surprise when the location of the reward was revealed. And although the

task was nonverbal, the experimenter and competitor chatted naturally with each other and with children throughout, while being careful to avoid references to the competitor's belief states. For example, when E displayed the sticker before hiding it, the competitor said excitedly, 'Oh, a smiley/dolphin/etc.! I'm going to get it!' If the child chose the incorrect container E made statements such as, 'Let's try again.'

True belief test. The true belief procedure was the same as for false belief, with the crucial difference that the competitor witnessed the switching of the containers. As before, the competitor left the room or turned her back without reaching after E hid the sticker. But in this case E just sat and waited, glancing occasionally at her watch, the child, and the door or the competitor's back. When the competitor resumed her position at the table, E switched the positions of the containers in full view of both the child and the competitor. The competitor then reached for the correct container, and the proper response for children was to also choose that container.

Control trials. In version B we also included control trials immediately preceding the test trials. These trials were added (after results were obtained for version A) in order to rule out alternative reasons why participants might have difficulty with our false belief test. The control trials were designed after ones used by Call and Tomasello (1999), and they evaluated two non-mentalistic component skills required to do well in our task, as follows:

Two *ignore-competitor* trials tested participants' capacity to disregard the competitor's reach when they knew it was wrong. The procedure was as for the false belief trials except that while the competitor's back was turned E removed the sticker from the container in full view of the child and moved it to the other container. When the competitor faced forward again she reached for the incorrect container. Children who could ignore the competitor's incorrect reach should choose the other container.

Two *invisible displacement* trials tested participants' knowledge that the reward moved from one location to the other when the containers were switched. While the competitor watched, E hid the sticker and then removed the screen blocking the child's view. The competitor immediately started reaching for the correct container but then became distracted, removed her arm, and turned around in her seat. While her back was turned, E switched the locations of the containers. When the competitor faced forward again she did *not* resume reaching as E slid the containers towards the child. If children understood that the competitor's reach indicated the initial location of the sticker, but that the sticker had since been moved to the other location along with its container, they should look for the sticker in the new location. The two types of trials were administered in alternating order, always beginning with an invisible displacement trial.

If participants failed the false belief test, then their performance on the control trials would be informative: if they also did poorly in the control trials, then they could have failed the false belief test *not* because they did not understand false beliefs, but because they did not possess the necessary non-mentalistic component skills.

Sally-Anne test. We administered the Sally-Anne test to confirm that the children in our sample were typical for their age regarding performance in a standard verbal test. One of the experimenters acted out the test for children using dolls named Max and Hannah, a plastic cookie, a small metal pot with a lid, and an opaque plastic jar. According to the standard procedure (Baron-Cohen *et al.*, 1985), Max put the cookie in the pot and left briefly, during which time Hannah entered and moved the cookie to the jar. Hannah then left, Max returned, and the experimenter asked children the customary test question and two control questions in a fixed order: (1) 'Where will Max look for the cookie first?' (2) 'Where is the cookie now?' and (3) 'Where was the cookie before Max left?' Children passed only if they answered all three questions correctly, either verbally or by pointing.

Coding and analysis

Choice. Our main measure was whether or not participants chose the correct container. Choice was defined as the container participants had settled on by the time E finished sliding the platform over to them. Children chose by reaching an arm through a hole in the Plexiglas barrier and touching a container. Occasionally they began to reach through both holes at once, in which case E asked them to put only one arm through at a time.

We used proportion scores because two trials had to be excluded from the analyses: one true belief trial because of procedural error (the competitor began reaching before the switch occurred), and one false belief trial because the child refused to make a choice. We calculated choice proportion scores for each participant in both the true and false belief conditions by dividing the number of correct trials by the total number of trials. Average scores were compared to the proportion of trials expected to be correct by chance (.50). However, we considered that the proportion of incorrect true belief trials might provide a more meaningful comparison for false belief performance than .50 chance in our task (see Carpenter *et al.*, 2002, and Lohmann, Carpenter & Call, 2005, for a similar argument regarding other tasks). That is, because of the high salience of the competitor's reaching cue, if participants had no understanding of false beliefs they would be likely to choose the container the competitor reached for in both conditions rather than choosing randomly (i.e. at chance levels as a group). If participants chose that container perfectly consistently they would achieve 100% success in true belief trials and 0% success in false belief trials. But even simple strategies can break down on occasion, for example because of

momentary distraction. This would result in some proportion of failures in true belief trials and a corresponding proportion of accidental successes in false belief trials. It is thus informative to compare the proportion of successes in false belief trials to the proportion of failures in true belief trials (what we call the *true belief complement*, or 1 minus the true belief score). If these two proportions do not differ, it suggests that participants are using the same strategy across conditions. However, if the false belief proportion is significantly higher than the true belief complement this suggests that participants are discriminating between conditions and thus have some understanding of the competitor's false beliefs. Note, however, that we cannot know for certain whether the most appropriate comparison is to chance or to the true belief complement so we also provide the more conservative comparison to chance.

Looking. Our second measure was whether participants looked *at least once* at the container the competitor was *not* reaching for during the couple of seconds it took E to slide the platform towards them. We coded only the subset of trials in which participants chose the same container as the competitor, for three reasons: (1) we were only interested in false belief trials in which participants displayed no understanding of false belief in their active choices, (2) trials in which participants chose the other container confounded choosing with looking, and including them could therefore artificially inflate the false belief results, and (3) by coding only the subset, we could directly compare true and false belief trials on an equal footing, because participants chose the container the competitor reached for in both.

Because we used within-subjects comparisons, only children who chose the same container as the competitor at least once in each condition were included in analyses. There were 13 such children in each of versions A and B. After excluding the other children, the following percentage of all trials remained for the looking analysis: 55% of true belief trials and 41% of false belief trials from version A; and 49% of true belief trials and 26% of false belief trials from version B. Our question was thus: before choosing the same container as the competitor, did children look at the other container in a greater proportion of false belief trials than true belief trials?

For each participant, we calculated a looking proportion score in each condition by dividing the number of trials in which the child looked by the number of trials retained for that child in that condition. Choice was coded live and videotapes were used to code looking. All analyses were non-parametric and all reported *p*-values are exact and two-tailed.

