Routledge
Taylor & Francis Group

Journal of Cognition and Development

ISSN: 1524-8372 (Print) 1532-7647 (Online) Journal homepage: https://www.tandfonline.com/loi/hjcd20

Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors

Hannes Rakoczy & Nese Oktay-Gür

To cite this article: Hannes Rakoczy & Nese Oktay-Gür (2020): Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors, Journal of Cognition and Development, DOI: 10.1080/15248372.2019.1709467

To link to this article: https://doi.org/10.1080/15248372.2019.1709467

JOURNAL OF COGNITION AND DEVELOPMENT
https://doi.org/10.1080/15248372.2019.1709467

Routledge
Taylor & Francis Group

EMPIRICAL ARTICLE

Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors

Hannes Rakoczy and Nese Oktay-Gür

University of Göttingen, Germany

ABSTRACT

When do children acquire a meta-representational Theory of Mind? False Belief (FB) tasks have become the litmus test to answer this question. In such tasks, subjects must ascribe a non-veridical belief to another agent and predict/explain her actions accordingly. Empirically, children pass explicit verbal versions of FB tasks from around age 4. The standard interpretation of this finding is that children at this age have acquired a solid capacity for meta-representation. New research with true belief (TB) control tasks, however, presents a puzzling phenomenon: While 3-year-olds pass these tasks but fail FB tasks, children from age 4 begin to show the reverse performance (passing FB but failing TB). Competence deficit accounts claim that these findings jeopardize the standard interpretation; they show that children may use simple heuristics rather than true meta-representation and that the original FB findings may thus have been false positives. Pragmatic performance limitation accounts, in contrast, claim that these findings do not document any conceptual limitations, but merely reflect children's confusion in light of the task pragmatics. In the present study, the two accounts were tested against each other in seven experiments with 4- to 7-year-old children. Pragmatic tasks factors of TB tests were systematically modified. Results show that children's difficulty with TB tasks indeed disappeared after some such modifications. This clearly speaks against competence limitation accounts and corroborates the standard interpretation of FB and related Theory of Mind tasks.

Theory of Mind (ToM), the capacity to ascribe subjective mental states such as belief, desires and intentions to others and oneself, is fundamental to human nature and sociality. At the conceptual heart of ToM lies meta-representation, that is, representing that others or oneself represent the world in certain ways that may or may not be accurate or fulfilled. The litmus tests for ToM, therefore, are false belief (FB) tasks in which participants are required to predict or explain how another agent will act on the basis of a mistaken belief (Wimmer & Perner, 1983). Decades of research with explicit verbal versions of such tasks have shown that children begin to master those tasks from around age 4 (Wellman, Cross, & Watson, 2001). The conclusion standardly drawn from these findings is that this transition marks the onset of robust, explicit meta-representation (Perner, 1991).

CONTACT Hannes Rakoczy hrakocz@gwdg.de Institute of Psychology, University of Göttingen, Waldweg 26, Göttingen 37073, Germany

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hjcd.

An empirical puzzle (from true belief tasks)

However, this conclusion has recently come under dispute both on theoretical and on empirical grounds. Theoretically, a number of alternative approaches to the standard picture have argued that children at age 4 and beyond may appear to use meta-representation but actually revert to much simpler heuristics[1]. Empirically, seemingly paradoxical findings from true belief (TB) control tasks appear to speak for these more skeptical approaches. In a typical FB task, a protagonist sees an object O being placed in box 1; O is then transferred to box 2 in her absence and the test question is either "Where will she look for O upon her return?" or "Where does she think O is?" (results from the two kinds of questions largely converge; Wimmer & Perner, 1983). TB control tasks are structurally analogous with the only difference that the protagonist also witnesses the object transfer and thus has a veridical belief. Originally, TB tasks were designed to ensure that younger children who still fail FB tasks do not do so because of general task demands related to the kinds of verbal vignettes used. Across a large body of studies, it has indeed been found that 3-year-olds typically pass TB but fail FB tasks (Wellman et al., 2001). Since their raison d'être has been to serve as control tasks, TB tests have hardly been systematically administered to older children until very recently. The few recent studies that did administer TB tasks to wider age ranges have uncovered surprising and paradoxical findings: Once children begin to master FB tasks, they begin to fail TB tasks. Children from age 4 to 7 or even beyond have been found to systematically answer TB test questions incorrectly (while performing close to ceiling on FB tasks). Only from age 8 to 10 do children then solve both FB and TB tasks. The development of TB performance thus constitutes a U-shaped curve. On an individual level, performance in FB and TB tasks is strongly negatively correlated until age 10 – initially, because children pass TB and fail FB, then because of the reverse pattern (Fabricius, Boyer, Weimer, & Carroll, 2010; Fabricius & Khalil, 2003; Hedger & Fabricius, 2011; Oktay-Gür & Rakoczy, 2017).

How to explain the puzzle? Competence vs. performance limitations

What does this puzzling pattern of findings mean? Broadly, there are two explanatory approaches. On the one hand, the pattern of findings may indeed support more skeptical positions and suggest a conceptual competence deficit: In contrast to the widely accepted standard picture, children around age 4 do not really make use of meta-representational concepts, but merely engage in simple heuristics (Fabricius et al., 2010; Fabricius & Khalil, 2003). Most notably, children may use a "perceptual access" heuristic roughly of the following content: when an agent has full access to a situation or unfolding chain of events, she will act successfully; otherwise she will act mistakenly. This heuristic would deliver the correct answer to FB tasks. And since in many TB tasks the protagonist also leaves the scene at some point and thus may be considered to lack full perceptual access to the unfolding events[2], this heuristic will deliver the wrong answer to TB questions.

On the other hand, however, the puzzling findings of TB failure in older children could simply reflect pragmatic performance (rather than competence) problems (Oktay-Gür & Rakoczy, 2017). In light of their triviality, TB questions may be confusing for older children. Upon closer inspection, TB test questions appear to combine a number of factors that may make them particularly odd in terms of pragmatics. First, they are *academic test questions* – questions asked not for genuine epistemic interest, in order to get new information, but asked

in order to test whether the interlocutor knows the answer that one knows perfectly well oneself. And it has long been suggested that this special question format may be difficult to understand for young children (e.g., Siegal, 1999). Second, they are highly trivial: Here is an agent who in clear daylight sees an object go to box 1, then to box 2, and now where will she look for it? Why would one ask a 6-year-olds child such things? Children with growing pragmatic awareness may wonder about a potential hidden agenda on the part of the speaker that they have been missing ("Why is she asking this question?"), and then conclude that the correct answer cannot be the obvious one ("She would not ask if this (the obvious one) were the answer"). Third, the test questions refer to a rational agent and her perspectives and actions. Typically, mental state talk about subjective perspectives and actions has a specific pragmatic point: Normally, we wonder and talk about subjective representation when it is relevant; that is, when there is the possibility of mis-representation, or at least of diverging representations between different agents. In the TB task, since no plausible possibility of mis-representation has been introduced, the question where the protagonist believes the object to be strikes one as highly confusing. Imagine, to illustrate the point, a situation in which three people A, B and C see a cat right in front of them, and A asks B "What does C believe this is?" Without a relevant and plausible context, B will probably be confused by this question. Why, she will wonder, is A asking such a question? Is this just a replica robot cat? Is C short-sighted and forgot her glasses?

It is thus possible that all of these factors together – in the form of trivial test questions with reference to subjective representation and action–make TB tasks pragmatically confusing for children in an intermediate stage and thus explain the U-shaped curve: in the intermediate stage from age 4 to 7 or so children (in contrast to 3-year-olds) have already developed basic meta-representational capacities that lay the foundation for pragmatic sensitivity; but in this intermediate stage they have not yet developed the higher-order, open-ended pragmatic flexibility of adults that can make sense of almost any speech act with the right kinds of additional premises (see the cat/short-sighted example above).

Now, which of the two kinds of accounts is correct? Much of the existing evidence (U-shaped curve in TB performance; negative correlations between FB and TB until age 8–10) is equally compatible with both kinds of accounts. Crucially, however, the accounts make radically diverging predictions regarding the stability and malleability of these patterns of findings. By their nature, competence deficit accounts claim that the patterns reflect deep conceptual competence limitations and should thus be stable across variations in irrelevant task characteristics. Pragmatic performance limitation accounts, in contrast, predict that the puzzling response patterns (TB failure, negative TB-FB correlations) can be made to disappear at least under some circumstances by modifying the pragmatic properties of the tasks.

First steps toward systematically testing competence and performance limitation accounts against each other have recently been undertaken in a study by Oktay-Gür and Rakoczy (2017, Study 3). This study showed that TB failure and negative TB-FB correlations disappear once the task is made less trivial and thus constitutes preliminary evidence for pragmatic performance limitation accounts. In the modified task, the triviality of the TB task was reduced by adding a second protagonist who had a false belief. While making the task generally much more complex (in terms of information to be tracked and thus working memory demands), the introduction of the second protagonist's false belief raises the possibility of mis-representation and thus, pragmatically, makes the question about

a true belief less trivial and confusing. Indeed, this modification resulted in proficient performance in 4- to 6-year-olds both in FB and in TB tasks, with now positive FB-TB correlations.

The present studies: rationale and overview

The rationale of the present set of studies is to test competence and performance limitation accounts against each other more comprehensively. In order to do so, potential pragmatic performance factors will be manipulated and modified more systematically. The results of such modifications speak to two questions. First, which kind of account is correct? To address this question, data on the stability vs. malleability of children's failure on TB tasks will be crucial. Since competence limitation accounts assume that such failure reflects a principled conceptual deficit, no superficial modifications of TB tasks should alter performance at all. Performance limitations accounts, in contrast, claim that children fail TB tasks not due to inherent conceptual limitations, but due to extraneous pragmatic performance factors (whatever those make turn out to be exactly). When some such factors are modified, children should thus, at least under some circumstances, perform successfully. That is, *any* kind of evidence that children can pass both FB and TB tasks once they are suitably modified is incompatible with competence limitation and speaks for pragmatic performance deficit accounts. Second, which are the crucial pragmatic performance factors, individually and/or in combination? Results from different types of modifications will provide preliminary evidence to address this second question.

