Cognition 166 (2017) 28–41

Contents lists available at ScienceDirect

Cognition

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

Children's difficulty with true belief tasks: Competence deficit or performance problem?

Nese Oktay-Gür *, Hannes Rakoczy

Institute of Psychology, University of Göttingen, Germany

ARTICLE INFO

Article history:
Received 4 July 2016
Revised 24 April 2017
Accepted 1 May 2017

Keywords:
Theory of mind
Social cognition
Pragmatics
Mental File Card Theory

ABSTRACT

According to the standard picture of explicit theory of mind (ToM) development, children begin to (explicitly) ascribe beliefs to others and themselves from around age 4. The empirical basis of this picture comes from numerous studies consistently showing that children master verbal false belief (FB) tasks from around age 4 while children much younger have no difficulty in mastering structurally analogous true belief (TB) tasks. The standard picture, though, has come under serious attack from recent studies using TB tasks with wider age ranges. These studies have found that, paradoxically, children begin to fail TB tasks once they master FB tasks. Such findings cast doubt on the standard picture and suggest, instead, that FB tasks may be solved by much simpler strategies than proper belief reasoning. In the present study, we tested for the development of FB and TB performance in comprehensive and systematic ways. In particular, we tested the competing predictions of competence accounts (according to which TB failure reflects lack of conceptual competence) versus performance limitation accounts (according to which the standard picture is true yet children from around age 4 fail TB tasks due to performance factors). Studies 1 and 2 showed that performance in a variety of novel TB tasks showed a clear U-shaped curve, with children until age 3 and from age 10 performing competently and children in between failing, with strong negative correlations between TB and FB. Crucially, these patterns were found for various kinds of TB tasks, including those for which existing competence limitation accounts would not even predict any difficulty. Study 3, therefore, directly tested performance limitation accounts in terms of pragmatic and related factors and found that these patterns (failure in TB and negative TB-FB correlations) disappear once the relevant performance factors have been removed from the TB tasks. Taken together, these findings suggest that previous TB findings constitute false negatives, clearly speak for performance limitation accounts and thus corroborate the standard picture of the development of explicit theory of mind.

1. Introduction

The social-cognitive capacity to ascribe mental states to others and ourselves, also known as Theory of Mind (ToM) is crucial to almost all aspect of our social lives. Concerning its measurement, false belief (FB) tasks have emerged as the developmental litmus tests for tapping basic ToM (Wimmer & Perner, 1983; for an overview see Wellman, Cross, & Watson, 2001). Such tasks require the prediction or explanation of an agent's rational action on the basis of her outdated or otherwise mistaken beliefs. Empirically, hundreds of studies have consistently shown that children younger than 4-years systematically fail FB tasks (while they do not have problems in passing analogous true belief control tasks) whereas children older than 4 years systematically pass. These converging results have standardly been interpreted as indicating a deep conceptual change or even revolution around age 4 (Perner, 1991).

This standard interpretation of a conceptual 4-year-revolution, however, has recently come under serious attack from different directions (see Rakoczy, 2015). On the one hand, much new research with implicit tasks (showing sensitivity to other agents' belief in looking time and interactive measures) has been taken to suggest that toddlers' incompetence may constitute false negatives (e.g. Buttelmann, Carpenter, & Tomasello, 2009; Onishi & Baillargeon, 2005; Southgate, Chevallier, & Csibra, 2010; Southgate, Senju, & Csibra, 2007; Surian, Caldi, & Sperber, 2007; for review see Baillargeon, Scott, & Bian, 2016; Baillargeon, Scott, & He, 2010; Baillargeon et al., 2015).

* Corresponding author at: University of Göttingen, Institute of Psychology, Waldweg 26, 37073 Göttingen, Germany.
E-mail address: nese.oktay-guer@uni-goettingen.de (N. Oktay-Gür).

http://dx.doi.org/10.1016/j.cognition.2017.05.002
0010-0277/

1.1. Do 4-year-olds really operate with a concept of belief? Skeptical concerns

From the opposite direction it has been argued that children's passing of standard FB tasks from around age 4 may actually constitute false positives and massively over-estimate children's ToM competence. It is this attack on the standard interpretation that will be the focus of the present paper.

A number of recent empirical findings constitute the empirical basis of this line of attack. One set of such findings suggests that children, when they master standard FB tasks, still do not understand a fundamental feature of beliefs and thus cannot properly be said to ascribe any beliefs at all: beliefs and other propositional attitudes are essentially aspectual, that is they only hold under certain aspects and not under others (Frege, 1980 [1892]; Searle, 1983; for an overview see McKay & Nelson, 2014). An agent may believe, for example, that Clark Kent is at home without thereby believing that Superman (in fact identical to Clark Kent) is at home. Yet many studies have suggested that children up to 6–8 years of age fail to respect this aspectuality in their belief ascriptions (Apperly & Robinson, 1998; Kamawar & Olson, 1999; Kamawar & Olson, 2009; Kamawar & Olson, 2011; Russell, 1987; Sprung, Perner, & Mitchell, 2007). For example, in one kind of scenario, children were presented with an obvious eraser in box 1, and a dice that was also non-obviously an eraser in box 2, and an agent who was unaware of the hidden identity of the dice. When asked where the agent would look for an eraser (correct answer: "box 1"), children failed to take into account the aspectuality of the agent's beliefs and answered incorrectly, indifferently or "both" (Apperly & Robinson, 1998). Recent work, however, suggests that once they are suitably modified and simplified, even 4-year-olds master such aspectuality tasks (Rakoczy, Fizke, Bergfeld, & Schwarz, 2015 see below).

A second set of findings that suggest that FB tasks may overestimate children's competence comes from true belief control tasks. In standard FB studies, true belief (TB) conditions, in which everything is more or less like in FB conditions with the exception that the protagonist is not mistaken, usually serve as mere baseline measures with younger children to rule out that they fail FB tasks because they somehow cannot cope with the narrative task structure. Standardly, 3-year-olds indeed have no problems in mastering TB tasks while systematically failing FB tasks. However, TB control tasks have rarely been used with older children who have come to master FB tasks – based on the background assumption that children master explicit TB tasks from early on and continue to do so but only come to master FB tasks around age 4 when they acquire true meta-representational capacities (e.g. Perner, 1991).

1.2. Older children's failure in true belief tasks

However, some recent research has used FB and TB tasks with wider age ranges and has produced surprising patterns of findings. Some of these studies have used change-of-location TB tasks matched to the standard change-of-location ("Maxi"/"Sally-Anne") FB tasks (Wimmer & Perner, 1983). In the TB versions, the protagonist failed to witness some relevant events, but luckily ended up having a true belief. For example, she put object O in box 1 and left. Her sister then removed O and thought about putting it into box 1 or box 2, finally deciding for box 1. Then the protagonist came back and the test question was where she believed O was/where she was going to search for O (Fabricius, Boyer, Weimer, & Carroll, 2010). These scenarios thus present something similar to what is known in philosophical epistemology as "Gettier cases" (after Gettier, 1963): cases where an agent has a justified true belief (that O is in box 1), in which, however, we would be hesitant to attribute to her knowledge of the fact in question, simply because her belief has not the right kind of history (she failed to witness too many crucial steps). Empirically, the results with these kinds of TB and FB tasks have produced striking findings: 3-year-olds passed TB and failed FB tasks, 4- to 6-year-olds showed the reverse pattern, and only children from age 6 passed both FB and TB tasks (Fabricius et al., 2010). Another recent set of studies has used FB and TB versions of aspectual belief tasks with a similar age range and has found similar patterns of performance between the ages of 3 and 6 (Perner, Huemer, & Leahy, 2015).

1.3. Failure in true belief tasks: Competence or performance limitation?

What do these patterns of findings in TB tasks show? In general, there are two potential kinds of explanations: Performance accounts assume that negative results in TB tasks present false negatives that do not reflect a lack of competence, but merely some performance limitation due to extraneous factors. Competence accounts, in contrast, claim that these negative results do reflect limitations of conceptual (meta-representational) competence. If they were true, competence accounts would have far-reaching implications. In particular, they would put into question the standard assumption that children acquire true meta-representational capacities by age 4.