Reliability

To assess inter-observer reliability, coders naïve to the hypotheses of the study independently coded 100% of children's experimental trials for choice and 35% (version A) and 30% (version B) of the subset trials for looking. These trials were chosen randomly. Excellent levels of agreement were achieved: Cohen's kappas were 1.00 for choice in both versions and .92 (version A) and .83 (version B) for looking.

Results and discussion

Preliminary analyses revealed no significant sex or age differences in false belief choice scores for either version A or B (all $ps \geq .43$). We therefore collapsed across these variables for all subsequent analyses.

Pretest

Children demonstrated from the outset a clear ability to use the competitor's reach to find the hidden reward. In version A the criterion to pass was at least 9/12 trials correct, and children's average number correct was 11.6 (97%). In version B, in which the criterion was three consecutive trials correct, children's average number of trials to reach criterion was 4.3.

Choice

Order of presentation of the true and false belief tests had no effect on performance in either test in either version A or B (Mann-Whitney *U* tests: all $Us \geq 40.50$, $Ns = 20$, $ps \geq .47$). There were also no statistically significant differences in children's true belief (TB) or false belief (FB) performance across versions (TB: $U = 136.00$, $N = 40$, $p = .062$; FB: $U = 153.00$, $N = 40$, $p = .20$). However, TB–FB difference scores (calculated by subtracting the false belief from the true belief proportion score for each child) *were* significantly different across versions ($U = 124.50$, $N = 40$, $p = .038$), indicating that children's degree of success in the true belief as compared to the false belief condition differed across versions. We therefore present the results combined and for versions A and B separately.

Figure 2a shows the mean proportion of correct choices by children in the true and false belief tests, both combined across versions and separately (also shown is the true belief complement). In the combined analysis, children performed significantly better than chance in the true belief test, verifying that they did not have any basic problems with the general procedure (Wilcoxon test: $T^+ = 582.00$, $N = 34$[6 ties], $p < .001$). They also performed better than chance in the false belief test ($T^+ = 387.00$, $N = 31$[9 ties], $p = .004$), indicating that they recognized the competitor's false belief.

In versions A and B separately, children's true belief scores still exceeded chance (version A: $T^+ = 187.00$, $N = 19$[1 tie], $p < .001$; version B: $T^+ = 116.00$, $N = 15$[5 ties], $p < .001$). In version B, children's false belief score was also significantly greater than chance ($T^+ = 134.00$, $N = 17$[3 ties], $p = .004$), as in the combined analysis. In version A alone, however, children's false belief score

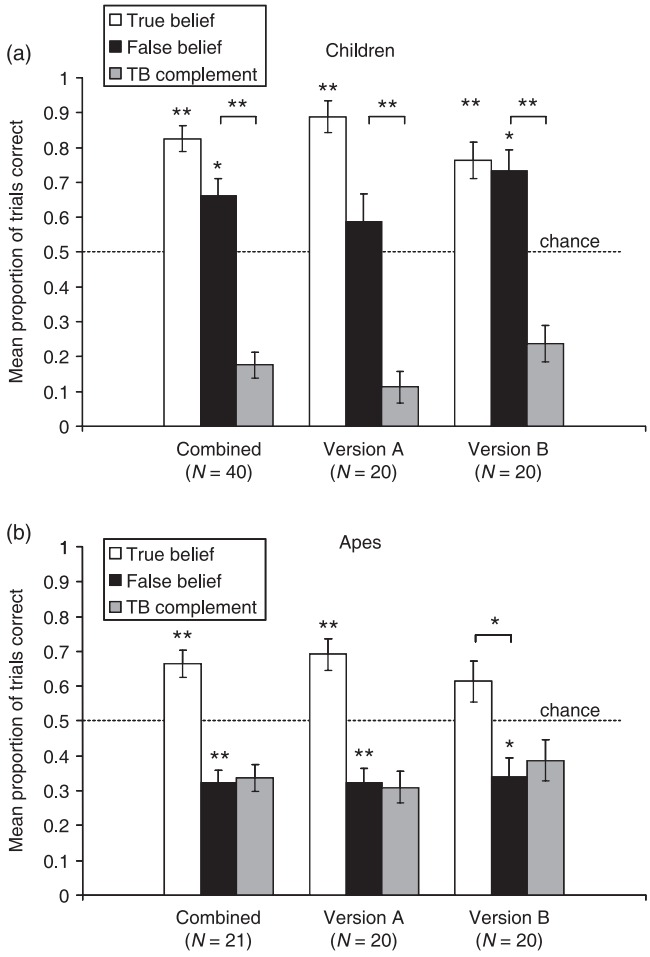

Figure 2 *Choice analysis: mean proportion of trials correct for (a) children and (b) apes in both versions of the procedure combined and separately (TB complement = true belief complement; see text for explanation). Bars show standard error.* * p < *.01;* ** p < *.001.*