Across seven studies with 4- to 7-year-old children, various such modifications were implemented regarding the potentially relevant pragmatic factors (the target dependent measures involve (i) trivial, (ii) academic test questions, (iii) about an agent's subjective perspective; see Table 1, for an overview of these new tasks and how they relate to the previous Study 3 by Oktay-Gür & Rakoczy, 2017). Study 1 removed trivial test questions, indeed any kind of questions altogether (thus addressing (i) and (ii)), by implementing a purely non-verbal FB/TB task based on a test developed for non-human primates (Call & Tomasello, 1999). Since this task has been shown to be more demanding generally (in terms of working memory etc.), so that children pass the FB version significantly later than in standard tasks, children of a wider age range (4 to 7 years) were tested and contrasted as a function of age.

Table 1. Overview of the new tasks (and of the previous task in Oktay-Gür and Rakoczy (2017, Study 3) with regard to the modifications that address potential pragmatic performance factors in Studies 1–5.

	Potential pragmatic performance factors			
	Trivial question	Academic test question	About an agent's perspective	Description of the (new) task
Oktay-Gür & Rakoczy, 2017 (Study 3)	-	✓	✓	Combined FB/TB scenario in which one agent's TB contrasts with another agent's FB
Study 1	-	-	✓	Non-verbal FB/TB task
Study 2	✓	✓	-	"False"/"True" Photo Task
Studies 3a-3c	✓	-	✓	Standard FB/TB tasks with test question asked by naïve agent
Study 4	-	✓	✓	FB/TB task where protagonist changes her belief
Study 5	-	✓	✓	FB/TB tasks in which E announces that some "baby questions" will be asked

"□"/"-" signify that the new task does/does not involve the potential pragmatic performance factor in question

Results revealed that children's performance in FB tasks increased with age (with only 7-year-olds performing above chance) but was equally proficient in TB across all age groups. In this non-verbal task, there was thus no evidence for any TB deficits.

Studies 2–5 investigated whether, and under which conditions, the TB deficits previously found in children who have come to master FB tasks can be alleviated or made to disappear in standard verbal tasks. Since previous studies have consistently documented that children between 4 and 6 show the strongest effects (begin to pass FB and to fail TB, with strong negative FB-TB correlations), we focused on this age range. Study 2, addressing (iii), implemented a task that was structurally analogous to FB/TB tasks with the crucial difference that there were no questions about an agent's subjective perspective. Instead, questions referred to a camera that produced veridical vs. outdated photos (based on the "False" Photo Task developed by Zaitchik, 1990). Children performed similarly on False Belief and False Photo tasks (passing both), but very differently on True Belief vs. True Photo tasks (failing the former while passing the latter). Studies 3a-c addressed (ii) and modified the nature of the question and attempted to structure them more like real rather than like academic test questions. This was done by having ignorant speakers ask. Despite these manipulations, children passed FB but failed TB tasks. Studies 4 and 5, finally, addressed (i) and manipulated the triviality of the test question. In Study 4, the protagonist's belief changed over time so that asking about a true belief of hers at one point in time appeared less trivial (given her contrasting false belief at another time). Despite this manipulation, children passed FB and failed TB tasks. In Study 5, before asking the test question the experimenter explained to children that some of the questions to be asked were designed for much younger children and would thus be very easy. With this manipulation regarding the perceived triviality of the test question, children (under some conditions) now passed both FB and TB tasks.

Study 1

If children's difficulty with TB tasks was based on conceptual competence limitations, then analogous difficulties should be observable in various task formats, including non-verbal ones. In contrast, if the puzzling TB findings documented so far reflect performance factors related to the pragmatics of the test questions, then once the test questions are removed, the difficulties should disappear. Study 1, therefore, implemented non-verbal FB/TB tests based on a task originally designed for non-human primates (Call & Tomasello, 1999) and previously used with children (Fizke, Barthel, Peters, & Rakoczy, 2014), neuropsychological patients (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005) and elderly adults (Bailey & Henry, 2008).

Method

Participants

Seventy-five 4- to 7-year-olds (4-year-olds, N = 20, 49–59 months, M = 55; 5-year-olds, N = 20, 62–73 months, M = 66; 6-year-olds, N = 20, 74–83 months, M = 77; 7-year olds, N = 15, 86–95 months, M = 90) were included in the final sample. One additional 6-year-old was tested but excluded from data analyses because she was uncooperative. Participants in this and all subsequent studies were recruited from a databank of children

whose parents had previously given consent to experimental participation. Children came from mixed socio-economic backgrounds and were native German speakers. Testing was conducted by a female experimenter (E) in the laboratory.

Design and procedure

Verbal ability. In this and all subsequent studies, children were given a vocabulary test at the beginning of the session (the vocabulary subscale of the Kaufman Assessment Battery for Children; Kaufman & Kaufman, 1999).

FB/TB tasks. Children received two FB and two TB trials in counterbalanced order. The basic structure of the procedure, modeled after Call and Tomasello (1999) is depicted in Figure 1. Before the test trials, children underwent a long introduction phase, also modeled after Call and Tomasello (1999) to make sure they understood the basic structure of the whole procedure including invisible hiding etc. (for details, see Appendix A).

The basic logic of the task is the following: Children see that a sticker is hidden in one of two perceptually indistinguishable boxes. Another agent (introduced beforehand as a very helpful communicator who supplies relevant information) can see in which box the sticker goes, but the child cannot. The two boxes are then swapped in the presence (TB) or absence (FB) of the agent. The agent then indicates toward the child where (she thinks) the object is, and the child's task is to find the object. Children must thus reason as follows: "She knows where it is and so it is where she indicates" (TB)/"She has not seen the swap, so holds a false belief, so it is not where she indicates but in the other box" (FB). The correct answers were thus to search in the box indicated by the communicator (TB) or in the other box (FB), respectively.

Results

The mean sums of trials answered correctly as a function of conditions and age groups are depicted in Figure 2. Since preliminary analyses did not reveal any order effects, order was not included as a factor in the main analyses. A 2 (belief: FB/TB) x 4 (age groups: 4-/5-/6- and 7-year olds) ANOVA revealed a main effect of belief ($F(1,71) = 10.01$, $p < .01$, $\eta_p^2 = .12$), a main effect of age group ($F(3,71) = 5.96$, $p < .01$, $\eta_p^2 = .20$) and an interaction ($F(13,71) = 4.69$, $p < .01$, $\eta_p^2 = .16$). Therefore, we analyzed children's performance separately for the different age groups. All age groups performed significantly above chance on TB (4-year-olds, $t(19) = 4.95$, $p < .001$, $d = 1.11$; 5-year olds, $t(19) = 2.18$, $p < .05$, $d = .49$; 6-year-olds, $t(19) = 3.33$, $p < .01$, $d = .74$; 7-year olds, $t(14) = 4.18$, $p < .01$, $d = 1.09$). In contrast, on FB only 7-year olds performed significantly above chance ($t(14) = 6.21$, $p < .001$, $d = 1.60$), while 4-year olds performed significantly below chance ($t(19) = -3.25$, $p < .01$, $d = -.74$) and 5- and 6-year-olds performed at chance (5-year olds, $t(19) = 1.83$, $p = .08$; 6-year olds, $t(19) = .25$, $p = .80$).

Correlations of FB and TB task performance as a function of age groups are depicted in Table 2. FB and TB are negatively correlated in some of the younger age groups (who pass TB while failing FB), but positively correlated in the 7-year-olds who pass both FB and TB. This developmental pattern is markedly different from the pattern found in standard verbal tasks: in those tasks FB and TB are negatively correlated in 3-year-olds who pass TB while failing