1.3.1. Competence limitation I: Perceptual Access Reasoning

One competence account, the so called *Perceptual Access Reasoning* (PAR) account assumes that the patterns of FB and TB findings show that children before age 6 do not use proper belief ascription but simpler conceptual strategies (Fabricius et al., 2010; Hedger & Fabricius, 2011; Recanati, 2012; Westra & Carruthers, 2016). Children's reasoning in FB/TB tasks, according to this account, undergoes three stages. In the first stage, before age 4, children use merely reality-based reasoning (agents search objects were they are) and thus pass TB while failing FB. In the second stage, between ages 4 and 6, children use so-called Perceptual Access Reasoning (PAR) according to which agents with full perceptual access to a situation get things right and agents lacking full perceptual access get things wrong. This strategy leads to correct performance in FB tasks. However, in TB tasks with Gettier-like cases such as the ones used by Fabricius et al. (2010) mentioned above, in which the agents fails to witness some crucial event and luckily ends up with a true belief, this strategy yields wrong answers. It is only in the third stage, from around age 6, that children then use belief reasoning proper, resulting in competent TB and FB performance. This specific account thus would predict U-shaped development in performance on a specific class of TB tasks, namely those in which the protagonist ends up with a TB despite limited perceptual access to a crucial step in the course of events.

1.3.2. Competence limitation II: Immature Mental File Card architecture

Another competence account predicts a similar U-shaped curve for TB performance, yet from a very different theoretical point of view, and for a different sub-class of TB tasks. The so-called *Mental File Card Account* by Perner and colleagues (Perner & Leahy, 2015; Perner et al., 2015) presents a formal theory of the sub-personal underpinnings of ToM reasoning with the help of the machinery of mental files (Recanati, 2012). The basic assumption is that representation of individuals in the world is realized via object files – representational structures that individuate referents (e.g. "Clark Kent") and that can include predicative information (e.g. "lives in a terraced house"). In discourse and thought, once a new object is encountered, a new object file is opened. Children operate with such basic object files from very early on in ontogeny, as can be seen, for example, in their object individuation and numerical

judgments (e.g. Carey, 2009; Feigenson, Carey, & Hauser, 2002). But more complex forms of handling object files emerge only later in development: around age 4, children begin to operate with *horizontal links* between object files: when children initially have two separate files for, say, Clark Kent and Superman, respectively, and then discover that they refer to the very same object, they can now link the two files so that predicates applying to one can be seen to apply to the other as well. This explains why children begin to properly understand identity statements only around this time (Perner, Mauer, & Hildenbrand, 2011). At the same time, children now begin to operate with *vertical links* between real files and so-called *vicarious files* that are used to represent the content of others' mental representations such as their beliefs. When the child, for example, sees object O in box 1, and sees the protagonist P witness this, she opens a vicarious belief file with the content "O is in box 1" attached to P. When O is then moved to box 2 in P's absence, the original vicarious file remains in place while the real O-file is updated to state that "O is in box 2". The fact that children around age 4 begin to operate with vertical linking explains why at this time they begin to master explicit FB tasks. But crucially, children around age 4 still cannot properly coordinate horizontal and vertical linking. And it is this missing linking that predicts that children this age should show the paradoxical pattern, for a specific sub-class of FB/TB tasks, of solving FB while failing TB tasks. The sub-class in question is that of aspectual FB/TB tasks. Imagine the following scenario, witnessed by protagonist P: Clark Kent (who is in fact Superman) enters house H. Superman then flies out of H and to the beach. P does (TB) or does not (FB) know the crucial identity Clark Kent = Superman, and the test question is: "Where does P think Clark Kent is?" (correct answers: in house H (FB)/at the beach (TB)). A child who can engage in horizontal linking (of the "Clark Kent" and "Superman" file cards) and in vertical linking (of the real "Clark Kent" file card and the vicarious file card attached to P "Clark Kent is in H"), will be able to answer FB correctly (because her vicarious Clark Kent file card attached to P still says "is in H"). But paradoxically, such a child will fail to answer TB correctly since when she learns that P knows about the Clark Kent = Superman identity, she will not coordinate horizontal links within the vertical links in such a way that now all the information pertaining to the vicarious "Clark Kent" and "Superman" files is treated as interchangeable within P's belief (if he knows that Superman is at the beach and knows that Clark Kent = Superman, he thereby knows that Clark Kent is at the beach). Rather, the two vicarious file cards of "Superman" and "Clark Kent" remain unlinked, and as a consequence, the 4-year-old child in the TB condition will give the incorrect answer that P will believe that Clark Kent is in the house H.

Empirically, a recent set of studies with FB and TB versions of aspectual belief tasks has found evidence for exactly these patterns of performance predicted by the Mental File Card Account. In these tasks, building on the simplified aspectual FB tasks by Rakoczy and colleagues (2015), there was an object that was both an A and a B. The object was put, as an A, into box 1, and then transferred, as a B, to box 2. All of this was witnessed by an agent and the test question was where this agent would look for the A – with the crucial difference between conditions being whether the agent did (TB) or did not (FB) know about the A = B identity. Again, the empirical results were striking: 3-year-olds tended to pass TB (by answering "box 1) and fail FB ("box 1" as well), 4- to 6-year-olds tended to show the reverse pattern, and only from age 6 did children reliably solve both FB and TB (Perner & Leahy, 2015; Perner et al., 2015).

1.3.3. Performance limitation accounts

In contrast to such competence deficit explanations, performance limitation accounts argue that the standard picture of a 4-year-conceptual revolution may be untouched by the negative findings from TB and other ToM tasks since these (false) negatives may simply reflect performance limitations. Concerning children's failure in aspectuality tasks, for example, recent findings suggest that children from age 4 have no problems in taking into account the aspectuality of an agent's belief once the task is suitable simplified and irrelevant performance factors (such as memory demands) have been removed (Rakoczy et al., 2015; see above).

Concerning children's failure in TB tasks, it is not yet clear what the relevant performance factors might exactly be that could account for the poor performance of 4- to 6-year-olds. One prominent possibility is that it is extraneous pragmatic factors associated with the specific tasks and their formats that make them confusing and unnecessarily difficult. Pragmatics performance factors have been invoked to explain (away) surprising performance limitations in many areas of cognitive development. For example, it has been argued that many classical Piagetian pre-operational failures stem from a lack of understanding test questions and related pragmatic rather than genuinely cognitive limitations (Siegal & Beattie, 1991). In the area of theory of mind, pragmatic factors have long been claimed to be a fundamental performance factor that helps explain young children's failure on verbal FB tasks (for the most recent work along such lines, see Helming, Strickland, & Jacob, 2014; Helming, Strickland, & Jacob, 2016; Westra, 2016; Westra & Carruthers, 2016). In the particular case of TB tasks under consideration here, pragmatic performance factors along the following lines may apply: these tasks are artificially difficult because children are asked (i) stunningly trivial questions, (ii) about the beliefs or actions of a protagonist who has perfect informational access and is thus not mistaken in any way and (iii) the question itself is asked as a test question (a question posed by someone who, of course, knows the answers and thus does not ask for information, but asks for the sake of testing whether the child knows the answer). The older children get, the more their ToM and corresponding pragmatic capacities get; in particular children become more sensitive to the pragmatic fact that we usually do not talk much about beliefs when they are true –the main point of belief talk, after all, being to refer to or at least raise the possibility of their falsity (Papafragou, Cassidy, & Gleitman, 2007). And so the older children get, the more they may start to wonder about a potential hidden agenda behind the TB questions ("It is so obvious, why is she asking me this stupid question?"), reasoning that they must have missed or misunderstood something ("So the correct answer must be different from the obvious one – otherwise, why would she ask me, after all?").

Other (potentially complementary) performance factors may have to do with the relevance or salience of the agent's beliefs in the TB scenarios. It may be that these scenarios are so boring that the agent's beliefs never really become sufficiently salient or relevant to the child. This should pose no problems to younger children (who, in the absence of a solid concept of belief, answer the test question on the basis of reality anyway); but it may well confuse older children who do have a concept of belief, yet with a fragile capacity for its application. The agent's belief, such a line of reasoning goes, are sufficiently salient and relevant in FB tasks, and so children successfully apply their belief concept there; but the agent's beliefs are not sufficiently relevant and salient in the TB tasks, which is why children there fail to translate their conceptual competence into successful performance.

1.4. Rationale of the present studies

The aim of the present studies was therefore to systematically investigate the development of children's patterns of TB performance, and to test whether these patterns can be best explained by competence or by performance limitation accounts. To do so,

we first investigated the development of FB and TB performance in a comprehensive design with different kinds of ToM tasks (standard location change and aspectuality) across a wide age range (from age 3 to adulthood) to see whether TB-performance generally and robustly yields a U-shaped curve while FB performance simply increases with age. Secondly, we derived and tested competing predictions of competence vs. performance limitation accounts: Competence limitation accounts predict U-shaped curves in TB tasks only in specific cases under limited circumstances: The Mental File Card Theory, predicts a U-shaped curve for TB performance only in the specific sub-class of aspectual TB tasks (but no such pattern for standard change-of-location TB tasks). The other competence limitation approach, the PAR account, predicts a U-shaped curve only for the sub-class of TB tasks in which the protagonist has "comparable lack of perceptual access" relative to FB tasks (Hedger & Fabricius, 2011, p. 432).