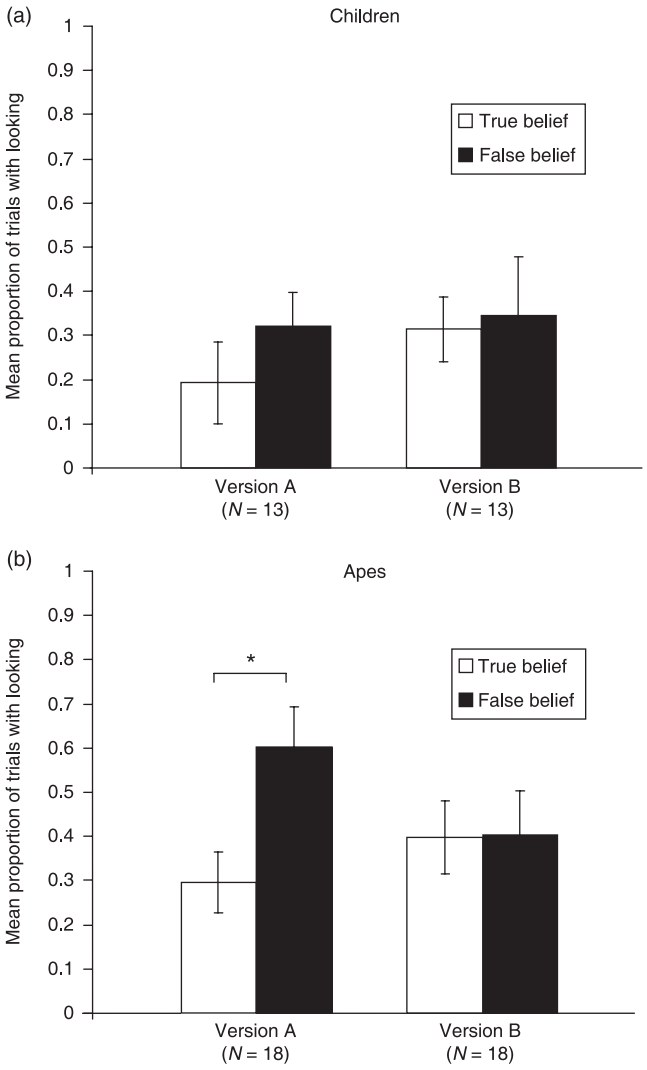

Figure 3 *Looking analysis: mean proportion of trials in which (a) children and (b) apes looked towards the unchosen container before choosing. Note that proportions are based only on trials in which participants chose the same container as the competitor, and for apes in version A the measure was face rather than eye orientation. Bars show standard error.* * p < *.05.*

was numerically but not significantly higher than chance ($T^+ = 73.00$, $N = 14$[6 ties], $p = .19$), suggesting that children found version A more difficult than version B. However, children's false belief score *was* significantly higher than the true belief complement in version A ($T^+ = 133.00$, $N = 16$[4 ties], $p < .001$), indicating that they did respond differentially across conditions (this was also true for version B alone [$T^+ = 136.00$, $N = 16$(4 ties), $p < .001$] and for the combined analysis [$T^+ = 523.50$, $N = 32$(8 ties), $p < .001$]). Thus, if the true belief complement, rather than chance, is used as the comparison with false belief, children's performance in version A suggests some recognition of the competitor's false belief state.

We did two final analyses to get a clearer sense of how children's performance varied across versions A and B. First, we set the passing criterion for a test at three or more trials correct out of four. According to this criterion, for the false belief test, 40% of children passed in version A and 70% passed in version B. For the true belief test, 90% of children passed in version A and 70% passed in version B. Second, we examined children's performance across versions using each child's very first trial only. In version A, 60% of the children whose first trial was a false belief trial got that trial correct, compared to 70% in version B. For true belief, the percentages were again 90% for version A and 70% for version B.

Looking

The looking measure was included to see if participants might still show some degree of false belief understanding, even if they did not perform well in their active choices. Because children's false belief performance was somewhat weak in version A, looking was of most interest in this version. Nevertheless, we analyzed both versions for comparison. Figure 3a shows that in version A, looking occurred in a numerically greater proportion of false belief than true belief trials (TB = .19, FB = .32), but

this difference was not statistically significant (T^+ = 21.00, N = 7[6 ties], p = .30). There was no difference in version B (TB = .31, FB = .35; T^+ = 33.00, N = 11[2 ties], p = 1.00), which was to be expected given that children did so well on the active choice measure in that version.

Sally-Anne test

All children who answered the Sally-Anne test question correctly also answered both control questions correctly. There were no significant differences in children's Sally-Anne performance across versions (U = 190.00, N = 40, p = 1.00). There were also no observable order effects in version B, in which the Sally-Anne test and our test were counterbalanced for order (for both tests, $U \geq 33.00$, $N_{first} = N_{last} = 10$, $p \geq .20$). We therefore report the results collapsed across both versions and order. Confirming that the children in our sample were age-typical regarding their performance on the verbal false belief test, the proportion of children passing the Sally-Anne test was significantly greater than the chance proportion of .50 (Binomial test: N = 40, p = .04). As a group, children's performance in the Sally-Anne test closely matched their performance in our nonverbal test: proportion scores (.68 and .66, respectively) were not significantly different (T^+ = 237.50, N = 30[10 ties], p = .92). Because the Sally-Anne test included only one trial, we also looked at children's performance in the first trial only of our nonverbal false belief test (in both versions combined). The proportion of children responding correctly on that trial was .60, also not significantly different from the Sally-Anne results (T^+ = 110.00, N = 19[21 ties], p = .65). These findings verify that, as a group, our child participants were equally challenged by the Sally-Anne test and our nonverbal test.

We were also curious to know how individual children's scores matched up across the Sally-Anne test and our nonverbal test, given that the two procedures were so different. We compared children's first nonverbal false belief trial with their Sally-Anne test trial. Twenty-one of the 40 children (53%) were consistent across tests, either failing or passing both. A test of agreement yielded a non-significant kappa of –0.02 (p = 1.00). Furthermore, when we included all four nonverbal false belief trials and split the children into two groups, those who passed the Sally-Anne test and those who failed it, the former did no better as a group on our nonverbal test than the latter (U = 170.00, N_{pass} = 27, N_{fail} = 13, p = .88). This pattern of results was also found for each version of the procedure analyzed separately (version A: U = 35.50, N_{pass} = 13, N_{fail} = 7, p = .47; version B: U = 34.00, N_{pass} = 14, N_{fail} = 6, p = .53). In short, performance was highly consistent across tests at the group level but was inconsistent at the individual level, probably because of differing task features unrelated to the false belief element of the tests. We return to this issue in the General discussion.

Control trials

Children performed well in the control trials, which were run in version B only. All children earned a perfect score in the two ignore-competitor trials. In the invisible displacement trials, the mean proportion of trials correct (.80) was far above chance (T^+ = 78.00, N = 12[8 ties], p < .001). Given that children did well in the false belief trials in version B (and also did well in the control trials) comparisons between the two types of trials are not so informative. However, since we report the comparison between invisible displacement and false belief performance for apes below, where it is informative, we provide it here as well. Children's performance on invisible displacement trials was not significantly different from their performance on false belief trials (T^+ = 71.00, N = 15[5 ties], p = .58). There was no significant correlation between performance on invisible displacement trials and false belief trials (Spearman's rank correlation: r_s = –0.20, N = 20, p = .42), which is not surprising given that many children were at ceiling levels with regard to invisible displacement.