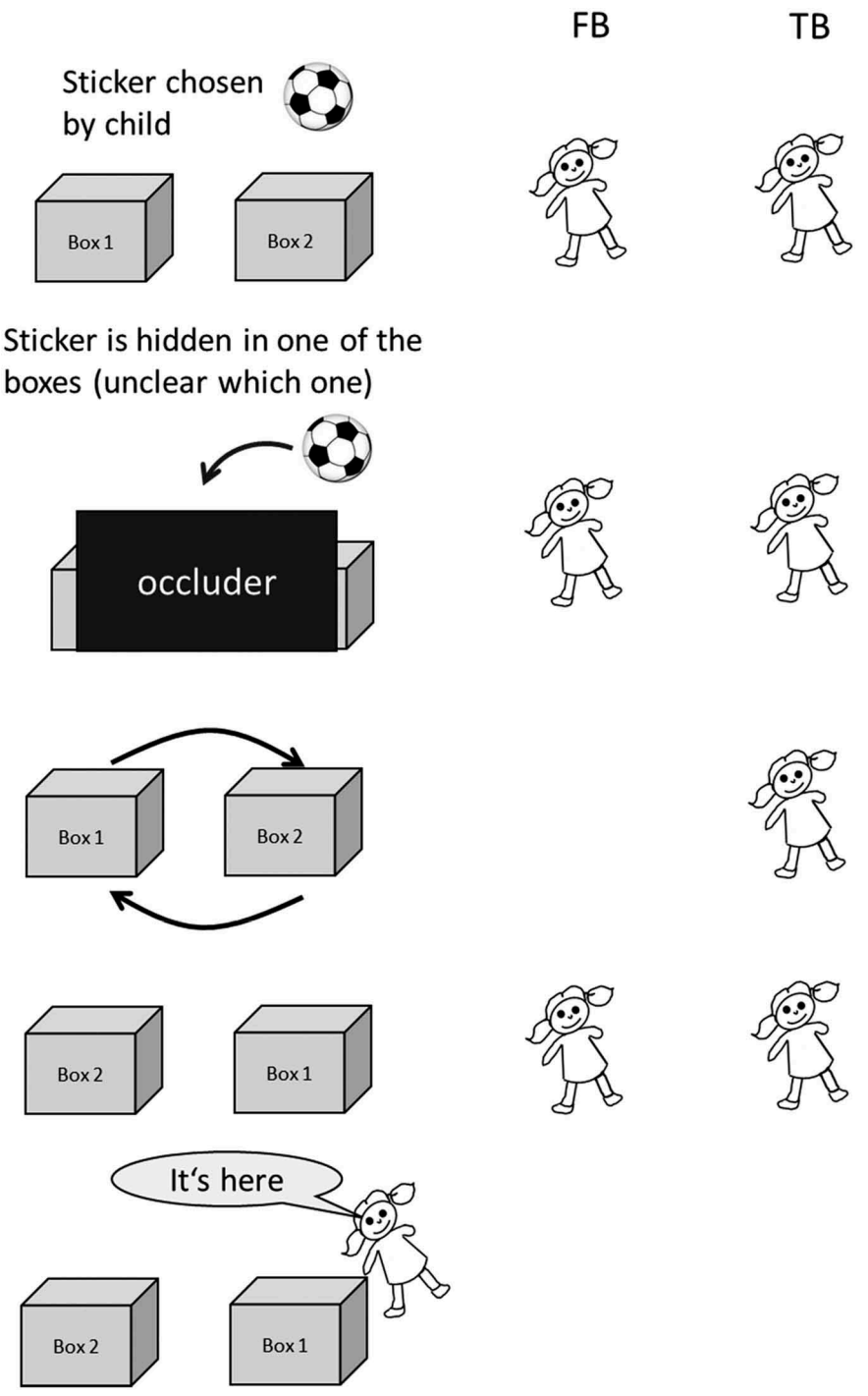

Figure 1. Schematic procedure of the FB/TB conditions of Study 1. Note that the two boxes are qualitatively identical and perceptually indistinguishable.

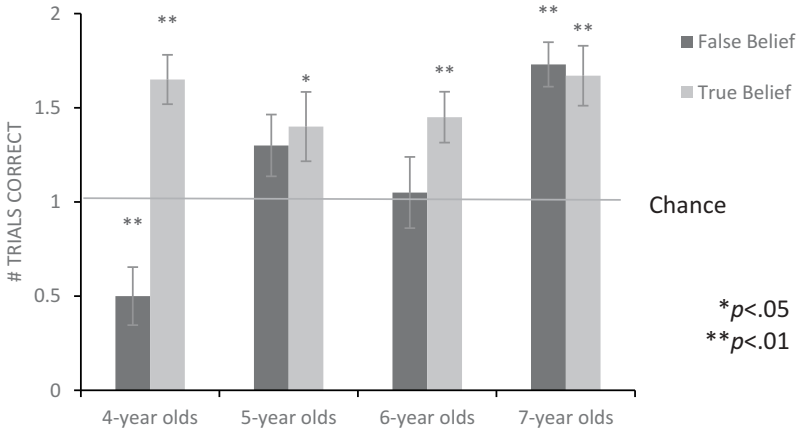

Figure 2. Mean sum of trials answered correctly as a function of age group and condition in Study 1. (*Note*: Error bars here and in all studies indicate S.E.).

Table 2. Raw and partial correlations (controlling for age and verbal ability) between False and True Belief in Study 1.

4-year-olds	5-year-olds	6-year-olds	7-year-olds
−.20	−.21	−.54*	.67**
(−.20)	(−.18)	(−.55*)	(.67**)[3]

*$p < .05$, **$p < .01$

FB, also negatively correlated in 4- to 7-year-olds who show the reverse pattern, and only become positively correlated in 10-year-olds who pass both (Oktay-Gür & Rakoczy, 2017).

Discussion

The main findings of the present study were the following. First, in terms of absolute proficiency, the present task was generally more difficult than standard verbal FB tasks. This, however, is both unsurprising given the inferential complexity of the task, and in line with previous findings in the literature (Call & Tomasello, 1999). Second and more importantly, in terms of relative performance patterns, the present task produced patterns of results very different from those in standard verbal FB/TB tasks: There was no evidence for the kind of U-shaped curve in TB performance in standard tasks (such that younger and older children pass while children in between fail). Third, the patterns of correlations were different (once children as a group solved the FB tasks at above chance levels, FB and TB were positively correlated). Taken together, these findings suggest that the puzzling pattern of findings in standard TB tasks seems peculiar to the verbal task format.

Study 2

Study 1 established that the peculiar TB performance pattern disappears once the test questions is removed altogether in a non-verbal format. Studies 2–5 investigate whether, and under which conditions, the TB pattern can be modified or made to disappear in

verbal tasks. Study 2 investigated whether the puzzling TB findings have indeed to do with the fact that verbal TB tasks involve trivial test questions *about an agent's perspective*. Such questions of subjective perspective usually have a pragmatic point that is missing in TB cases: they are normally only asked in contexts in which the possibility of misrepresentation or of diverging representation between individuals is relevant.

In order to test whether this pragmatic oddity makes TB tasks difficult, we compared them to structurally analogous tasks that also involved trivial test questions, yet with the crucial difference that these questions did not refer to an agent's perspective. To implement contrasts as minimal as possible, we made use of a task that was designed to be structurally isomorphic to FB tasks, the so-called "False Photo Task"[4] (Zaitchik, 1990), and adapted it to include both "False Photo" and "True Photo" conditions that closely matched FB and TB. In the False/True Photo tasks the same events happen as in TB/FB: an object is placed in box 1 and then transferred to box 2 before the test question is asked. The agent is replaced by a camera, and belief-formation by taking a picture, either before (FB) or after the object's change of location (see Figure 3). The logic of this study is the following. If it is the triviality of the test question as such that underlies children's poor performance in TB tasks, then across Belief and Photo tasks the same pattern should emerge: children fail the trivial TB/True Photo but pass the less trivial FB/False Photo version. In contrast, if part of the problem lies in the fact that the trivial test question is about an agent's perspective in such a way that the pragmatic point of such questions is missing, then the following pattern should emerge: children should show the usual pattern in FB and TB (passing the former while failing the latter, with negative correlations between them), but should be equally proficient in both True and False Photo tasks (potentially with positive correlations between them).

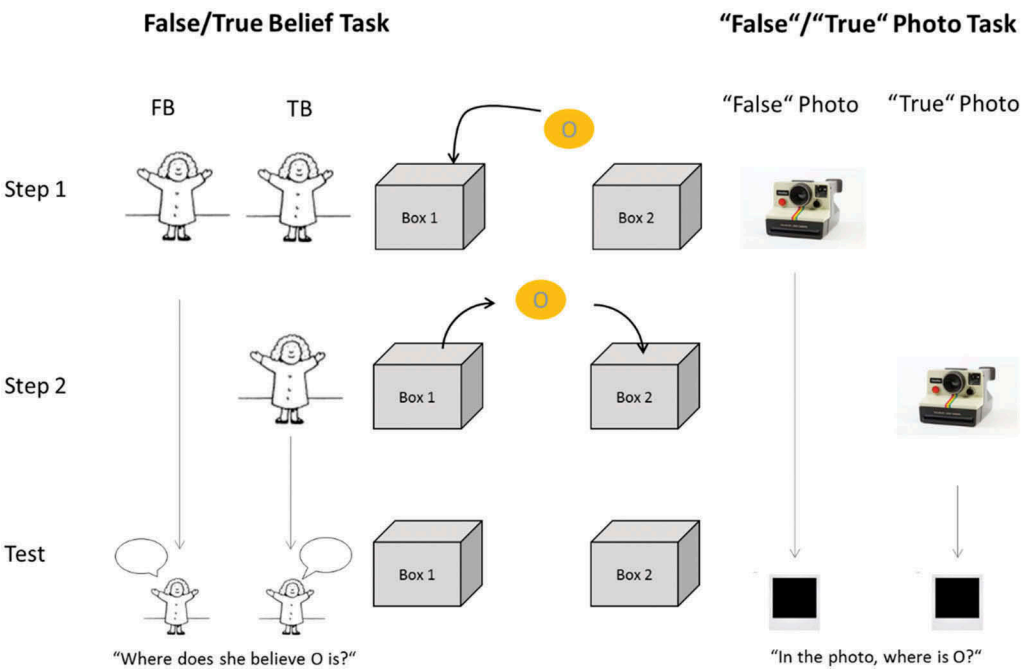

Figure 3. Schematic procedure of the tasks used in Study 2.

Method

Participants

Thirty-one 4- to 6-year-olds (49– 72 months, $M = 60$, $SD = 8$) from mixed socioeconomic background were tested. Children were recruited from a databank of children whose parents had previously given consent to experimental participation. Children were tested by a female experimenter (E) in the laboratory.

Design and procedure

The basic design was a 2 (Task: Belief/Photo) x 2 (Condition: False/True) within subjects design. Children received two trials per condition resulting in eight trials per child. The order of the tasks was counterbalanced across participants.

False/true belief task. Four trials of standard change-of-location tasks with different stimuli were administered per child, 2 in TB and 2 in FB versions (Wimmer & Perner, 1983). The protagonist and the child were introduced to an object O. The object was then placed in one of two boxes (box1) before the protagonist left. Either in her absence (FB condition) or after her return (TB condition), O was moved to the other box (box2) and the following control and test questions were asked.