Performance limitation accounts in terms of extraneous task factors surrounding TB tasks (such as salience/relevance, and/or pragmatics), in contrast, would predict U-shaped curves in TB tasks to be a much more general phenomenon. First of all, the pattern of TB and FB performance should be analogous over different types of tasks (standard change-of-location and aspectual), including those for which either the PAR account or the Mental File Card account do not even apply. Second, some performance factor accounts would assume that the sensitivity to the crucial performance factors (that make the TB tasks difficult) depends on ToM (which in turns is tapped in FB tasks); and they would thus predict an inverse relation between FB and TB performance: for both change-of-location and aspectual tasks, children's FB and TB performance should be negatively correlated.[1] Such a prediction follows clearly from pragmatic performance factor accounts: sensitivity to pragmatics is known to depend developmentally on ToM (e.g. Happé, 1993; Winner & Gardner, 1993), and thus increase in ToM (indicated in FB performance) should go along with increase in pragmatic competence, and thus with pragmatic confusion in TB tasks, and thus in general with decrease in TB performance.

Thirdly, once the critical performance factors have been removed or alleviated, children's difficulty with TB tasks (and the negative correlations between TB and FB) should vanish.

These predictions were tested in 3 studies against those of the competence limitation accounts. Studies 1 and 2 investigated the development of performance in standard and aspectual TB and FB tasks from early childhood to adulthood. The results revealed analogous patterns of U-curves in TB in standard change-of-location and aspectual tasks, increase in FB tasks, and negative correlations of FB and TB between ages 3 and 6. In Study 3, new FB/TB tasks were devised that removed potential performance factors (such as the pragmatic oddity and the relevance and salience of the agents' beliefs), and children from age 4 now performed competently on both FB and TB trials.

[1] One important qualification is in order here: Clearly, pragmatic performance factor accounts assume that pragmatically based failure in TB tasks is a transient phenomenon (after all, older children and adults finally do master TB tasks again). Presumably, at some point, children's pragmatic capacities have developed to a higher level at which they now understand why people may engage in trivial and seemingly pointless test questions. At this stage, then, children should be able to apply their belief concept in all kinds of pragmatic situations and thus perform equally competently in FB and TB tasks (with positive correlations between TB and FB). This means that the predicted negative correlation should only be expected in the intermediate period in which children have acquired a concept of belief, are capable of applying it in the FB tasks, yet are pragmatically still vulnerable in the TB tasks. When exactly this period ends is an empirical question (previous research suggests perhaps around age 6, whereas the current findings point to a much more protracted development; see below).

2. Study 1

In a first step, in Study 1, we tested performance in standard change-of-location and aspectuality TB and FB tasks in a comprehensive within-subjects design in children around the alleged 4-year-revolution, from 3;6 to 5;6 years of age.

2.1. Methods

2.1.1. Participants

Twenty-three 3 to 5- year olds (43–65 month, M = 55; 13 male) from mixed socioeconomic background were included in the final sample: Children were recruited from a databank of children whose parents had previously given consent to experimental participation. Children were tested by a male experimenter either in a quiet room of their day care or in the laboratory.

2.1.2. Design and procedure

The basic design was a 2 (belief: TB-FB) X 2 (condition: standard - aspectuality) within-subjects design. Each child received two trials in each of the four conditions, resulting in eight trials in total. The order of the true and FB blocks as well as the order of the tasks within the blocks was counterbalanced across subjects.

2.1.2.1. Verbal ability. At the beginning of the session, children were given a vocabulary test (the vocabulary subscale of the Kaufman Assessment Battery for Children; Kaufman & Kaufman, 2001).

2.1.2.2. Standard 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 X [e.g., a plastic duck]. 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), the object was moved to the other box (box2) and the following control and test questions were asked:

- Control Question 1: Where did we put the X [e.g., the duck] in the beginning? [correct answer: box1]
- Control Question 2: Where is the X now? [correct answer: box2])
- Test question: Where will the protagonist look for the X? [correct answer: box 2 (TB)/box 1(FB)]

What is crucial about the TB version of the standard task is that neither the PAR nor the Mental File Card account would predict that it should be difficult. The latter only applies to aspectual TB tasks, and the former does not apply because the protagonist leaves before the crucial events unfold (transfer of the object) and thus has no lack of relevant perceptual access.

2.1.2.3. Aspectuality task. Four trials of aspectual FB/TB tasks with different stimuli (dual-function and dual-identity) were administered per child, 2 in TB (1 dual function/1 dual identity) and 2 in FB versions (1 dual function/1 dual identity). The basic logic of these tasks (closely modeled after Study 3 of Rakoczy et al., 2015) is depicted in in Fig. 1: In the presence of a protagonist an object was put into a box (box 1) under aspect A [e.g. pen]. In the presence (TB) or absence (FB) of the protagonist 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 now 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

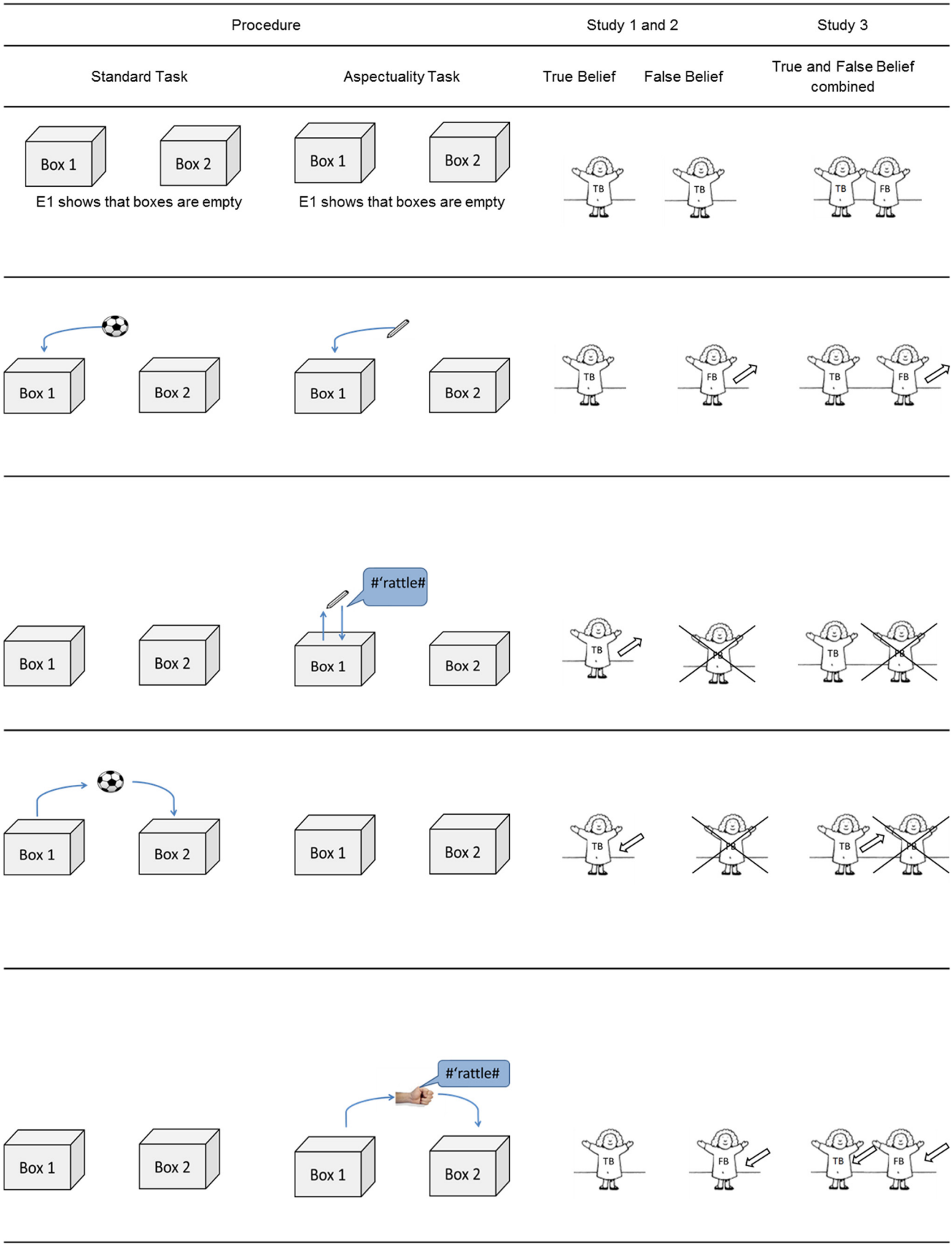

Fig. 1. Procedure of the different tasks used in Studies 1–3.

only the B-identity (rattle) could be heard). In both belief conditions (TB and FB) the protagonist witnessed the transfer of the object. The critical difference between the conditions was that in FB the protagonist did not know that the objects she saw at different time points as A and B were identical. The following control and test questions were asked:

- Control Question 1: Does the protagonist know that the A (e.g. pen) is the B (e.g. rattle)? [correct answer: yes (TB)/no(FB)]
- Control Question 2: Where did we put the A [e.g., pen] in the beginning? [correct answer: box1]
- Control Question 3: Where is the A now? [correct answer: box2]
- Test question: Where will the protagonist look for the A? [correct answer: box 2 (TB)/box 1(FB)]

When children failed to answer the control questions correctly, the experimenter repeated the test question. If children insisted on their wrong answer they were not corrected.