Overall, the results for study 1 (at least in version B) were typical for children in the age group we tested. By choosing the container the competitor reached for more often in the true belief trials and the one she did not reach for more often in the false belief trials, children demonstrated that they recognized the competitor's false belief about the location of the reward. The next question was whether apes would also show such insight. We believed that the competitive nature of our task would give them a better chance of succeeding than in previous tasks using cooperative paradigms.

Study 2: Apes

Method

Participants

Sixteen chimpanzees (*Pan troglodytes*) and five bonobos (*Pan paniscus*)[2] housed in social groups at the Wolfgang Köhler Primate Research Center (WKPRC) in Leipzig, Germany participated in both versions A and B. An additional chimpanzee was tested but dropped because he would not make clear choices. Participants included six males and 15 females, all captive-born, nine mother-reared and 12 nursery-reared. Their ages ranged from 4 to 28.5 years when testing began (mean = 14.8 years). All apes had previously taken part in a variety of social and physical cognition studies and were familiar with

[2] We had no *a priori* reason to think that bonobos (which are in the same genus as chimpanzees and just as closely related to humans evolutionarily) would perform any differently in the task than chimpanzees. They were included simply to boost our number of ape participants.

the object-choice paradigm. Eleven chimpanzees had participated several years earlier in a study in which they used a human competitor's reach to find hidden food (Hare & Tomasello, 2004). Apes were not food-deprived and water was provided *ad libitum* throughout testing. One chimpanzee's version A data were excluded from analyses because she repeatedly tried to look into the containers before choosing. Another chimpanzee's version B data were excluded because of procedural error. The final number of apes in each version was thus 20.

Experimental set-up

The set-up was the same as the children's except for the following: the table (92 × 80 cm) was placed in a three-sided windowed testing booth between two adjoining enclosures. The competitor and ape each sat in a different enclosure, with the experimenter just outside the booth at the side of the table. Rewards were grapes or banana slices. The Plexiglas barriers (49 cm high, 69 cm wide) had holes in them spaced 59 cm apart from center to center. For the competitor, the holes were large enough to fit her arms through; for safety reasons the apes' holes were just large enough for their fingers. The opaque screen used to block apes' view was 28 cm high and 64 cm wide.

Design

All apes received version A first and then B. In both versions apes received a warm-up and pretest. In version A, apes then received eight true belief and eight false belief trials (four of each per session for two sessions). At the start of each new test session we gave apes six further pretest trials to refresh their memory of the procedure. In version B, apes received four true belief and four false belief trials in a single testing session. Each ape thus received a total of 12 true and 12 false belief trials. In both versions, the true and false belief trials were administered in blocks of four trials, counterbalanced for order across apes, with the location of the reward randomized as for children. Version B also included four control trials immediately preceding the test trials. Finally, in a later session, two warm-up trials and four post-test control trials were also administered.

Procedure

Warm-up. Apes were also tested by two female experimenters. The warm-up was the same as for children except that, because we could not verbally explain the task, we gave apes three warm-up trials at the start of version A (and then two more at the start of version B). To establish the competitive context, the competitor successfully took the food in the second trial (or sometimes the third trial in version A) before the ape could do so. She also adopted a demeanor more appropriate for competing with apes. While watching E hide the food, she performed behaviors indicative of attention and excitement in chimpanzees, such as head bobbing and grunting. When the competitor won the reward she gave the impression of consuming it greedily, and when she lost the reward she reacted angrily by banging on the windows or walls and shouting.

Pretest. Again, because we could not verbally explain the task to the apes we gave them a greater number of pretest trials than the children. In version A apes received 18 trials in a session and needed to get at least 13 correct to pass. They were given repeated sessions as needed until they reached criterion. Because the same apes participated in version B, they received the shorter three-consecutive-trials pretest in that version, like children.

True and false belief tests. These were exactly as for children except that the competitor, for obvious reasons, did not give an excuse for leaving the room or turning around. In version A she left when a bell rang, and in version B she simply turned around and became absorbed in some activity (e.g. scratching her leg or tying her shoe).

Control trials. Immediately preceding the test trials in version B, apes received two ignore-competitor and two invisible displacement trials (alternating, as for the children). However, for the invisible displacement trials, because Call and Tomasello (1999) found that apes did better when a marker indicating the reward's location was present both during the switch and while apes made their choice, the competitor left her arm resting in the hole while her back was turned. When she faced forward again after the switch, she resumed reaching (now towards the incorrect container) as the ape chose a container.

Post-test control trials. In retrospect, we noted that apes' modified invisible displacement control trials contained an element of false belief: the competitor's reach towards the incorrect container while the ape chose. Therefore, after testing was complete we gave apes four additional invisible displacement trials. In two *back-turned* trials (more similar to the children's), the competitor left her arm in the hole while her back was turned and the switch happened, as before, but when she faced forward again she removed her arm and did not resume reaching while the ape chose. In two *facing-forward* trials, the competitor did not turn her back during the switch, and she removed her arm from the hole just as E began to switch the containers. These trials also alternated, always beginning with a back-turned trial.

Coding and analyses

Choice. The container chosen (coded live) was defined as the one the ape was oriented towards and actively

poking at when E had finished sliding the platform. Occasionally, an ape tried to choose both containers, in which case E waited until the ape had clearly chosen one container. If necessary, she slid the containers away from the ape and back again. A number of trials were excluded from analyses because of procedural error (three trials), the ape did not clearly choose a container (three trials), or the ape refused to participate (eight trials, all from the same individual). Proportion scores were calculated as for the children.