- Control Question 1: Where did we put O in the beginning? [correct answer: box 1]
- Control Question 2: Where is O now? [correct answer: box 2])
- Test question: Where does the protagonist believe O is? [correct answer: box 2 (TB)/box 1 (FB)]

"False"/"true" photo task. Each child received four trials of a Photo Task (modeled after Zaitchik, 1990). The child was introduced to a digital camera and was allowed to take a picture of something (e.g. a drawing she draw in the warm-up phase before the testing). The child and E then together looked at the digital photo on the screen of the camera. In the next step, E introduced a protagonist and two rooms of her flat (two transparent boxes, one designed as a living room and one as a bedroom). The protagonist then placed an object O first in one of the rooms (opened box, placed object and closed box again). The protagonist moved the object to the other box either before (true) or after (false) a photo of the scene was made by the experimenter (see Figure 3). Then, the following control and test questions were asked:

- Control question 1: "Where was O at the beginning?" [correct answer: box 1]
- Control question 2: "Where is O now?" [correct answer: box 2])
- Test question: "In the photo, where is O?".[correct answer: box 2 ("True" photo)/box 1("False" Photo)]

Results

Children answered the control questions correctly in 87% of the FB trials, 98% of the TB Trials and in 100% of False and True Photo trials[5]. The mean sums of trials answered correctly as

a function of task and condition are depicted in Figure 4. Since preliminary analyses did not reveal any order effects, order was not included as a factor in the main analyses. A 2 (Task: Belief/Photo) x 2 (Condition: False/True) repeated measures ANOVA revealed a main effect of task ($F(1,30) = 26.06$, $p < .001$, $p^2 = .47$), a main effect of condition ($F(1,30) = 19.38$, $p < .001$, $p^2 = .39$) and an interaction of task and condition ($F(1,30) = 7.87$, $p < .01$, $p^2 = .21$). Children performed significantly above chance on FB ($t(30) = 3.72$, $p < .01$, $d = .66$), False Photo ($t(30) = 6.52$, $p < .001$, $d = 1.17$) and True Photo ($t(30) = 3.05$, $p < .01$, $d = .55$), while significantly below chance on TB ($t(30) = -3.77$, $p < .01$, $d = -.68$).

Correlations of the different tasks and conditions are depicted in Table 3. FB and TB were highly significantly negatively correlated (even corrected for age and verbal ability) while False and True Photo were positively correlated when corrected for age and verbal ability.

Discussion

The main findings of the present study were the following: when comparing structurally analogous False/True Belief and "False"/"True" Photo tasks, fundamentally different performance patterns emerged: the puzzling pattern in Belief tasks was reproduced (children passed FB and failed TB, and the two were negatively correlated); but in the

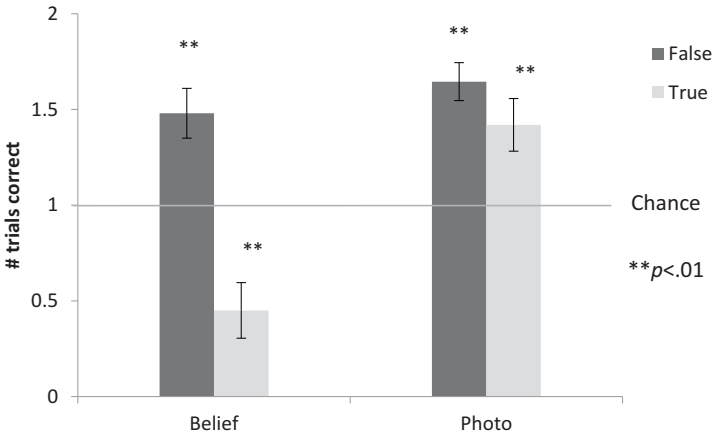

Figure 4. Mean number of trials answered correctly as a function of condition in Study 2.

Table 3. Raw and partial correlations (corrected for age and verbal ability) of the different tasks and conditions in Study 2.

	True Belief	"False" Photo	"True" Photo
False Belief	−.61**	.03	−.02
	(−.58**)	(.05)	(−.03)
True Belief		.07	.01
		(.06)	(−.10)
"False" Photo			.29
			(.37*)

*$p < .05$; **$p < .01$

Photo task, children passed both False and True conditions, and those were also positively correlated. The upshot of these findings for the question under study is thus the following: It cannot be the fact that trivial test questions are used that explains TB difficulty; but more specifically it seems to be crucial that these are trivial test questions *about an agent's perspective*.

Studies 3a–3c

Studies 1 and 2 suggest that children's peculiar difficulty with TB tasks only arises in verbal task formats and only when trivial test questions *about an agent's perspective* are asked. Next, we investigated in more fine-grained ways whether children's difficulty can also be alleviated once the questions are about an agent's perspective yet are real rather than test questions (Studies 3a–3c) or once they are made less trivial (Studies 4–5). Studies 3a–3c modified the context in which the central questions were asked in ways which were supposed to turn them into real questions (where the speaker genuinely asks for information) rather than test questions (where the speaker does not need information but just wants to know whether the child knows …). This was done by having the questions asked by a naïve human adult (Study 3a), by a "very unknowledgeable" hand-puppet (Study 3b) or by a naïve hand-puppet (Study 3c).

Study 3a

Method

Participants. Seventy-three 4- to 6-year-olds (48–84 months, $M = 66$) from mixed socio-economic background were included in the final sample. Five additional children were tested but excluded from data analyses because they were uncooperative. Children were recruited from a databank of children whose parents had previously given consent to experimental participation. Children were tested by two female experimenters (E1 and E2) in the laboratory.

Design and procedure. Two different types of FB/TB tasks were used: Standard change of location tasks in which an object changes location in the presence (TB) or absence (FB) of the protagonist (Wimmer & Perner, 1983). And modified aspectuality tasks in which an object with two identities changed location in the presence of the protagonist, with the difference between conditions being whether the agent does (TB) or does not (FB) know about both identities (after Rakoczy, Fizke, Bergfeld, & Schwarz, 2015; see below). The basic design was a 2 (Condition: Standard/Aspectuality) x 2 (Belief: FB/TB) design, with condition as a between- and belief as a within-subjects factor. Each child received two trials of each belief in a given condition resulting in four trials in total. The order of TB and FB blocks was counterbalanced across subjects.

Standard false/true belief task. The FB/TB Task used in this experiment was very similar to standard change-of-location FB tasks (Wimmer & Perner, 1983) as those used in previous studies. The crucial difference was that a naïve experimenter asked the test question who did not witness the events before (so that the target question was more like a real question asked to get information than a test question asked in order to see

whether the interlocutor knows the answer). E1 enacted the scenario (object is in box 1, then transferred to box 2) up to the protagonist's return. Upon the protagonist's return just before the experimenter could start asking control and test questions, E2 entered the room and informed E1 about an urgent phone call for her. E1 then suggested that E2 could go on to play with the child, the protagonist and the object and left the scene. Now E2 continued with the following introduction: "Hmm … You played with [the protagonist] and the [object], I see", and asked the following control and test questions.

- "Where did you put [the object] in the beginning?" (Control Question 1)
- "Where is [the object] now?" (Control Question 2)
- "Did [the protagonist] see that you moved [the object] from there over there?" (Control Question 3)
- "If [the protagonist] wants [the object], where will she look for it?" (Test Question)

E2 then continued with the next trial, and here, the roles were reversed. E2 enacted the scenario up to the protagonist's return, then E1 entered, told E2 there was a phone call for her and asked the control and test questions.

Aspectuality task. The basic logic of these tasks (closely modeled after Study 3 of Rakoczy et al., 2015) is the following. In the presence of a protagonist an object was put into a box (box 1) under aspect A [e.g. pen]. In the protagonist's presence (TB) or absence (FB) it was revealed that the object had another identity B [e.g. rattle] and it was stored in the same box again. In the presence of the protagonist the object was then transferred to box 2 under its identity B (for example, the experimenter covered the object with her hands while taking it out of its initial box, rattled with it and then moved it to the other box such that the A-identity (pen) remained invisible throughout and only the B-identity (rattle) could be heard). In both belief conditions (TB and FB) the protagonist witnessed the object's transfer. The crucial modification relative to the previous version of this task was that the crucial control and test questions were asked by a naïve experimenter who entered the room after the scenario had been acted out by the other experimenter

- Control question 1: "Where did you put [the A (e.g., pen)] in the beginning?"
- Control Question 2: "Where is [the A (e.g., pen)] now?"
- Control Question 3: "Does [the protagonist] know that [A (e.g. pen)] is also a [B (e.g., rattle)]?"
- Test question: "If [the protagonist] wants [the A (e.g., pen)], where will she look for it?")

Results and discussion

Children answered the control questions correctly in 93% of the Standard FB/TB and in 83% of the Aspectuality FB/TB trials. The mean sums of trials answered correctly as a function of conditions are depicted in Figure 5.