Concerning the TB version of the aspectuality task, only one competence limitation account, the Mental File Card approach, predicts that it should be difficult, whereas the PAR account does not even apply – again, because the protagonist leaves before the crucial events and thus has no lack of relevant perceptual access.

2.2. Results

Children answered the control questions correctly in the following percentages of all trials: Standard FB: Control question 1: 96%; Control question 2: 98%; Standard TB: Control questions 1 and 2: each 100%. Aspectuality FB: Control question 1: 72%; Control questions 2 and 3: each 100%. Aspectuality TB: Control question 1: 70%; Control question 2: 96%; Control question 3: 100%. Children's performance in the control questions was thus generally close to ceiling with the exception of control questions 1 in the aspectuality FB/TB tasks. A closer analyses of performance as a function of age revealed that control question 1 of the aspectual TB tasks was solved by 88% of the 3-year olds (N = 4), by 71% of the 4-year olds (N = 12), and by 57% of the 5-year olds (N = 7). Control question 1 of the aspectual FB tasks was solved by 13% of the 3-year olds (N = 4), by 80% of the 4-year olds (N = 12), and by 93% of the 5-year olds (N = 7).

2.2.1. Consistency across trials and contingency between tasks

The consistencies in performance of children over trials 1 and 2 of the same kind of task were high for all tasks and conditions The percentages of children who showed the same performance in both trials of a given type of tasks were 83% in Standard FB, 91% in Aspectual FB, 83% in Standard TB and 96% in Aspectual TB; all $\Phi s > 0.65$). Therefore, sum scores of trials answered correctly per condition [0–2] were used for further analyses.

2.2.2. Main analyses

The mean sums of trials answered correctly as a function of conditions are depicted in Fig. 2. As a group, children did not perform differently from chance in any of the tasks (standard FB, $t(22) = 0.23$, $p = 0.82$; standard TB, $t(22) = 0.23$, $p = 0.82$; aspectuality FB, $t(22) = 0.65$, $p = 0.53$ and aspectuality TB, $t(22) = 0.42$, $p = 0.68$). This, however, was not due to random performance, but due to the fact that in each task, most children performed either consistently correctly or consistently incorrectly.

Since preliminary analyses (a 2 (belief: FB/TB) × 2 (task: standard/aspectuality) × 2 (order) ANOVA on the mean sum of trials correct) failed to find any main or interaction effects for the order (FB-TB vs. TB-FB) of test blocks (all $ps > 0.55$), this factor was skipped from further analyses. A 2 (belief: FB/TB) × 2 (task: standard/aspectuality) ANOVA on the mean sum of trials answered correctly did not yield any significant main effects (belief, $F(1,21) = 0.00$, $p = 0.95$ and task, $F(1,21) = 0.41$, $p = 0.53$) nor an interaction effect ($F(1,21) = 0.03$, $p = 0.88$)

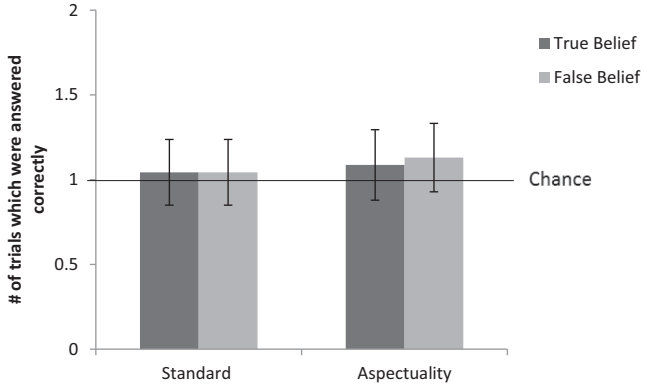

Fig. 2. Mean number of trials answered correctly in the different tasks.

2.2.3. Correlations between the tasks

The correlations between the different tasks are depicted in Table 1. There were high positive correlations between tasks of the same belief condition (TB: $r = 0.64$; FB: $r = 0.81$) and high negative correlations between TB and FB versions of the tasks (standard: $r = -0.72$; aspectuality: $r = -0.86$), even if controlled for age and verbal ability (see Table 1).

2.3. Discussion

The present study tested for patterns of FB and TB performance in children aged 3–5 and found preliminary evidence that speaks for performance limitation accounts: children's FB performance across different types of tasks was highly consistent, as was their TB performance. And performance between TB and FB tasks was inversely related, with substantial negative (even partial) correlations.

These patterns of finding match nicely the predictions of performance limitation accounts (in particular, those in term of pragmatic factors), but seem hard or impossible to reconcile with any of the two competence limitation accounts under consideration here: The PAR account would actually not predict any difficulty in the kinds of TB tasks used here since the protagonist did not fail to have perceptual access to any crucial events. And the Mental File Card account would only predict such a TB pattern for the aspectual, but not for the standard change-of-location tasks.

However, the current results taken by themselves have some limitations: first of all, the age ranges tested were quite narrow (roughly 4 years ± 1). Secondly, given the 2 × 2 within-subjects design with 2 trials per task, the session was rather taxing. It might have produced some noise, and it might not be suitable for testing even younger 3-year-olds. Study 2, therefore, tested a wider age range of subjects, ranging from age 3 to adulthood, with a modified design resulting in a shorter and less taxing session.

3. Study 2

3.1. Method

3.1.1. Participants

171 subjects were included in the final sample (3-year olds,[2] 37–41 months, $M = 39$, n = 14; 3.5-year olds, 42–47 months,

[2] We included two separate groups of younger and older 3-year-olds since in previous studies and pilot work in our lab, a considerable proportion of older 3-year-olds already passed FB tasks. We thus targeted young 3-year-olds specifically since we wanted to make sure that the youngest age group performs close to floor in FB.

Table 1
Correlations (and partial correlations correcting for age and language ability) between the different tasks in Study 1.

		Standard	Aspectuality	
		TB	FB	TB
Standard	FB	−0.64* (−0.48*)	0.70** 0.75*)	−0.64** (−0.63*)
	TB		−0.77** (−0.70*)	0.64* (.61**)
Aspectuality	FB			−0.86** (−0.87*)

* $p<0.05$.
** $p<0.01$.

M = 44, n = 26; 4-year olds, 48–59, M = 54, n = 26; 5-year olds, 61–70 months, M = 66, n = 20; 6-year olds, 72–85 months, M = 79, n = 25; 8-year olds, 96 to 107 months, M = 102, n = 20; 10-year olds, 122–143, M = 127, n = 22; adults, 21–38 years; M = 26 years; n = 18). Participants came from mixed socioeconomic backgrounds and were recruited from a databank of children whose parents had previously given consent to experimental participation (children) or via recruiting in a teaching class (adults). 14 additional children were tested but excluded from data analysis because they were uncooperative (n = 2), due to insufficient linguistic abilities (n = 2), or due to experimental error (n = 10). Children were tested by either a male or female experimenter in their daycare or in the lab. Adults were tested in the lab and received chocolate for participation.

3.1.2. Design and procedure

In order to overcome the limitations of Study 1, children's and adults' performance in standard and aspectual FB and TB tasks was investigated in a similar design as in Study 1 with the following modifications: task type (standard vs. aspectuality) was used as a between subjects factor, resulting in a 2 (FB/TB) × 2 (standard/aspectuality) design, with the former as within- and the latter as between-subjects factor. Each participant received now two trials per condition and thus received only four trials overall instead of eight (order of FB/TB tasks counterbalanced across subjects). The procedure in the standard and aspectuality tasks remained the same as in Study 1, and children (except for the 10-year olds, for whom the task was not age-appropriate anymore) received the same task of verbal ability (K-ABC). When children failed to answer the control questions correctly the experimenter repeated the test question. If children insisted on their wrong answer in this experiment they were corrected (but, conservatively, their first answer to the control question was used for further analysis and coded as "incorrect").