Looking. For version A, we were unable to obtain adequate inter-observer reliability for eye direction because the apes' eyes were difficult to see on the video due to poor lighting. We therefore used face orientation as the looking measure. While we could not be positive that it correlated perfectly with eye direction, it would seem unnatural for apes to turn their face without also turning their eyes. Nevertheless, they could orient in the direction of a container without focusing on it, in which case face orientation would overestimate looking. On the other hand, they could also direct their eyes towards a container without turning their face, in which case face orientation would underestimate looking. However, both of these possibilities should be equally likely across conditions, and so the face orientation measure should not introduce any systematic biases into the coding. For version B, we improved the lighting and were therefore able to code looking from eye direction. We used this measure rather than face orientation, although it was different from version A, because we felt it better to use the more precise measure when possible.

For looking, we coded the first four true belief and first four false belief trials in each version (i.e. the first block only in version A and all trials in version B). We limited coding to these trials because the difficulty of the coding made the task extremely time consuming, and also because children received only one block of trials in both versions (so this way we could directly compare children and apes). We also considered early trials to be far more meaningful than later ones, as responding in later trials could be affected by earlier experiences.

Eight trials were excluded from the looking analysis because they were not recorded or the ape's face was offscreen during the relevant part of the trial. As with children, only the subset of trials in which apes chose the same container as the competitor was coded. Again, apes needed to choose the same container as the competitor at least once in each condition to be included in analyses. In each version, 18 apes met this criterion. The percentage of all trials included in the looking analysis was as follows: 58% of true belief and 59% of false belief trials from version A (block 1), and 59% of true belief and 60% of false belief trials from version B. All analyses were nonparametric and all reported *p*-values are exact and two-tailed unless noted.

Reliability

To assess reliability, coders who were naive to the hypotheses and blind to condition coded 40% (version A) and 35% (version B) of test trials for choice. Both kappas were 1.00. They also coded 55% of subset trials for looking in version A and 31% in version B, with respective kappas of .81 and .91. A greater number of version A trials were coded because the coding was more difficult and we wanted to be more confident of the results. Trials were chosen randomly.

Results and discussion

Preliminary analyses revealed no significant sex, age, species, or rearing differences in false belief performance in either version of the procedure (all ps $\geq$ .18). We therefore collapsed across these variables for all analyses.

Pretest

In version A, apes' average number of pretest sessions to reach the criterion of 13/18 trials correct was 1.9 (range = 1–4) and the average passing score was 14.5 (80.6%). In version B, apes' average number of trials to reach the criterion of three consecutive trials correct was 5.6.

Choice

There were no order effects (all Us $\geq$ 35.50, Ns = 20, ps $\geq$ .28) and no differences in true or false belief performance across versions A and B (TB: T^+ = 63.50, N = 14[5 ties], p = .51; FB: T^+ = 55.50, N = 14[5 ties], p = .87). There was also no significant difference in the TB–FB difference scores across versions, indicating that apes' pattern of responding was similar in each (T^+ = 75.00, N = 15[4 ties], p = .41). We nevertheless report the results for versions A and B combined and separately, for direct comparison with the children.

In false belief trials, apes' performance (Figure 2b) contrasted sharply with children's: they were significantly *worse* than chance in both the combined and separate analyses (combined: T^+ = 183.00, N = 19[2 ties], p < .001; version A: T^+ = 116.00, N = 15[5 ties], p < .001; version B: T^+ = 70.00, N = 12[8 ties], p = .014). False belief performance was also not significantly different from the true belief complement (combined: T^+ = 86.00, N = 18[3 ties], p = 1.00; version A: T^+ = 55.50, N = 14[6 ties], p = .87; version B: T^+ = 79.00, N = 16[4 ties], p = .59). In their active choice responses, apes therefore gave no sign that they recognized the competitor's false beliefs.

In contrast, in the true belief control condition, apes performed significantly better than chance in both the combined analysis and in version A alone (combined: T^+ = 117.00, N = 15[6 ties], p < .001; version A: T^+ = 89.00, N = 13[7 ties], p = .001), suggesting that they had no major problems with the basic structure of the task. Their version B performance analyzed separately only

approached significance in a one-tailed analysis (T^+ = 69.50, N = 13[7 ties], $p_{\text{two-tailed}}$ = .11, $p_{\text{one-tailed}}$ = .054), although it was numerically higher than chance and significantly better than false belief (T^+ = 136.50, N = 17[3 ties], p = .003). Apes' true belief performance may have deteriorated by the time they participated in version B because they had by then repeatedly used the strategy of choosing the container the competitor reached for, only to have it fail in the false belief trials. This could have prompted them to choose less systematically in later trials, bringing their performance down closer to chance levels.

To be consistent with our analysis of the children's data, we also examined apes' true belief and false belief performance in each version using a pass criterion of 75% of trials correct (at least 6/8 trials correct in version A and at least 3/4 correct in version B). For the false belief test, 0% of apes passed in version A and 10% passed in version B. For the true belief test, 50% of apes passed in both versions. And analyzing just the very first trial, in version A, 11% of the apes whose first trial was a false belief trial got that trial correct, compared to 40% in version B. For true belief, the percentages were 44% for version A and 80% for version B. Because the same apes participated in both versions, however, comparing first-trial performance across versions is less informative for the apes than for the children.

An important issue is how the apes' previous experience using a competitor's reach to find hidden food may have affected their performance in the false belief test. Apes with more experience using the reaching cue successfully in the past might have more difficulty choosing contrary to the reach in the false belief test. We found no evidence of this when comparing the false belief performance of the 11 apes who took part in Hare and Tomasello's (2004) pointing-versus-reaching test with those who did not (for each version separately and combined, $U \geq 31.00$, N_{separate} = 20, N_{combined} = 21, $p \geq .15$). Additionally, looking at differential experience with the reaching cue within our study, there was no significant correlation between the number of pretest trials to reach criterion and false belief performance (separately and combined, $|r_s| \leq .16$, N_{separate} = 20, N_{combined} = 21, $p \geq .49$). Previous experience thus did not affect the apes' performance.