A 2(Belief: False/True) x 2 (Task: Standard/Aspectuality) x 2(order of FB/TB trials) mixed factors ANOVA revealed a main effect of Task ($F(1,69) = 37.79$, $p < .001$, $\eta_p^2 = .36$), a main effect of Belief ($F(1,69) = 23.86$, $p < .001$, $\eta_p^2 = .26$) and an interaction

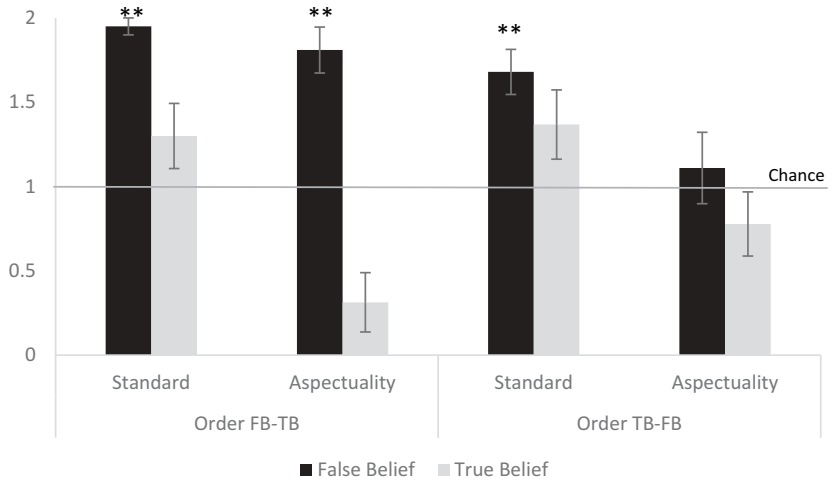

Figure 5. Mean number of FB and TB trials answered correctly as a function of condition and order in Study 3a.

belief x order ($F(1,69) = 6.86$, $p < .05$, $\eta_p^2 = .09$) such that the differences between TB and FB were more marked when FB was administered first.

Analyzing FB and TB trials separately as a function of order and task revealed that children only performed significantly across chance-level for the standard FB task (order FB-TB: $t(19) = 19.00$, $p < .001$, $d = 4.24$; order TB-FB: $t(18) = 5.12$, $p < .001$, $d = 1.17$) and for the aspectual FB task when it was administered first (order FB-TB: $t(15) = 5.98$, $p < .001$, $d = 1.5$; order TB-FB: $t(17) = 0.52$, $p = .61$), but at or below chance for TB trials. Across conditions, TB and FB tasks were significantly negatively correlated, $r = -.29$, $p < .05$ (partial correlation controlling for age and verbal ability: $r = -.34$, $p < .05$).

In the present study, there was thus no unambiguous evidence that children's difficulty with TB tasks (partly) had to do with the fact that the focal question was a test rather than a real question. Perhaps, however, the present implementation of real rather than test questions did alleviate the pragmatic confusion yet introduced other complications. In particular, having two experimenters exit and enter the room in alternating order may have created excessive memory demands and confusion.

Study 3b

Study 3b, therefore, used an alternative way of asking real (instead of test) questions, by using a naïve puppet as speaker.

Method

Participants. Thirty-four 4- to 6-year-olds (49–80 months, $M = 62$, $SD = 8$) from mixed socioeconomic background were included in the final sample. Four additional children were tested but excluded from data analyses because they broke up or due to experimental errors. Children were recruited from a databank of children whose parents had previously given consent to experimental participation. Children were tested by a female experimenter either in the laboratory or in an appropriate room in their day care.

Design and procedure. The basic design was again a 2 (Condition: Standard/Aspectuality) x 2 (Belief: FB/TB) design, with condition as a between- and belief as a within-subjects factor. Each child received two trials of each belief in a given condition resulting in four trials in total. The order of TB and FB blocks was counterbalanced across subjects.

Standard and aspectuality FB/TB tasks. The same tasks as in Study 3a were used, except for the following modifications: Before the actual protagonist was introduced, the Experimenter introduced a naïve hand puppet who "often needs some time to understand what is going on and therefore tends to ask stupid questions". Children were then encouraged to help the naïve puppet whenever she had any questions. The experimenter then introduced the actual protagonist, the boxes and the object. In the early stages of the scenarios the hand puppet repeatedly asked trivial questions with obvious answers[6]. Finally, after the change of location (witnessed (TB) or not witnessed (FB) by the protagonist) the hand puppet asked test and control questions. These were the same as in Study 3a, with one exception: In the Aspectuality task, an additional control question (Control Question 0) was asked by the naive puppet when the dual identity of the object was revealed ("Oh, wait a minute! Does the puppet know that [the X] is also a [Y]?").

Results and discussion

Children answered all control questions correctly in the 75% of the Standard FB/TB and in 73% of the Aspectuality FB/TB tasks. The mean sums of trials answered correctly as a function of conditions are depicted in Figure 6. Since preliminary analyses did not reveal any order effects, order was not included as a factor in the main analyses. A 2(Belief: False/True) x 2 (Task: Standard/Aspectuality) mixed factors ANOVA revealed a main effect of belief type ($F(1,32) = 4.937$, $p < .05$, $\eta_p^2 = .13$), no main effect of task type and no interaction effect. Analyses of each condition separately revealed that children performed below chance in the standard TB tasks ($t(15) = 2.15$, $p < .05$, $d = .54$) only, and did not differ from chance in the other conditions ($ps > .10$). Across conditions, TB and FB tasks were significantly negatively correlated, $r = -.61$, $p < .01$ (partial correlation controlling for age and verbal ability: $r = -.57$, $p < .01$).

Like Study 3a, the present study failed to find any evidence for the crucial role of test (rather than real) questions. Perhaps, however, the current procedure was not sensitive enough. Perhaps having the agent ask trivial questions dispositionally, was not the best way to convey and make salient to the child the puppet's ignorance and thus the fact that she asks genuine rather than test questions.

Study 3c

Study 3c, therefore, presents yet another attempt at implementing the facts the speaker asking the question was ignorant and thus that the question was a genuine one. Like in Study 3a, an ignorant agent was used who failed to witness the crucial steps of the scenario. Like in Study 3b, however, the agent was a hand-puppet. In this way, the problems of Study 3a (having experimenter exit/enter the room) could be avoided. In addition, we simplified and shortened the session by only using standard FB/TB tasks and skipping the aspectuality FB/TB tasks (the two kinds of tasks mostly delivered the same

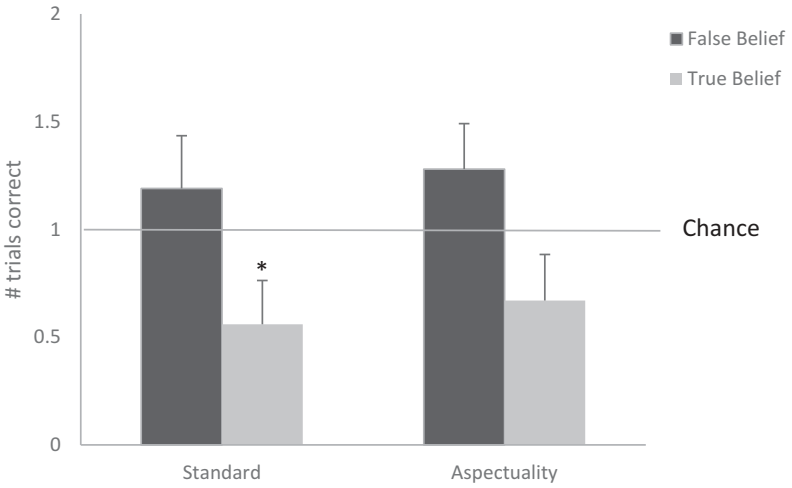

Figure 6. Mean number of trials answered correctly as a function of condition in Study 3b.

absolute performance patterns, and, most importantly, have delivered identical relative performance patterns of FB in contrast to TB conditions).

Methods

Participants. Nineteen 4- and 5-year-olds (49–71 months, M = 60) from mixed socio-economic background were included in the final sample. One additional child was tested but excluded from data analyses because she refused to cooperate. Children were recruited from a databank of children whose parents had previously given consent to experimental participation. Children were tested by a male experimenter (E) in an adequate room in their day care.

Design and procedure. In a within-subjects design, each child received two standard FB and two standard TB trials (resulting in four trials per child in total), with the order of TB and FB counterbalanced across subjects.

Standard FB/TB task. The same standard FB/TB tasks as in Studies 3a/3b was used, except for the following modifications. Before the experimenter introduced the protagonist, he announced that he brought a friend of his to join the game they were going to play, but that this friend was sleeping. Therefore, he asked the child to pay special attention to everything in order to be able to explain what was going on to the sleepy friend once he woke up. In both FB and TB, then after the change of location took place and the protagonist returned, E said that he heard his friend wake up and introduced him to the scene and the friend puppet asked the control and test questions (the same as in Studies 3a/b).

Results

In the Standard Task FB children answered Control Question 1 correctly in 90% of the trials, Control Question 2 in 95% and Control Question 3 in 75% of all trials. In the

Standard TB Task, Control Question 1 was answered correctly in 81% of all trials, Control Question 2 in 94% and Control Question 3 in 81%. The mean sums of trials answered correctly as a function of belief condition are depicted in Figure 7. Since preliminary analyses did not reveal any order effects, order was not included as a factor in the main analyses. Children performed significantly better in FB than in TB, t(19) = 2.25, $p < .05$, $d = .52$), with performance being significantly above chance in FB ($t(18) = 3.64$, $p < .01$, $d = .84$) but at chance in TB (t (18) = 0, $p = 1$). FB and TB performances were not significantly correlated ($r = -.10$, $p = .69$; controlled for age and verbal ability, $r = .02$, $p = .94$).

Discussion studies 3a-3c

Across the three Studies 3a-3c we failed to find any evidence that children's difficulty with TB tasks is (at least partly) due to the fact that they were asked test rather than genuine questions. Absence of evidence, of course, is not evidence of absence. So, we cannot at the current stage of inquiry conclude that asking test questions does not play a role. Perhaps the manipulations administered here simply failed to make the fact that genuine questions were being asked convincing or salient enough. One specific possibility is the following: The interlocutor who asked the target question ("where will the protagonist look … ?") was naïve in the sense that she had not witnessed the unfolding of the events. However, the experimental protocol mandated that she asked some control questions before the target questions, and so in some sense she could be said to be no longer naïve (because children may have assumed that the answers to the control question did convey, at least implicitly, all the relevant information to her). Future research will need to test whether alternative ways of implementing the interlocutor's naivete may succeed in modifying the perceived pragmatics of the target question as a mere test question; and whether, as a consequence, children's difficult with TB tasks can be alleviated.