3.2. Results

Children answered control questions correctly in the following percentages of given trials: Standard FB/TB: Control question 1: 93% correct; Control question 2: 99%; Aspectuality FB/TB: Control question 1: 82%; Control question 2: 95%; Control question 3: 99%. Overall, 81% (n = 118) children answered all control questions correctly. Adults answered all control questions correctly. A closer analysis of control question performance as a function of age revealed the following patterns: 3-year olds (N = 14) performed moderately on Standard FB/TB control questions (success in at least 61% of the trials), competently on Aspectual TB control questions (success in more than 90% of the trials) but poorly on control question 1 in Aspectuality FB (10% correct) while performing moderately on control question 2 and 3 in Aspectuality FB (success in at least 70% of the trials). 3.5 year olds (N = 26) solved control questions in at least 86% of the trials, except for control question 1 in Aspectuality TB (56% correct). 4-year olds (N = 26) also performed worst on Aspectuality TB control question 1 (67% correct), while they solved all other control questions in at least 89% of the trials. 5-year olds (N = 20) showed a similar pattern, solving all but control question 1 in Aspectuality TB (45%) in at least 95% of the trials. All other age-groups (5-, 6-, 8- and 10-year-olds) solved all control questions in at least 90% of the trials.

3.2.1. Consistencies across trials

The consistency in performance of children over trials 1 and 2 of the same type of task was very high for all conditions (Φs > 0.48). Therefore, sum scores of trials answered correctly per condition [0–2] were used for further analyses.

3.2.2. Performance as a function of condition

The mean sum of trials answered correctly as a function of conditions is depicted in Fig. 3. As can be seen from the figure, both for standard change-of-location and for aspectuality tasks, the development of TB performance marks a clear U-shaped curve whereas FB performance shows increase with age. Since adults performed at ceiling with no variance whatsoever, they serve as a validation or reference group but cannot be entered into any inference-statistical analyses. These analyses thus focus on the remaining seven age groups. Since preliminary analyses (a 2 (belief: FB/TB) × 2 (task: standard/aspectuality) × 7 (age groups) × 2 (order) ANOVA on the mean sum of trials correct) failed to find any main or interaction effects for the order (FB-TB vs. TB-FB) of test blocks (all $ps > 0.18$), this factor was skipped from further analyses.

A 2 (FB/TB) × 2 (standard/aspectuality) × 7 (age groups: 3-/3.5-/4-/5-/6-/8- and 10-year-olds) ANOVA on the mean sum of trials answered correctly yielded a main effect of belief type ($F(1,139) = 15.96$, $p < 0.001$, $\eta\ p^2 = 0.10$), a main effect of age ($F(6,139) = 11.07$, $p < 0.001$, $\eta\ p^2 = 0.32$) and no effect of task type (standard/aspectuality) ($F(1,139) = 0.10$, $p = 0.74$). Crucially, there was an interaction effect of belief type (FB/TB) and age ($F(6,139) = 7.02$, $p < 0.001$, $\eta\ p^2 = 0.23$) and no other interaction effect.

3.2.3. Performance as a function of age

To test for a potential age related development we conducted age-related regression analyses for FB and TB. These analyses revealed that children's performance increased with age and that the age-related FB development is best fitted by a linear model ($F(1,77) = 15.00$, $p < 0.01$):. In TB, in contrast, children's performance followed a U-shaped curve and age-related development was best fitted by a quadratic model ($F(2,77) = 10.09$, $p < 0.01$).

To test for children's performance in FB and TB as function of age in more fine-grained ways, post hoc follow-up tests against chance in FB and TB tasks were computed separately for the different age groups. These analyses yielded the following results: for FB tasks, only 3-and 3.5-year olds did not perform above chance (3-year olds, $t(13) = -1.88$, $p = 0.08$; 3.5-year olds, $t(25) = 0$, $p = 1$) while all other age groups did so (4-year olds, $t(25) = 3.64$, $p < 0.01$, $d = 0.71$; 5-year olds, $t(19) = 6.66$, $p < 0.01$, $d = 1.49$; 6-year olds, $t(24) = 3.65$, $p < 0.01$, $d = 0.73$, 8-year olds, $t(19) = 4.77$, $p < 0.001$, $d = 1.07$ and 10-year olds, $t(22) = 10.00$, $p < 0.001$, $d = 2.13$).

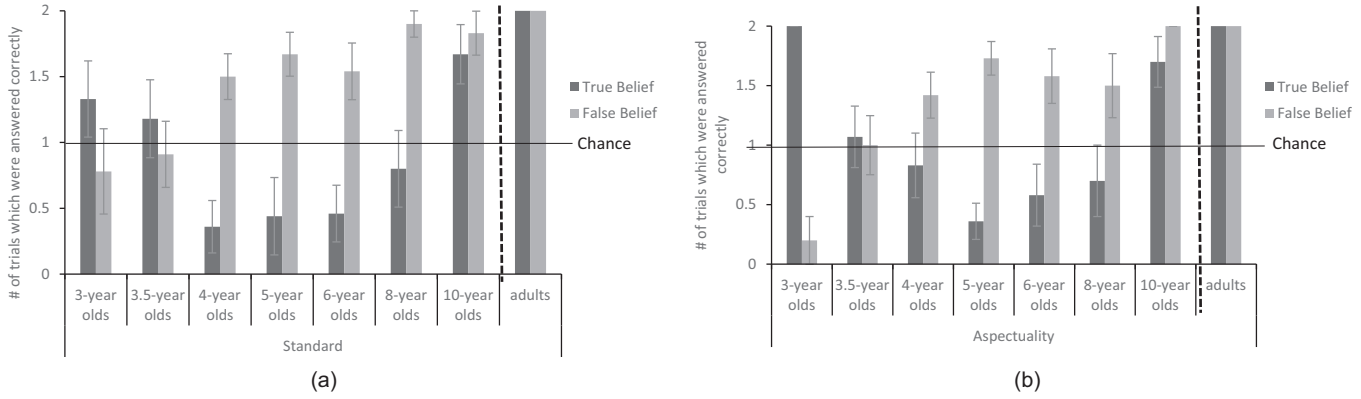

Fig. 3. Mean number of trials answered correctly in (a) the standard and (b) the aspectuality FB/TB tasks as a function of age group.

TB performance, in contrast, revealed a rather different (U-shaped) pattern: 3- and 10-year olds performed significantly above chance (3-year olds, $t(13) = 2.83$, $p < 0.05$, $d = 0.61$; 10-year olds, $t(21) = 4.47$, $p < 0.001$, $d = 0.95$), 3.5- and 8-year olds at chance (3.5-year olds, $t(25) = 0.40$, $p = 0.69$ and 8-year olds, $t(19) = -1.23$, $p = 0.23$), and 4-, 5- and 6-year olds performed below chance (4-year olds, $t(25) = -2.52$, $p < 0.05$, $d = -0.50$; 5-year olds, $t(19) = -3.94$, $p < 0.01$, $d = -0.88$ and 6-year olds, $t(24) = -2.92$, $p < 0.01$, $d = -0.58$).

3.2.4. Correlations between tasks

Across all age groups, TB and FB tasks were negatively correlated – both in terms of raw and partial correlations (see Table 2). Separate analyses as a function of age groups suggests that these correlations were mainly driven by the 3- to 6-year-olds (with the exception of the 5-year olds).

3.3. Discussion

The main findings of Study 2, replicating and extending those of Study 1, were the following: first, children performed on comparable levels, on standard change-of-location and aspectuality FB tasks, and the same was true for the two types of TB tasks. Second, TB performance (both for standard and for aspectuality tasks) followed a U-shaped curve such that 3-year-olds and children from age 10 performed competently, with children in between failing. In FB tasks (both for standard and for aspectuality tasks), in contrast, performance increased with age such that children younger than 4 failed while children from 4 passed. Third, FB and TB performance was negatively correlated until the age of 8–10 (when children began to master both types of tasks).

These results are very much in line with the predictions of performance limitations accounts (and not readily explainable by either of the two competence limitation accounts). Taken by themselves, however, they remain somewhat indirect. More direct evidence would be desirable from studies that manipulate the alleged performance factors, showing that children's failure in TB tasks (and the negative TB-FB correlations) can be alleviated once the relevant task demands have been removed. Study 3 was designed to test for such evidence.

4. Study 3

The rationale of Study 3 was to test for children's TB and FB performance in novel tasks in which the TB versions are less affected by potential performance factors. One prime candidate for the unnecessary complexity of the TB tasks used previously and in Studies 1 and 2, is the lack of relevance or salience of the protagonist's TB (nothing belief-relevant happens in these scenarios, so why should one pay attention to or care about the protagonist's epistemic situation?). Another one is pragmatic oddity (why would one ask such trivial test questions about an agent's beliefs and actions if there is no point in talking about beliefs since the possibility of mistake has not even been raised?).