Looking

In contrast to the choice results, the apes' looking results (see Figure 3b) were slightly more suggestive of some level of false belief understanding. The pattern of looking responses was significantly different across the two versions of the procedure, justifying a separate analysis of each (T^+ = 47.00, N = 10[5 ties], p = .045). In version A, apes oriented their face toward the unchosen container in a significantly greater proportion of false belief than true belief trials (T^+ = 95.50, N = 15[3 ties], p = .041). These findings are consistent with the interpretation that apes may have recognized the competitor's false belief on some level. However, this result was not replicated in version B, in which apes looked toward the unchosen container equally often in both conditions (T^+ = 46.50, N = 13[5 ties], p = .96). We considered a possible alternative explanation for the positive looking result in version A: apes who chose the container the competitor reached for in a false belief trial would not receive the reward. This failure could make them unsure of this strategy in subsequent trials and cause them to be more uncertain about their choices. To test this possibility, for each ape in the version A looking analysis, we selected out their first true belief subset trial and first false belief subset trial (in which responding could not be influenced by earlier failures) and compared these. Looking occurred in 33% of true belief trials and 61% of false belief trials. Although this difference was not statistically significant (T^+ = 63.00, N = 13[5 ties], p = .27), the tendency for apes to look at the unchosen container more often in the false belief condition was clearly present from the earliest trials.

Control trials

Because apes failed the false belief test, it was important to see whether their failure could have been due to some non-mentalistic component of the task. With the exception of one chimpanzee (who got one trial wrong), all apes earned perfect scores in the two ignore-competitor trials. For the pretest invisible displacement trials, the mean proportion correct was .40, not significantly different from the chance proportion of .50 (T^+ = 38.50, N = 10[10 ties], p = .34). However, the false belief element of the pretest invisible displacement trials (see Procedure) may have negatively affected apes' performance. The results of the post-test invisible displacement trials support this idea, as apes' performance improved in those trials. The mean proportion of trials correct (.60) was numerically higher than chance, and this difference approached significance (T^+ = 61.50, N = 12[8 ties], p = .087). It should be noted that the apes had received many trials by this point, and being tired with the whole affair may have weakened their performance in the post-test trials. This is consistent with the deterioration in true belief performance and the loss of differential looking across conditions in version B.

We also directly compared apes' performance on invisible displacement trials and false belief trials. Apes' pretest invisible displacement scores were not significantly better than their false belief scores (T^+ = 52.00, N = 13[7 ties], p = .66), likely because of the false belief element in the pretest invisible displacement trials, as noted above. However, apes' performance in the post-test invisible displacement trials was significantly better than their false belief performance, whether we collapsed across the two different types of trials or analyzed them separately (in all cases, $T^+ \geq 82.00$, $N \geq 13[\leq 7$ ties], $p \leq .017$).

In addition, we compared false belief proportion scores in apes who passed the post-test invisible displacement trials and apes who failed them (using a pass criterion

for the invisible displacement trials of at least three out of four trials correct). If difficulty with invisible displacement explained apes' poor false belief performance in general, then apes who failed these control trials should have lower false belief scores. Mean false belief proportion scores for these two groups were .39 and .30, respectively, not significantly different ($U = 39.50$, $N_{passed} = 9$, $N_{failed} = 11$, $p = .50$). There was also no significant correlation between post-test invisible displacement scores and false belief scores ($r_s = .22$, $N = 20$, $p = .37$). In short, any difficulties apes might have had with invisible displacement were not solely responsible for their poor performance in our false belief test. The false belief element of the task, rather than its non-mentalistic components, appears to have been the limiting factor on apes' performance.

General discussion

We directly compared the performance of children and apes in a new nonverbal false belief test. The test was more species-relevant for apes than previous ones (Call & Tomasello, 1999; O'Connell & Dunbar, 2003) because it was competitive rather than cooperative and did not involve communicative elements. Nevertheless, the picture that emerged was consistent with earlier findings: whereas $4\frac{1}{2}$- to 5-year-old children passed the false belief task, both ape species failed. Children responded differentially in true belief and false belief trials in both versions of the procedure, and in version B their choices were unambiguously consistent with recognition of the competitor's false beliefs. They most often chose the container the competitor reached for in true belief trials but the other container in false belief trials. In contrast, apes most often chose the container the competitor reached for in both true and false belief trials, across both versions of the procedure. This strategy served them well in the pretest and true belief trials, in which the competitor always reached for the correct container; but it was disastrous in the false belief test, in which their performance was worse than if they had simply guessed.

The question that arises at this point is: did apes' poor performance in our task reflect a true lack of understanding of false beliefs, or could other factors have interfered with their performance? One might argue that our task was too difficult for apes for reasons unrelated to false belief understanding. It is possible, for example, that apes found the competitor's reach so compelling that they could not inhibit a tendency to choose the container she reached for, even when they suspected she was wrong. There is some evidence that inhibitory issues of this sort can influence apes' responding in experimental tasks (Boysen & Berntson, 1995; Boysen, Berntson, Hannan & Cacioppo, 1996). For two reasons, however, we do not think this was a significant problem in the current study. First, apes had no trouble ignoring the competitor's reach when they knew it was wrong in the ignore-competitor control trials. Second, they had a couple of seconds to make a choice after the competitor began reaching, while the containers were being slid within their reach. Even if their initial impulse was to reach for the same container as the competitor, they had the opportunity to reverse their decision and choose the other container. They rarely did so, and certainly no more than children, who also rarely changed their choice. One might also argue that competing with humans, rather than conspecifics, was unnatural for the apes. This seems an unlikely explanation, however, given that apes have succeeded in other tasks requiring some degree of mental-state understanding, even when interacting with humans rather than conspecifics (Call, Hare, Carpenter & Tomasello, 2004; Hare, Call & Tomasello, 2006; Hare & Tomasello, 2004; Melis, Call & Tomasello, 2006). Other nonhuman primates have also succeeded in tasks in which they needed to infer humans' perceptual states in competitive situations (Flombaum & Santos, 2005; Vick & Anderson, 2003).