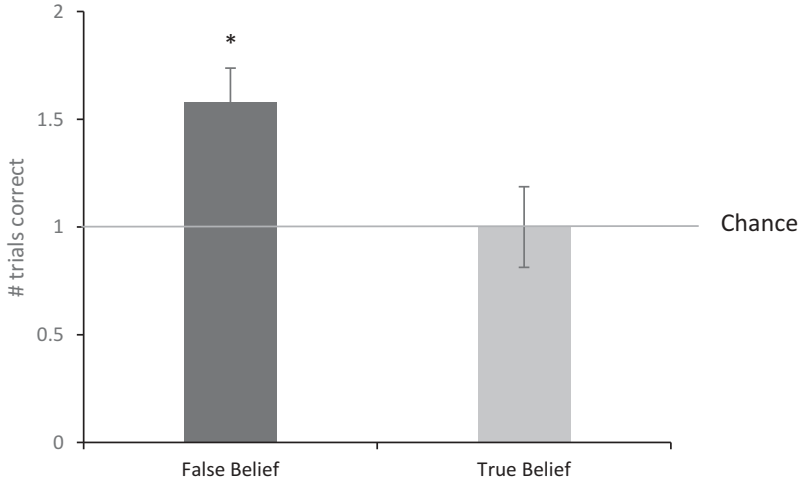

Figure 7. Mean number of trials answered correctly as a function of condition in Study 3c.

Study 4

In the last two studies, we finally tested whether children's difficulty with TB tasks was (partly) due to the triviality of the test questions. Preliminary evidence for this possibility comes from the recent finding by Oktay-Gür and Rakoczy (2017) that once another agent who holds a false belief is added to the scene children have no difficulty with TB tasks anymore – supposedly because the presence of the mistaken agent makes the question of who believes what less trivial. In Study 4, we tested whether a different yet related way of making the TB question less trivial alleviates children's difficulties with the TB task: over time, the agent in the scene changes/updates her beliefs so that she has both false and true beliefs over time and so that asking about her true belief at one point should be less trivial under some circumstances. More specifically, when (in condition "TB first") the agent at time 1 has a true belief and the child is asked about the agent's belief, and then the agent acquires a false belief at time 2, the original pattern should be reproduced (because at time 1 the possibility of mis-representation has not been raised yet). But in the reverse order (condition "FB first": FB at time 1 followed by TB at time 2) at the time the TB question is asked, the possibility of mis-representation has already been made very salient, and the TB question should thus be much less trivial.

Methods

Participants

Thirty-two 4- to 6-year-olds (48–71 months, M = 58, SD = 8) from mixed socioeconomic background were tested. Five additional children were tested but excluded from data analyses because they interrupted the procedure (N = 3) or ended the session (N = 2). Children were recruited from a databank of children whose parents had previously given consent to experimental participation. Children were tested by a female experimenter (E) either in the laboratory or an appropriate room in their day care.

Design and procedure

The basic design was a 2 (Belief Type: FB/TB) x 2 (Belief order: FB-TB/TB-FB) design, with belief order as a between- and belief as a within-subjects factor. Each child received two tasks in which a protagonist had a TB at one point and an FB at another, with two corresponding questions (resulting in four questions/data points per child).

FB/TB task. The FB/TB task used in this study was closely modeled after the belief tasks used before with the following crucial modifications (see Figure 8): TB and FB questions were combined within a given scenario that unfolded over time such that there were two transfers of the object and participants were asked about the protagonist's belief twice at separate points in the story. One of the location changes was witnessed by the protagonist, resulting in a true belief; the other one was not witnessed, resulting in a false belief. Two conditions differed in the order of true and false beliefs: In the *FB first* condition, the first steps were just like in a Standard FB task, followed by an additional step in which the objects' real location was shown before it was transferred back to its original location, all witnessed by the agent whose belief now became true. In the *TB first* condition, the first steps were just like in a standard TB tasks, followed by an additional step in which the object was transferred back to its original location, unbeknownst to the agent who thus

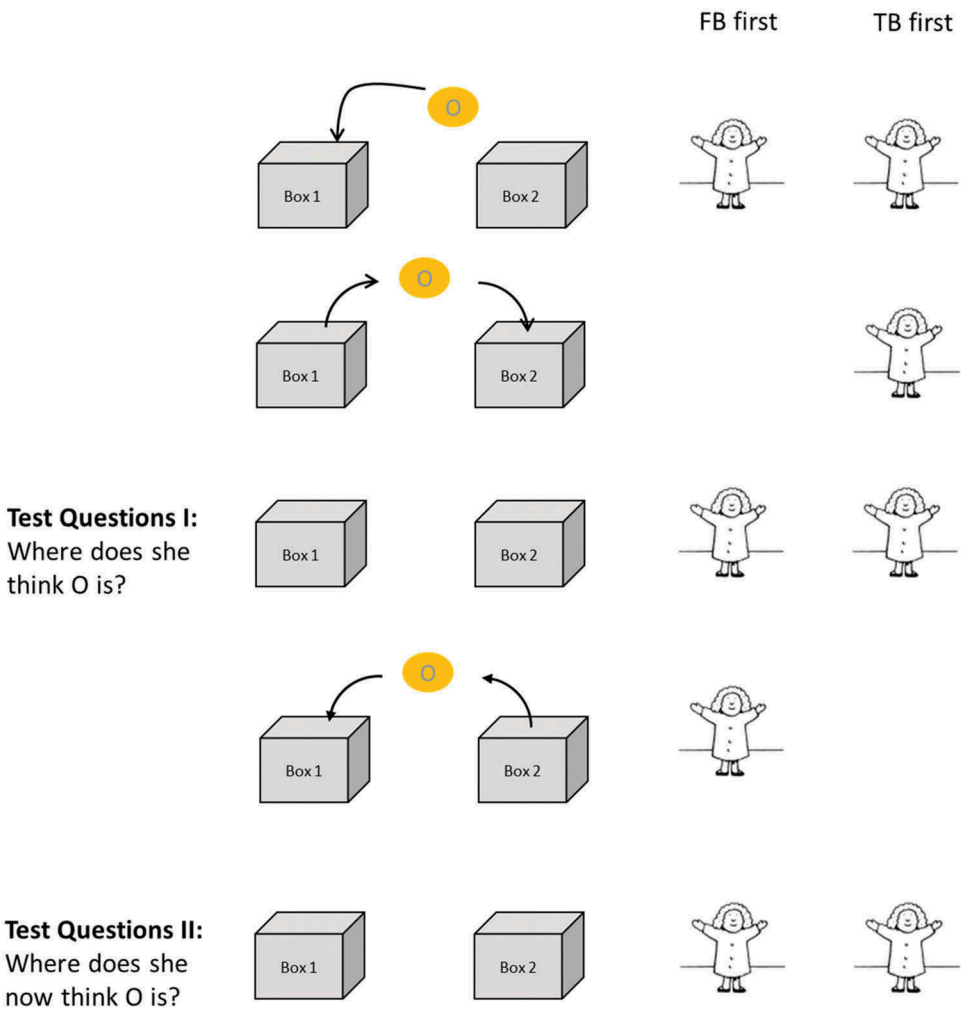

Figure 8. Schematic procedure of the different conditions in Study 4.

came to hold a false belief. In both conditions, children were asked the same two control and the same test question at two points in time (see also Figure 8):

- Control question 1: Where was O before?
- Control question 2: Where is O now?
- Test question: Where does (protagonist) now think O is?

Results and discussion

Children answered control questions correctly in 91% of the trials. The mean sums of trials answered correctly as a function of Belief Type and Belief Order are depicted in Figure 9. A 2 (belief order: FB-TB vs.TB-FB) x 2 (belief: FB-TB) mixed factors ANOVA revealed a main effect of belief ($F(1,30) = 101.52$, $p < .001$, $\eta_p^2 = .77$), no main effect of

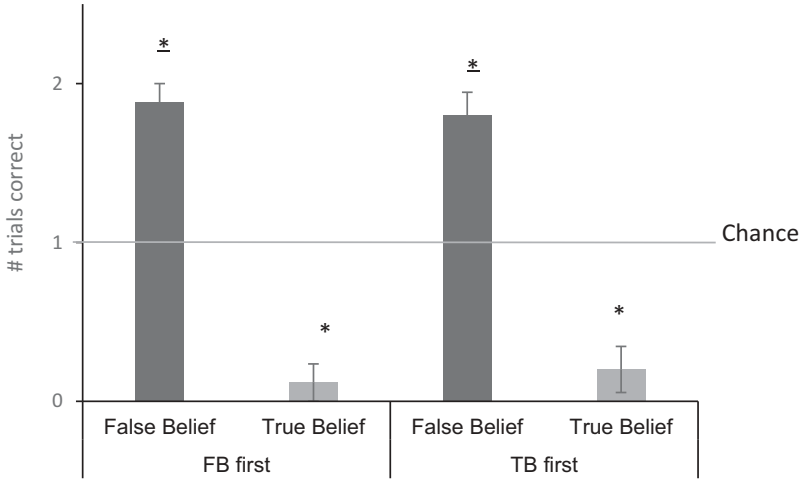

Figure 9. Mean number of trials answered correctly (0–2) as a function of condition in Study 4.

belief order ($F(1,30) = 0$, $p = 1$) and no interaction effect between these factors ($F(1,30) = .24$, $p = .63$). Overall children performed above chance on FB ($t(31) = 9.23$, $p < .001$, $d = 1.64$) and significantly below change on TB ($t(31) = -9.23$, $p < .001$, $d = -1.64$). FB and TB were significantly negatively correlated ($r = -.64$, $p < .001$; partial correlation controlling for age and verbal abilities, $r = -.61$, $p < .001$).