In order to remove or at least reduce these potential performance factors, we devised tasks (both standard change-of-location and aspectuality) with two protagonists one of whom failed to witness a crucial event and thus had a false belief while the other one had full perceptual access and thus true beliefs. The basic idea is that in this context, the contrast between one agent's FB and the other one's TB makes the TB much more salient and relevant. And from a pragmatic point of view, asking about the TB of an agent – given the contrast to the other agent's FB and the fact that this other agent brings into play the possibility of mistake and thus a motivation for belief-talk- is now much less trivial and thus confusing (for similar preliminary findings that adding a second protagonist may help to make FB-ascription more salient and relevant, see Lewis, Hacquard, & Lidz, 2012; Pham, Bonawitz, & Gopnik, 2012).

Table 2
Correlations (and partial correlations correcting for age and language ability) between TB and FB overall and as a function of age group and task type in Study 2.

	Overall	Standard	Aspectuality
Correlations TB – FB			
All children	−0.42** (−0.54)**	−0.40** (−0.50)**	−0.45** (−0.58)**
3-year olds	−0.55* (−0.89)*	−0.50 (−0.80)*	n/c
3.5-year olds	−0.82** (−0.81)**	−0.71* (−0.77)*	−0.89** (−0.87)**
4-year olds	−0.43** (−0.34)	−0.71* (−0.58)	−0.17 (−0.04)
5-year olds	−0.10 (−0.11)	0.19 (−0.21)	0.04 (0.00)
6-year olds	−0.74** (−0.70)**	−0.72* (−0.69)*	−0.78* (−0.65)*
8-year olds	0.04 (0.06)	0.31 (0.06)	−0.07 (−0.07)
10-year olds	−0.10 (−0.06)[a]	−0.14 (−0.39)[a]	n/c

n/c - not computable due to at least one constant variable.
* $p<0.05$.
** $p<0.001$.
[a] Only controlled for age.

The underlying reasoning and prediction from the point of view of the performance limitation account is the following: If children from around age 4 have the meta-representational capacity to ascribe beliefs (including aspectual beliefs), both true and false, to agents, and if this competence is masked in some TB tasks by pragmatic or other performance factors, then removing these factors (by making the tasks less pragmatically confusing, more relevant, etc.) should have the following effects: children from around age 4 should now master different versions of FB and TB tasks in much the same way; that is, performance in FB should be as proficient as in Study 2; but performance in TB should be significantly better than in Study 2; negative correlations should disappear, and the tasks should be positively correlated instead. For 3-year-old children who do not yet have the competence to operate with fully-fledged belief concepts, though, these manipulations will have little effect (they will continue to solve TB and fail FB tasks, and the tasks will remain negatively correlated).

4.1. Method

4.1.1. Participants

101 children were included in the final sample (3-year olds, age = 37–47 months, M = 44, n = 20; 4-year-olds, age = 48–59 months, M = 53, n = 41 and 6-year olds, age = 73–83, M = 78, n = 40). Children came from mixed socioeconomic backgrounds and were recruited from a databank of children whose parents had previously given consent to experimental participation. 4 additional 4-year olds were tested but excluded from data analyses because they were uncooperative (n = 3) or due to experimental error (n = 1). Children were tested by a female experimenter either in a quiet room of their day care or in the laboratory.

4.1.2. Design and procedure

The basic design was a 2 (belief: TB-FB) × 2 (condition: standard-aspectuality) within-subjects design. Each child received four trials in total, two trials of standard change-of-location tasks and two trials of aspectuality tasks, with each trial containing TB and FB questions. The order of the tasks as well as which protagonist was holding the TB/FB was counterbalanced across subjects. Again, the vocabulary subscale of the Kaufmann Assessment Battery for Children was used to measure children's verbal ability. The same tasks as in Studies 1 and 2 (standard and aspectuality) were used, with the following modification: instead of one protagonist per trial holding either a FB of a TB, we introduced two protagonists per trial, of which one was holding a FB and the other one holding a TB. This was realized as described below.

4.1.3. Standard task

The standard task differed from the task used in Studies 1 and 2 in the following ways (see Fig. 1): instead of one protagonist, we introduced two protagonists [e.g. ape and horse]. In the presence of both protagonists an object [e.g. ball] was put in a box [box1]. Then one of the protagonists [e.g. the ape] left the situation. In her absence and the presence of the other protagonist [e.g. the horse] the object was then transferred to the other box [box2] and the horse left the situation, too. In the absence of both protagonists the first and second control questions were asked (in cases in which children did not answer correctly, the experimenter explained the relevant part of the story to them again and corrected them).

- Control Question 1: Who was present when we transferred the object from box 1 to box 2? [correct answer: the horse]
- Control Question 2: What about the other one? Was she present? [correct answer: no]

Then both of the protagonists returned and the following control and test questions were asked:

- Control Question 3: Where did we put the object in the beginning? [correct answer: box 1]
- Control Question 4: Where is the object now? [correct answer: box 2]
- Test Question 1: What does the horse think where the object is? [depending on her belief]
- Test Question 2: What does the ape think where the object is? [depending on her belief][3]

4.1.4. Aspectuality task

The basic logic of the different aspectuality tasks (FB/TB) did not differ from the ones used in Studies 1 and 2. The difference was again that we used two protagonists within a trial, who left the scene at different time points: the one holding the FB left the scene before the dual identity of the object was revealed; the one holding the true belief left the scene after learning the dual identity. The test and control questions in the aspectuality task were the same as in the standard task with the following differences in the first and second control question:

- Control Question 1: Who knows that the A [e.g. pen] is also a B [e.g. rattle]? [correct answer: depending on who was holding the TB]
- Control Question 2: What about the other one? Does she know? [correct answer: no]

4.2. Results

4.2.1. Control questions

The percentages of children who spontaneously answered the different kinds of control questions correctly and thus needed no feedback is depicted in Table 3.

62% of all children (N = 63) answered all control questions correctly. While 95% of the 6-year olds (N = 38) and 57% of the 4-year olds (N = 24) did so, only one 3-year old answered all control questions correctly.

4.2.2. Main analyses (whole sample)

4.2.2.1. Consistency across trials. The consistency in performance of children over trials 1 and 2 of the same type of task and belief was high for all conditions ($\Phi s > 0.34$). Therefore, sum scores of trials answered correctly per condition [0–2] were used for further analyses.

4.2.2.2. Performance as a function of condition. The mean sum scores of trials in which children answered TB questions correctly and the sum scores of trials in which they answered FB questions correctly as a function of age and task type are depicted in Fig. 4.

A 2 (belief type: TB/FB) × 2 (task type: standard change-of-location/aspectuality) × 3 (age group) mixed-factors ANOVA on these mean sum scores of correct trials yielded no main effect of task type ($F(1,98) = 1.76$, $p = 0.10$), a main-effect of belief type ($F(1,98) = 8.23$, $p < 0.01$, $\eta\ p^2 = 0.08$) and a main effect of age-group ($F(1,98) = 9.05$, $p < 0.001$, $\eta\ p^2 = 0.16$. Furthermore there was an interaction effect between belief type and age-group ($F(2,98) = 12.17$, $p < 0.001$, $\eta\ p^2 = 0.20$).

To test for children's competence as a function of task type and age, separate planned comparisons against chance were conducted. These analyses revealed that all age-groups performed significantly above chance on all TBs (3-year olds: standard TB,

[3] Note that test questions 1 and 2 remained always the same. It was counterbalanced whether the horse or the ape was holding the FB.

Table 3
Children's performance on the control questions as a function of questions and age group in Study 3.

% trials correct	Standard task				Aspectuality task			
	CQ1: Presence (%)	CQ2: Other one? (%)	CQ3: Location 1 (%)	CQ4: Location 2 (%)	CQ1: Knowledge (%)	CQ2: Other one? (%)	CQ3: Location 1 (%)	CQ4: Location 2 (%)
3-year-olds	63	90	55	55	53	63	93	60
4-year-olds	99	96	88	91	89	93	96	96
6-year-olds	100	100	98	98	100	100	100	100

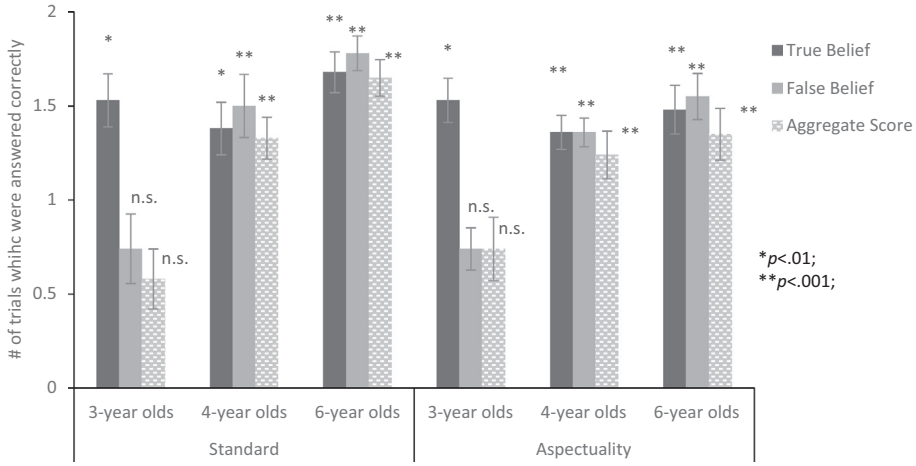

Fig. 4. Mean number of trials in which TB and FB questions were answered correctly and aggregate scores as a function of age and task type. [note that the chance level of guessing correctly differed between TB/FB (chance level = 50%, i.e. 1) and the aggregate score combining both measure (chance level = 25%, i.e. 0.5)].