Along with active choice, we measured looking in the subset of trials in which participants chose the same container as the competitor. The positive results for apes in version A of the procedure gave some indication that they may not have been entirely oblivious to the competitor's belief states. In that version, apes looked at (i.e. oriented their face toward) the container the competitor was *not* reaching for in a greater proportion of false belief trials than true belief trials. Precisely what or how much apes understood is still unclear. They appeared to recognize that something was going on in the false belief trials that should make them unsure of how to respond. It is possible that their looking responses were indicative of uncertain or implicit false belief understanding. It is also possible that the deceptive attitude of the experimenter as she performed the switch made them suspect something underhand was happening, without knowing exactly what that something might be. Or, granting the apes a little more understanding, they may have recognized that the competitor had to be ignorant of the reward's location because she had not witnessed the switch, yet she reached confidently as if she knew its location. The apes' uncertainty about what to do in these circumstances may have caused them to vacillate between options and look at both containers.

Another important issue is that the differential looking found in version A was not replicated in version B, in which the competitor had her back turned during the switch. Conceivably, the competitor's leaving the room in version A could have brought the apes closer to understanding that she could not have possibly known about the location switch, and therefore must believe the food to still be in its original location. Yet, it is curious then that children had an easier time recognizing the competitor's false belief in version B than in version A. Another possibility is that apes, by the time they received version B of the procedure, had become tired or bored with the task, or confused by

the repeated alternating back and forth between blocks of true belief and false belief trials.

In short, the ape looking results are admittedly fragile and open to a number of interpretations (although it should be noted that the latter issue also applies to many looking studies with children; e.g. see Perner & Ruffman, 2005). Given that they were based on a reduced data set, on face orientation rather than actual eye direction, and were not replicated in version B, we recommend that they be interpreted very cautiously. Nevertheless, they are important in pointing to a possible fruitful direction for future research into apes' understanding of the mind. It would be premature at this stage to do more than speculate on what the differential looking results might mean. It is important to first see if the effect replicates across various experimental procedures and different groups of apes. Given the desirability of determining whether false belief understanding on any level is a uniquely human capacity, researchers may be interested in undertaking studies to resolve the issue.

Apes' performance aside, some unexpected results for the children warrant discussion. For example, why was children's false belief performance in version A of the procedure weaker than in version B? One possibility is that children found the competitor's false belief more salient in version B, in which they could see the deceptive act carried out, literally, behind the competitor's back. It may also have been easier for children to stay focused on the task and the switching of the containers without the distraction of the competitor leaving the room, as she did in version A. Additionally, children received far more pretest trials in version A than in version B, and so they may have been losing interest in the task by the time the experimenter administered the test trials. Informal observations of children's behavior during testing suggested that these factors might have very well been an issue in version A. Another possibility is that in version B, but not version A, participants received pretest invisible displacement control trials, and these may have primed children to pay more attention to the deception during the test trials. Finally, we should note that while competitive tasks might be optimal for apes this does not necessarily make them optimal for children (which was not, in any case, our goal for this study). Nevertheless, children did well in version B, so the competitive nature of the task is an unlikely explanation for their poor performance in version A.

Another issue is that while children as a group performed similarly in both our false belief test and the standard verbal Sally-Anne test, there was little correspondence in individual performance across the two tests. The fact that some individual children found our nonverbal task easier than the Sally-Anne test, and vice versa, is not so surprising, however, as the tasks differed greatly on many dimensions. In the Sally-Anne test participants passively observed the sequence of events, did not compete, saw the item as it was moved (and so knew directly its location), had to predict the protagonist's future action, and were not immediately rewarded for answering correctly. In our nonverbal test, participants actively participated in a competition, saw only the containers but not the item being moved, had to recognize that the competitor's current actions were incorrect, and won a reward if they responded correctly. Some combination of these factors could have very conceivably led to differences in individual performance across tests.

While some researchers have found positive correlations in performance on different false belief tasks (e.g. Call & Tomasello, 1999; Carlson *et al.*, 2005), inconsistencies in individual performance – even across standard verbal tasks or on the same task administered at two different times – are not uncommon in the literature (e.g. Charman & Campbell, 1997; Holmes, Black & Miller, 1996; Mayes, Klin, Tercyak, Cicchetti & Cohen, 1996; Naito, 2003). And although Call and Tomasello (1999) found a positive correlation between children's performance on their nonverbal task and a verbal task, they did not partial out the possible confounding effects of age. In addition, the verbal task they administered was a live enactment of a situation highly similar to their nonverbal task, including the same actors and the same stimuli. Finally, in some studies in which different false belief tests have been administered to the same participants, correlations between tests have neglected to be reported at all. The issue of individual variation in performance across false belief procedures clearly demands closer attention, and also speaks to the importance of not relying on any one task to determine false belief competence in individuals.

Conclusions

While evidence continues to mount that false belief understanding is an exclusively human capacity, it is still too early to conclude this with certainty. A decade ago it was not clear that apes had any mentalistic understanding at all (Povinelli & Eddy, 1996; Tomasello & Call, 1997). Since then, innovative paradigms have provided evidence that apes understand what others intend (e.g. Buttelmann, Carpenter, Call & Tomasello, 2007; Call *et al.*, 2004), what others can see (e.g. Hare *et al.*, 2000; Hare *et al.*, 2006; Melis *et al.*, 2006), and what others know based on visual access to past events (e.g. Hare *et al.*, 2001). Similarly, for children, the age of emergence of false belief understanding has traditionally been set at somewhere between 4 and 5 years old, and still is by many researchers (see Wellman, Cross & Watson, 2001). But new, creative ways of asking much younger children what they understand has seriously challenged this view. One method has been to measure looking behavior, on the assumption that it may reveal understanding not reflected in more active responses (e.g. Clements & Perner, 1994; Garnham & Ruffman, 2001; Onishi & Baillargeon, 2005; Southgate *et al.*, 2007; Surian *et al.*, 2007). Our looking results for the apes

suggest that there might be some value in pursuing this approach with nonhuman participants as well.

Another crucial step has been the creation of nonverbal tests of false belief understanding that use active response measures (e.g. Call & Tomasello, 1999). Nonverbal tests provide a useful complement to standard false belief tasks. An aggregate approach to testing false belief is desirable, as any one task on its own may not give an entirely accurate picture, especially when making judgments about individuals' false belief competence (Hughes, Adlam, Happé, Jackson, Taylor & Caspi, 2000). The nonverbal test we created for this study (particularly version B, which yielded unambiguously positive results from children) provides a useful resource for those who wish to test animals, pre-verbal children or other special populations, or to combine multiple measures of false belief understanding.