The present findings thus closely reproduce the puzzling pattern of results from previous TB tasks (success in FB, failure in TB, negative correlations between them) and did not supply evidence that making the test question less trivial makes the pattern disappear. Perhaps, however, the present attempt to reduce triviality was simply not suitable.

Study 5

Study 5, therefore, aims at reducing test question triviality in an alternative way, not by modifying the enacted scenario, but by adapting the pragmatic framing. Children received standard, separete FB and TB tasks, but were told by the experimenter beforehand that participants of very different ages, many of them much younger children, and even babies, would participate and that questions for all kinds of age groups would be given to all of them so that some of the questions would be "really very easy, made for much younger children". This explanation supplies a reason for asking such trivial questions and should thus alleviate pragmatic confusion on the part of the child.

Method

Participants

Thirty-two 4- to 6-year-olds (49–82 months, $M = 64$, $SD = 10$) from mixed socioeconomic background were tested. Children were recruited from a databank of children whose

parents had previously given consent to experimental participation. Children were tested by a female experimenter (E) in the laboratory.

Design and procedure

Children received a block of two trials of a standard FB and a block of two trials of a standard TB task, with order of blocks varied between children. The design was thus a 2 (Belief type: FB/TB) x (order: FB first/TB first) design, with order as a between- and belief type as a within-subjects factor.

Intent clarification introduction. At the beginning, E told the child the following: "Today we are going to play three games. Some of them are really easy, like for 2-year-olds, some are still easy, like for 3-year-olds and some are just for your age, which means that they are not a problem for you. Let's start with something for babies!" E introduced a protagonist who was going to join the first game and showed the child and the protagonist an object (e.g. a bell) and asked the child what the object was. Then she asked what the protagonist thought what the object was. This was introduced to show that in the present context asking a question about the belief of an agent did not imply that the protagonist's belief had to be different from the child's belief/the true state of affairs. Two such introductory trials were administered. On the first trial, many children (14 out of 18) still struggled with the pragmatics of the question, either denying to give an answer at all or claiming that the protagonist thought that the object was something else (e.g. the given object was a bell and children said that the protagonist thought it was a nut). After E then clarified once more that the task was really easy and the protagonist was asked what she thought what the object was and she gave the same answer as the child, in trial 2 virtually all children (17 out of 18) answered correctly that the protagonist thought the same as they themselves. After the two introductory trials, E introduced the FB/TB test trials in the following way: "Now I have another game for you! This may be easier, a little bit harder or as easy as the one we just played."

Standard FB/TB tasks. Four trials of standard change-of-location tasks with different stimuli were administered per child, 2 in TB and 2 in FB versions (Wimmer & Perner, 1983). The protagonist and the child were introduced to an object O. The object was then placed in one of two boxes (box 1) before the protagonist left. Either in her absence (FB condition) or after her return (TB condition), the object was moved to the other box (box 2) and the following control and test questions were asked.

- Control Question 1: Where did we put O in the beginning? [correct answer: box 1]
- Control Question 2: Where is O now? [correct answer: box 2])
- Test question: Where does the puppet think O is? [correct answer: box 2 (TB)/box 1(FB)]

Results and discussion

Children answered all control questions correctly in all trials. The mean sums of trials answered correctly as a function of Belief and order are depicted in Figure 10. A 2 (belief order: FB-TB vs. TB-FB) x 2 (belief: FB/TB) mixed factors ANOVA revealed

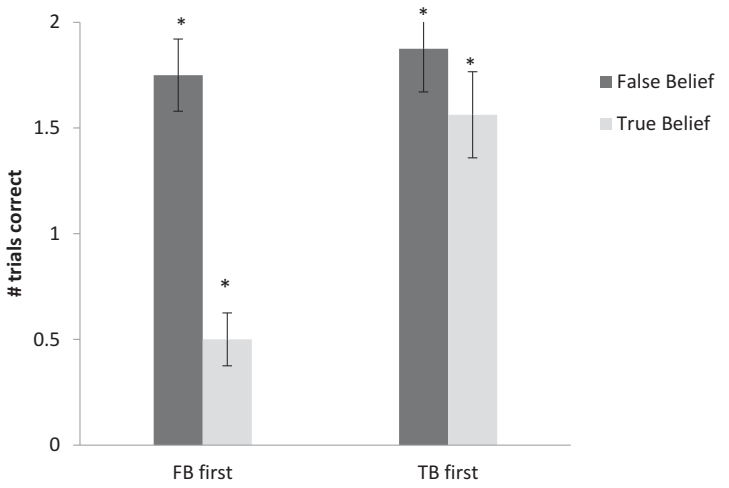

Figure 10. Mean number of trials answered correctly as a function of condition in Study 5.

a main effect of belief ($F(1,30) = 13.19$, $p < .001$, $\eta_p^2 = .31$), a main effect of belief order ($F(1,30) = 19.98$, $p < .001$, $\eta_p^2 = .40$) and an interaction effect between these factors ($F(1,30) = 4.75$, $p < .05$, $\eta_p^2 = .14$). Therefore, separated tests against chance were conducted for the different belief order groups. While children who started with FB still performed at above chance levels in FB ($t(15) = 4.39$, $p < .001$, $d = .1.10$) but at below chance levels in TB ($t(15) = 2.45$, $p < .05$, $d = .61$x), children who started with TB performed significantly above chance in both FB and TB (FB: $t(15) = 7.00$, $p < .001$, $d = .1.74$ and TB: $t(15) = 2.76$, $p < .05$, $d = .69$). Overall, there was (a trend for) a negative correlation between FB and TB ($r = -.33$, $p = .07$; partial correlation, $r = -.41$, $p < .05$). Inspection of correlations as a function of order revealed that there was a strong negative correlation between FB and TB in the sub-sample with the FB-TB order ($r = -.72$, $p < .01$; partial correlation, $r = -.791$, $p < .001$), but not in the TB-FB order sub-sample ($r = -.14$, $p = .60$; partial correlation, $r = -.07$, $p = .81$).

The findings of Study 5 thus show that under some circumstances the puzzling TB performance patterns can be made to disappear. This means that children do not have irremediable difficulty with trivial test questions about an agent's perspective and thus no principled competence limitation. At least in the "TB first" condition, children performed proficiently in both FB and TB. What remains unclear from the present data is why the effect of the experimental modification was not observed in the "FB first" condition. Two (not mutually exclusive) possibilities are, first, that children forgot about the explanation for the triviality of the question by the time they were asked the second test question of a given trial (the TB question in the "FB first" condition). Second, the manipulation regarding the TB task may have been masked by carry over effects in the FB first condition. After the first FB question, children may have perseverated from first to second test question (thus repeating the answer that was correct in FB in TB where it was then wrong). Future research will need to explore these possibilities more systematically.

General discussion

Summary of main findings

The present studies investigated why children from age 4, when they have begun to pass classical False Belief tasks, reveal a very puzzling pattern of performance on trivial control True Belief tasks (fail TB while passing FB tasks, with negative correlations between the tasks). Competence limitation accounts were tested against pragmatic performance limitation accounts. The former assume that children's failure in TB tasks reveals a conceptual competence limitation (children operate with simpler heuristics that happen to deliver correct answers in FB and wrong answers in TB). Performance limitation accounts, in contrast, assume that children fail TB tests because of their confusing task pragmatics. A number of factors – in particular, that these tasks administer trivial test questions about an agent's perspective – may make them pragmatically odd and thus merely mask children's competence. Therefore, in seven studies, these potentially relevant factors were experimentally manipulated.

The main findings were the following: First of all, several manipulations of the crucial factors did actually affect performance, alleviating difficulty in TB so that children performed as proficiently on FB and TB tasks (in some cases with positive FB-TB correlations). This pattern by itself clearly speaks against any principled conceptual competence limitation and in favor of performance limitation accounts. Second, not all such modifications did actually make tasks easier and made performance problems disappear. Removing test questions altogether (Study 1), removing the question's reference to an agent's perspective (Study 2) as well as reducing the question's triviality by giving some background explanation and context (Study 5) did make tasks easier. Attempts to change the question from test to genuine question (Studies 3a-3c) as well as attempts to reduce the triviality of the test questions by introducing belief change over time (Study 4), in contrast, did not change children's TB performance. Why some of the manipulations proved effective while others did not, we cannot tell from the present data. As discussed above, the attempts to implement genuine instead of test questions in Studies 3a-3c simply may not have been credible and stringent enough. Similarly, introducing belief change over time (Study 4) may simply not be a suitable measure for children this age to highlight the possibility of mis-representation and thus reduce the triviality of the TB question. At the current stage of inquiry, needless to say, this remains purely speculative. Future research will need to investigate more systematically under which conditions TB test questions are or are not pragmatically confusing.

Theoretical implications

Taken together, the present findings are compatible with performance limitation accounts regarding the puzzling performance patterns of older children in TB tasks, but do not accord in any obvious ways with competence limitation accounts: Once some limiting performance factors are removed or modified, older children perform as proficiently on TB and on FB tasks, and the two tasks are positively correlated. These findings thus also corroborate the standard interpretation of FB tasks: that they show that children from age four operate with solid meta-representational capacities (see Rakoczy, 2017).