M = 1.45, $t(19) = 2.93$, $p < 0.01$, $d = 0.65$ and aspectuality TB, M = 1.50, $t(19) = 3.68$, $p < 0.01$, $d = 0.82$; 4-year olds: standard TB, M = 1.41, $t(40) = 3.96$, $p < 0.001$, $d = 0.61$ and aspectuality TB, M = 1.37, $t(40) = 3.06$, $p < 0.01$, $d = 0.48$ and 6-year olds: standard TB, M = 1.67, $t(39) = 7.46$, $p < 0.001$, $d = 1.17$ and aspectuality TB, M = 1.47, $t(39) = 3.68$, $p < 0.001$, $d = 0.58$).

On FB, however, 3-year olds did not perform above chance in both task types (standard FB, M = 0.80, $t(19) = -1,07$, $p = 0.30$ and aspectuality FB, M = 0.75, $t(19) = -1,56$, $p = 0.14$) while 4- and 6-year olds did so (4-year olds: standard FB, M = 1.49, $t(40) = 5.23$, $p < 0.001$, $d = 0.82$ and aspectuality FB, M = 1.37, $t(39) = 3.57$, $p < 0.01$, $d = 0.50$; 6-year olds: standard FB, M = 1.78, $t(39) = 10.22$, $p < 0.001$, $d = 1.63$ and aspectuality FB, M = 1.55, $t(39) = 4.44$, $p < 0.001$, $d = 0.70$).

4.2.2.3. Correlations. The correlations of TB and FB for the different tasks and age-groups are depicted in Table 4. As can be seen from the table, FB and TB performance was highly correlated for the 4- and 6-year-olds ($rs > 0.54$) but not for 3-year olds ($rs < 0.3$).

4.2.3. Aggregate scores analyses

In a second analysis, we computed aggregate scores that took into account whether children solved both TB and FB within a given trial. A trial only received an aggregate score "correct" if children answered both TB and FB in this trial correctly (with a chance level of guessing correctly of 1/4). The sum aggregate scores as a function of condition and age group are depicted in Fig. 4.

A 2 (task type: standard/aspectuality) × 3 (age group) ANOVA on these mean aggregate scores revealed that there was no main effect of task type ($F(1,98) = 0.71$, $p = 0.40$), but a main effect of age ($F(2,98) = 12.34$, $p < 0.001$). Post-hoc Tuckey-B tests revealed that this was due to the fact that 3-year olds performed worse than 4-year olds ($p < 0.01$) and 6-year olds ($p < 0.001$), while 4- and 6-year olds' performance did not differ ($p = 0.26$).

Table 4
Correlations (and partial correlations correcting for age and language ability) between TB and FB versions of a given task in Study 3.

	Standard FB/TB	Aspectuality FB/TB
All children	0.47* (0.43)*	0.50* (0.53)*
3-year olds	0.18 (0.10)	−0.05 (0.10)
4-year olds	0.59* (0.56)*	0.78* (0.56)*
6-year olds	0.75* (0.73)*	0.54* (0.73)*

* $p < 0.001$.

Post-hoc tests against chance showed that 3-year olds did not perform above chance (standard, M = 0.55, $t(19) = 0.33$, $p = 0.75$ and aspectuality, M = 0.70, $t(19) = 1.22$, $p = 0.24$) while 4- and 6-year olds did so in both the standard and aspectuality task (4-year olds: standard, M = 1.37, $t(40) = 7.94$, $p < 0.001$, $d = 1.25$ and aspectuality, M = 1.27, $t(40) = 6.10$, $p < 0.001$, $d = 0.95$ and 6-year olds: standard, M = 1.65, $t(39) = 11.69$, $p < 0.001$, $d = 1.85$ and aspectuality, M = 1.35, $t(39) = 6.22$, $p < 0.001$, $d = 0.98$).

Aggregate scores for the standard and the aspectuality tasks were correlated ($r = 0.40$) even if controlled for age and verbal ability ($r = 0.32$).

4.2.4. Control analyses (only the sub-sample with correct control questions)

Since the present task was rather taxing in terms of memory demands, in particular for the younger children, a substantial number of 4-year-olds, and even the majority of 3-year-olds answered at least one control question incorrectly. Therefore, in a secondary more conservative control analysis, these children were removed

from the analyses. These control analyses on the remaining subsample of children answering all control questions correctly (one 3-year-old, 24 4-year-olds, and 38 6-year-olds) largely replicated the results of the main analyses for the 4- and 6-year-olds (given the sample size of n = 1 of the remaining 3-year-olds, this age group could not be included in the control analyses) (for details, see Appendix A).

4.2.5. Complementary analysis: comparison between Study 3 and Study 2

In order to test whether the removal of the potential performance factors in Study 3 made a crucial difference to TB (but not to FB) performance, we compared the FB and TB performance of the 4-and 6-year-olds across between Study 3 and Study 2. These analyses revealed that the 4-year olds in Study 3 outperformed those in Study 2 in TB ($t(53) = 4.86$, $p < 0.001$; $d = 1.46$), but not in FB ($t(53) = 0.19$, $p = 0.85$). The same was true for 6-year olds who performed better in Study 3 than in Study 2 in TB ($t(51) = 6.01$, $p < 0.001$, $d = 1.78$ but did not differ in their FB performance ($t(51) = 1.31$, $p = 0.20$).

4.3. Discussion

The main results of Study 3 were the following: First, the modified TB version was much easier than the previous versions: 4- and 6-year olds performed competently on the present TB tasks and significantly better than they did in Study 2. Second, children's performances on FB and TB of the different tasks were now positively correlated, with strong convergence between tasks. Third, 3-year olds performance remained largely unchanged (although these findings remain somewhat difficult to interpret given the poor performance on control questions), in the sense that they performed competently on TB but still failed FB tasks. Taken together, these findings are thus clearly in line with the predictions of performance limitation accounts.

5. General discussion

5.1. Summary of the main findings

The aim of the present study was to investigate the development of FB and TB performance systematically and comprehensively, thereby testing different theoretical accounts against each. Empirically, the main findings were the following: First, Studies 1 and 2 taken together revealed that performance in different kinds of FB tasks (standard change-of-location and aspectuality) develops analogously and in strongly consistent and correlated fashion, as does performance in the TB versions of these tasks. Second, FB performance increased with age, whereas TB performance followed a clear U-shaped curve. Third, FB and TB performance were strongly negatively correlated between the ages of 3 and 6 (before children began to then reliably master both TB and FB tasks from age 8 to 10). Fourth, Study 3 showed that both children's difficulty with TB tasks and the inverse relation of FB and TB vanish once the TB tasks were suitably modified.

5.2. Theoretical implications: competence versus performance limitation accounts

Theoretically, the present studies aimed at testing competence limitations accounts concerning children's difficulty with TB tasks against performance limitation accounts. Competence limitation accounts argue that children's failure in TB tasks (i) reveals a true competence limitations, (ii) shows that they do not yet operate with a fully-fledged concept of belief – even if they seem to in FB tasks- and (iii) implies that the standard picture according to which children acquire truly meta-representational concepts of propositional attitudes around age 4 is mistaken. Two kinds of competence limitation accounts predict the kinds of patterns found here – U-shaped curve in TB and negative TB-FB correlations—for some classes of TB tasks. The Perceptual Access Reasoning (PAR) account predicts such patterns for TB tasks in which the protagonist lacks perceptual access to some relevant step in the course of events. The Mental File Card account predicts such patterns for a specific class of aspectual TB tasks (in which the target object can be referred to under two descriptions A and B, where the protagonist does (TB) or does not know (FB) about the A = B identity; Perner et al., 2015). Performance limitation accounts, in contrast, assume that children's failure in TB tasks constitute false negatives: Children from age 4 really do operative with a fully-fledged concept of belief, yet fail TB tasks due to extraneous tasks demands, most likely having to do with the pragmatics of the task.