Acknowledgements

We would like to thank Anja Gampe, Antje Girndt, Eileen Graf, Katharina Hamann, Dana Langner, Jana Reifegerste, Elena Rossi and Franziska Zemke for help with data collection and coding. We are also grateful to Juan Carlos Gómez and anonymous reviewers for helpful comments on an earlier draft of this paper.

References

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

Boysen, S.T., & Berntson, G.G. (1995). Responses to quantity: perceptual versus cognitive mechanisms in chimpanzees (*Pan troglodytes*). *Journal of Experimental Psychology: Animal Behavior Processes*, **21**, 82–86.

Boysen, S.T., Berntson, G.G., Hannan, M.B., & Cacioppo, J.T. (1996). Quantity-based interference and symbolic representations in chimpanzees (*Pan troglodytes*). *Journal of Experimental Psychology: Animal Behavior Processes*, **22**, 76–86.

Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition*, **10**, 439–448.

Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science*, **10**, F31–F38.

Call, J., Agnetta, B., & Tomasello, M. (2000). Cues that chimpanzees do and do not use to find hidden objects. *Animal Cognition*, **3**, 23–34.

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. *Developmental Science*, **7**, 488–498.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development*, **70**, 381–395.

Carlson, S.M., Wong, A., Lemke, M., & Cosser, C. (2005). Gesture as a window on children's beginning understanding of false belief. *Child Development*, **76**, 73–86.

Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology*, **20**, 393–420.

Charman, T., & Campbell, A. (1997). Reliability of theory of mind task performance by individuals with a learning disability: a research note. *Journal of Child Psychology and Psychiatry*, **38**, 725–730.

Clements, W.A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, **9**, 377–395.

Flombaum, J.I., & Santos, L.R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, **15**, 447–452.

Garnham, W.A., & Ruffman, T. (2001). Doesn't see, doesn't know: is anticipatory looking really related to understanding of belief? *Developmental Science*, **4**, 94–100.

Gómez, J.C. (2004). *Apes, monkeys, children, and the growth of the mind*. Cambridge, MA: Harvard University Press.

Gómez, J.C., & Teixidor, P. (1992). Theory of mind in an orangutan: a nonverbal test of false belief appreciation? In *XIV Congress of the International Primatological Society* (August), Strasbourg.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, **59**, 771–785.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, **61**, 139–151.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, **101**, 495–514.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative tasks. *Animal Behaviour*, **68**, 571–581.

Holmes, H.A., Black, C., & Miller, S.A. (1996). A cross-task comparison of false belief understanding in a head start population. *Journal of Experimental Child Psychology*, **63**, 263–285.

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test–retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry*, **41**, 483–490.

Karin-D'Arcy, M.R., & Povinelli, D.J. (2002). Do chimpanzees know what each other see? A closer look. *International Journal of Comparative Psychology*, **15**, 21–54.

Lohmann, H., Carpenter, M., & Call, J. (2005). Guessing versus choosing – and seeing versus believing – in false belief tasks. *British Journal of Developmental Psychology*, **23**, 451–469.

Mayes, L.C., Klin, A., Tercyak, K.P. Jr., Cicchetti, D.V., & Cohen, D.J. (1996). Test–retest reliability for false-belief tasks. *Journal of Child Psychology and Psychiatry*, **37**, 313–319.

Melis, A.P., Call, J., & Tomasello, M. (2006). Chimpanzees (*Pan troglodytes*) conceal visual and auditory information from others. *Journal of Comparative Psychology*, **120**, 154–162.

Myowa-Yamakoshi, M., Yamaguchi, M.K., Tomonaga, M., Tanaka, M., & Matsuzawa, T. (2005). Development of face recognition in infant chimpanzees (*Pan troglodytes*). *Cognitive Development*, **20**, 49–63.

Naito, M. (2003). The relationship between theory of mind and episodic memory: evidence for the development of autonoetic consciousness. *Journal of Experimental Child Psychology*, **85**, 312–336.

O'Connell, S., & Dunbar, R.I.M. (2003). A test for comprehension of false belief in chimpanzees. *Evolution and Cognition*, **9** (2), 131–140.

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

Perner, J., Leekam, S.R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: the case for a conceptual deficit. *British Journal of Developmental Psychology*, **5**, 125–137.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: how deep? *Science*, **308** (5719), 214–216.

Povinelli, D.J., & Eddy, T.J. (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, **61** (3, Serial No. 247).

Povinelli, D.J., Rulf, A.B., & Bierschwale, D.T. (1994). Absence of knowledge attribution and self-recognition in young chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, **108**, 74–80.

Reaux, J.E., Theall, L.A., & Povinelli, D.J. (1999). A longitudinal investigation of chimpanzees' understanding of visual perception. *Child Development*, **70**, 275–290.

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

Santos, L.R., Seelig, D., & Hauser, M.D. (2006). Cotton-top tamarins' (*Saguinus oedipus*) expectations about occluded objects: a dissociation between looking and reaching tasks. *Infancy*, **9**, 147–171.

Sapp, F., Lee, K., & Muir, D. (2000). Three-year-olds' difficulty with the appearance–reality distinction: is it real or is it apparent? *Developmental Psychology*, **36**, 547–560.

Smith, J.D., Shields, W.E., & Washburn, D.A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences*, **26**, 317–339.

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

Suda, C., & Call, J. (2006). What does an intermediate success rate mean? An analysis of a Piagetian liquid conservation task in the great apes. *Cognition*, **99**, 53–71.

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

Tomasello, M., & Call, J. (1997). *Primate cognition*. Oxford: Oxford University Press.

Uller, C. (2004). Disposition to recognize goals in infant chimpanzees (*Pan troglodytes*). *Animal Cognition*, **7**, 154–161.

Vick, S.-J., & Anderson, J.R. (2003). Use of human visual attention cues by olive baboons (*Papio anubis*) in a competitive task. *Journal of Comparative Psychology*, **117**, 209–216.

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

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**, 103–128.

Received: 8 August 2007
Accepted: 7 May 2008