Open questions

At the same time, the present findings leave open many fundamental questions. From a (cross-)linguistic point of view, if the U-shaped curve reflects pragmatic performance factors to do with the explicit questioning, do such effects for all kinds of languages invariably? From a cognitive point of view, how exactly can we explain the U-shaped curve in TB performance according to a pragmatic performance limitation account? Such an account claims that children, once they have developed the requisite ToM skills and the pragmatic capacities that build on them, suffer from pragmatic confusions in TB tasks. The present results show that children age 4–6 can be helped to overcome these confusions by relevant task modifications. But why do older children (from around age 8–10 in previous studies) and adults overcome these confusions even in standard tasks without any modifications? Again, at the current stage of inquiry we do not know and can only speculate. Two potential speculations that appear promising, to which we now turn, revolve around abductive inference and recursive higher-order intentionality:

Intuitively, older children and adults operate with an open-ended, flexible, generative capacity for abductive pragmatic inferences (inferences to the best explanation) that allows them to make sense of virtually any potential speech act. Coming back to our example of A, B, C and the cat, when A asks B "What does C think this (the cat) is?", B may, for example, remember that A has just started to take a class in epistemology and now keeps on wondering about the foundations of our perceptual beliefs. Or B may wonder whether there is now a new class of robots around that look like cats etc. – there is virtually no limit to what could count as an additional relevant premise that helps to make sense of A's intention and the corresponding speech act. Study 5 shows, for example, that children can take into account such additional premises ("some of the questions are made for much younger children") when they are explicitly provided. One possibility is thus that what changes over developmental time is the general capacity to spontaneously generate or examine potential additional premises in abductive inferences to make sense of a given speech act. This increasing pragmatic flexibility may be intimately connected with children's growing experience, in particular in schooling contexts, in which learning to respond to pragmatically deviant speech acts such as academic test questions etc. is a ubiquitous challenge (e.g., Mercier, 2011).

Another possibility is that what develops is, more specifically, a form of complex, recursive higher-order intentionality ascription. According to many accounts in pragmatics, notably those in the Gricean tradition, understanding speech acts, in particular indirect and complex ones, is a matter of pragmatic inferences about the underlying higher-order, recursive speaker intentions such as "She intends me to believe that p on the basis of my recognition of her intention to make me believe that p" (Grice, 1975; Sperber & Wilson, 1986). From Theory of Mind research we know that children's ascription of intentional attitudes becomes more complex over developmental time: from around 4 years, children reliably master standard FB tasks which require the ascription of first order beliefs ("She believes the object is in box 1"); from around age 5–6, children then master more complex tasks that require the ascription of second-order beliefs ("She believes that he believes that the object is in box 1") (Perner & Wimmer, 1985; Sullivan, Zaitchik, & Tager-Flusberg, 1994). But of course this is by far not where things end. Common sense and recent research suggest that adults have no difficulty to

engage in much more complex forms of recursive intentionality ascription ("she thought that he had understood that she wanted him to believe that she held no such prejudices regarding his intentions towards her feelings"), up to at least 7th order (O'Grady, Kliesch, Smith, & Scott-Phillips, 2015). From a developmental point of view, however, little is known about the ontogenetic trajectories of ever more complex forms of recursive intentional ascription in between age 6 (2nd order) and adulthood (7th order or so). Thus, it is possible that important developmental progressions in this period actually drive increasing pragmatic sensitivity and competence which in turn explains why children's trouble with trivial TB tasks disappear at some point.

Future research will thus need to investigate more systematically whether the developmental changes in TB performance around age 8–10 marks a more general transition in pragmatic capacities; and if so, whether developmental changes in children's growing general inferential complexity, or in their more specific capacity for recursive higher-order intention ascription, or a combination of both can account for these pragmatic transitions.

Summary & outlook

The present findings show that the puzzling difficulties children from age 4 suffer in trivial TB tasks are an artifact of task pragmatics. Once the pragmatic structure of the tasks is suitably modified, children perform as well on TB as on FB tasks and performance in the two tasks converges. These results speak against any principled competence limitations, and corroborates the standard picture that (at the latest) from age 4 children have developed a solid capacity for meta-representation. How this capacity is then put to practice in pragmatically ever more complex situations over developmental time is still an open question for future research.

Acknowledgments

We would like to thank Natalie Bleijlevens for help with testing and coding. Thank you very much to Marlen Kaufmann and Konstanze Schirmer for the organization of the studies.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contribution to theory of mind: Neuropsychological evidence from a false belief task with reduced language and executive demands. *Journal Of Cognitive Neuroscience*, *16*, 1773–1784.

Bailey, P. E., & Henry, J. D. (2008). Growing less empathic with age: Disinhibition of the self-perspective. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, *63B*(4), P219–P226. doi:10.1093/geronb/63.4.P219

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development*, *70*(2), 381–395. doi:10.1111/cdev.1999.70.issue-2

Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Developmental Psychology*, *46*(6), 1402–1416. doi:10.1037/a0017648

Fabricius, W. V., & Khalil, S. L. (2003). False beliefs or false positives? Limits on children's understanding of mental representation. *Journal of Cognition and Development*, *4*(3), 239–262. doi:10.1207/s15327647jcd0403_01

Fizke, E., Barthel, D., Peters, T., & Rakoczy, H. (2014). Executive function plays a role in coordinating different perspectives, particularly when one's own perspective is involved. *Cognition*, *130*(3), 315–334. doi:10.1016/j.cognition.2013.11.017

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 41-58). New York, NY: Academic Press.

Hedger, J. A., & Fabricius, W. V. (2011). True belief belies false belief: Recent findings of competence in infants and limitations in 5-year-olds, and implications for theory of mind development. *Review of Philosophy and Psychology*, *2*(3), 429–447. doi:10.1007/s13164-011-0069-9

Kaufman, A., & Kaufman, N. (1999). *Kaufman assessment battery for children* (4th ed., P. Melcher & U. Preuß, trans.). Frankfurt am Main: Swets Test Services.

Mercier, H. (2011). On the universality of argumentative reasoning. *Journal of Cognition and Culture*, *11*, 85–113. doi:10.1163/156853711X568707

O'Grady, C., Kliesch, C., Smith, K., & Scott-Phillips, T. C. (2015). The ease and extent of recursive mindreading, across implicit and explicit tasks. *Evolution and Human Behavior*, *36*(4), 313–322. doi:10.1016/j.evolhumbehav.2015.01.004

Oktay-Gür, N., & Rakoczy, H. (2017). Children's difficulty with true belief tasks: Competence deficit or performance problem? *Cognition*, *166*, 28–41. doi:10.1016/j.cognition.2017.05.002

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

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that … ": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471. doi:10.1016/0022-0965(85)90051-7

Rakoczy, H. (2017). In defense of a developmental dogma: Children acquire propositional attitude folk psychology around age 4. [journal article]. *Synthese*, *194*(3), 689–707. doi:10.1007/s11229-015-0860-8

Rakoczy, H., Fizke, E., Bergfeld, D., & Schwarz, I. (2015). Explicit theory of mind is even more unified than previously assumed: Belief ascription and understanding aspectuality emerge together in development. *Child Development*, *86*(2), 486–502. doi:10.1111/cdev.12311

Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way: A case of a selective deficit in inhibiting self-perspective. *Brain: A Journal of Neurology*, *128*(5), 1102–1111. doi:10.1093/brain/awh464

Siegal, M. (1999). Language and thought: The fundamental significance of conversational awareness for cognitive development. *Developmental Science*, *2*(1), 1–14. doi:10.1111/1467-7687.00048

Sperber, D., & Wilson, D. (1986). *Relevance*. Cambride, MA: Harvard University Press.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*(3), 395–402. doi:10.1037/0012-1649.30.3.395

Wellman, H., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684. doi:10.1111/1467-8624.00304

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs - representation and constraining function of wrong beliefs in young childrens understanding of deception. *Cognition*, *13*(1), 103–128. doi:10.1016/0010-0277(83)90004-5

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition*, *35*, 41–68. doi:10.1016/0010-0277(90)90036-J

Appendix A. Detailed procedure of the non-verbal FB/TB task in Study 1

General setup

Children were introduced to a sticker game in which they could win stickers they were allowed to select from a box containing several stickers. A protagonist who joined the game to help the child was introduced. Each trial began by the child choosing the sticker. The sticker was then hidden in one of two boxes behind an enclosure invisible to the child while the protagonist was behind the enclosure and could see the sticker's location. After the enclosure was removed, the protagonist gave an advice by saying "I think it is in here" and pointing to one of the two boxes. The experimenter then moved the boxes toward the child who was allowed to choose freely in which of the boxes she wanted to look. If the child found the sticker she won it, if she looked in the wrong box or tried to cheat, the sticker was lost and placed in a savings box being no more available.

Warm-up trials and control questions

Warm-up trials were closely modeled after Call and Tomasello (1999), Apperly and colleagues (Apperly, Samson, Chiavarino, & Humphreys, 2004) and Fizke et al. (2014). Children received two trials of a control in which the puppet saw the sicker being hidden and immediately after the removal of the enclosure gave his advice ("Control Trial"). In the following two trials ("Invisible Displacement") after the protagonist gave his advice the boxes were exchanged visible to the child and then the child was allowed to search. In the next two trials ("Ignore Communicator"), after the removal of the enclosure the protagonist claimed that he had to leave. In the protagonist's absence the sticker was moved from one to the other box by visibly (to the child) transferring it to the other box. After the puppet's return, the puppet gave a (wrong) advice in accordance to his belief. And the child was again allowed to search. Before the test trials began again two Control Trials were played. These warm-up trials aimed to make clear that the child was allowed to disobey the protagonist's hint.

Performance on control questions

Children solved 84% of all Control Trials at the beginning, 74% of all Invisible Displacement trials, 95% of all Ignore Communicator Trials as well as 88% of all Control Trials at the end.