The present findings make a clear case against the two competence limitation accounts under consideration, and for performance limitation accounts. The findings of Studies 1 and 2 are not compatible with the Mental File Card account which would only predict the patterns found here (U-curve in TB and negative FB-TB correlations) for aspectual but not for standard change-of-location tasks. Nor are they compatible with the PAR account since the PAR analysis does not even apply to the TB tasks used here: in our tasks, there was no relevant event that the protagonist failed to witness (she simply left the room before the crucial events–this was done in order to match FB and TB tasks superficially - but then nothing happened in her absence, and after her return she witnessed and thus had full perceptual access to the relevant events). Thus, for the TB tasks used here it was not the case that there was anything like "comparable lack of perceptual access" (Hedger & Fabricius, 2011, p. 432) relative to FB cases, and thus the PAR account would not predict that children find our TB tasks difficult at all. Now, one might modify the PAR account such that it does predict difficulty in the present TB tasks as well. One might, for example, add the extra premise that *any* kind of absence of the protagonist at any time in the scenario is sufficient for attributing lack of perceptual access to her and thus assuming she will act incorrectly. But such a move would not only amount to a fundamental revision of the PAR account, but seems utterly ad hoc and without any independent motivation or plausibility. In any case, however, the PAR account is even more clearly refuted by the results of Study 3 in which everything concerning the protagonist acting on her TB remained the same in terms of perceptual access.[4] The only modification was that in the same scenario there was now also another protagonist acting on her FB. This modification should make no difference whatsoever from the theoretical point of view of the PAR account (yet it should make potentially a big difference from the perspective of performance limitation accounts since the true belief of the TB protagonist, in its contrast to the false belief of the other protagonist, is now much more salient and relevant, and the

[4] Actually, the TB condition in Study 3 is the only one in the present studies for which a PAR analysis, with some additional assumptions, seems remotely plausible at all. Remember that in Studies 1–2, the protagonist left before the crucial events and then witnessed all subsequent events up to the test question. Since, however, in Study 3, the TB agents leaves the room *after* the crucial transfer of the object and re-enters right before the test questions, one could argue that the protagonist now enters a new perceptual situation and thus suffers from "interrupted perceptual access", and accordingly, PAR reasoners should predict that the protagonist will get things wrong and thus fail this task (for such an argumentation, see Hedger & Fabricius, 2011). That is, if the PAR account can be applied to the present studies at all, then at most to Study 3. Overall, the account would then predict the following pattern of results in the three studies: The TB tasks in Study 3 should be difficult, but those in Studies 1–2 should pose no problems. In fact, however, the empirical pattern that was found is the exact reverse – and thus even less compatible with the PAR account than each finding taken by itself.

TB test question is now much less pragmatically confusing). But empirically it did make a fundamental difference. Similarly, from the perspective of the Mental File Card account, everything in Study 3 remains the same concerning the TB protagonist in terms of the file card demands (coordination of horizontal and vertical linking) and thus the results should be the same as in the TB conditions of Studies 1 and 2.

Taken together, the results of Studies 1–3 thus clearly speak against the two competence limitation accounts tested here, and for some kind of performance limitation account. Three challenges, in our view, will need to be addressed by future research in this area: First, though the present findings are incompatible with existing competence limitation accounts, may there be some other future competence limitation account suitable for explaining the present patterns of findings with TB tasks? We have to admit that we have no idea what such an account might look like, but it should be noted that the present findings do not strictly rule out the possibility of alternative competence limitation accounts. Second, can converging evidence for performance and against competence limitation accounts be found with different methodological approaches? One complementary line of such research could test competence vs. performance accounts further by using completely non-verbal tasks that are stripped of task pragmatics altogether and in which keeping track of another agent's true of false belief is relevant for strategic decision-making and not just for answering test questions. For example, Call and Tomasello (1999) and Kaminski, Call, & Tomasello, 2008 have devised such task for apes in which subjects gamble with or against another player and have to base their decisions on what they belief the other believes. These tasks could be adapted for use in analogous FB and TB conditions with children: While competence limitation accounts predict the same patterns of performance as in verbal FB/TB tasks, performance limitation accounts would predict that such nonverbal TB tasks with increased relevance and decreased pragmatic performance factors should pose no problems. Third, if some version or other of performance limitation accounts is correct – as strongly suggested by the present data– what exactly are the limiting performance factors that mask children's competence in TB tasks?

5.3. Which type of performance limitation account?

In our view, two kinds of – potentially complementary – candidates are the most plausible performance factors responsible for the paradoxical difficulty of TB tasks: The true beliefs of the single agent in TB scenarios like those used in Studies 1 and 2 may not be sufficiently salient and relevant for young ToM-reasoners who do have a concept of belief but are still fragile in its application. This idea could well explain why 4- to 6-year-olds have difficulty answering TB questions whereas no such difficulty is found in 3-year-olds (lacking a concept of belief, they answer in reality-based ways) and in children from age 10 (their application of their belief concept has somehow become less fragile and sensitive to superficial task factors such as salience). And it could explain why these difficulties vanish in Study 3 in which the TB agent's true belief is made more salient and relevant. What remains unclear is how this idea should explain the negative correlation between FB and TB tasks in children between the ages of 4 and 6. Why would advanced ToM capacity, as indicated in better FB performance, go along with worse performance in TB tasks? It seems that there is currently no obvious and plausible story following from the relevance/salience idea taken by itself to motivate such a prediction.

However, the second prime candidate for underlying performance factors that make TB tasks difficult –pragmatic ones- in fact does make this very prediction on theoretical grounds. The basic idea is that TB tasks with only one protagonist are pragmatically odd and thus confusing (see Siegal and Beattie (1991), for general explanations along such lines, and Helming et al., 2014; Helming et al., 2016, for some recent application to ToM tasks): Asking a test question about the beliefs and actions of a single protagonist who has not even failed to witness anything, and in the complete absence of the possibility of mistake, is confusingly trivial and in violation of the basic point of belief discourse, namely referring to or at least highlighting the possibility of mistakes. Children with some, yet limited degree of pragmatic capacities, which in turn developmentally depends on their developing ToM, will then get confused by such questions ("why would she ask this?") and assume they must have mis-understood or mis-construed something ("I must have missed something, and she must be somehow wrong after all"). This idea equally predicts all the patterns of results found here in the way the relevance/salience idea does: 4- to 6-year-olds, with some ToM capacities and thus some, yet fragile pragmatic capacities get confused by TB questions whereas no such confusion is found in 3-year-olds (lacking a concept of belief, they are not subject to pragmatic confusion in this way, simply answering in reality-based ways) and in children from age 10 (their pragmatic capacities have developed further and have become less fragile so that they, like adults, can deal also with trivial, pointless and odd test questions). It can explain why the 4- to 6-year-olds' difficulties with TB tasks vanish in Study 3: asking about the TB agent's true belief in light of and in contrast to the other agent's false belief makes this question much less trivial, pointless and confusing. And crucially, it can explain the negative correlations of FB and TB performance in the 4- to 6-year-olds: The more advanced children get in their ToM (as indicated in FB performance), the more pragmatically sensitive they get, and thus the more subject to pragmatic confusion by odd, trivial and pointless questions.

This – at the current stage speculative – interpretation leaves open a number of crucial questions, of course: First, what exactly is it about the TB questions that make them pragmatically confusing? Their triviality and status of test question (asked not in need of information, but in order to test whether the addressee knows the answer) seem to be crucial elements, yet clearly ones that are not sufficient (children this age have no difficult with many other utterly trivial test questions, including, for example, the control questions ("where is the ball now?") used in FB tasks here and everywhere). Quite plausibly, what is an additional ingredient is the fact that the questions are about a rational agent's beliefs and actions. Belief discourse usually has a specific point: we would not talk about someone's belief or predict her belief-based actions unless there was at least a possibility of her being wrong. Children would probably never learn any belief verbs if it was not in such contexts (Papafragou et al., 2007). It may thus be the combination – a question asked as test question that is utterly trivial and misses the main point of the discourse it embodies – makes the TB questions in previous studies and in Studies 1 and 2 here difficult.[5]

Second, why then do older children at some point (in the present case, children from age 10 in Study 2) overcome these pragmatic confusions? Again, at this point we can only speculate. But it is highly plausible to assume that children's pragmatic understanding simply reaches some higher level of complexity and

[5] This possibility is currently being tested in a follow-up study in our lab. In a 2 × 2 design, 4- to 6-year-olds are confronted with FB and TB tasks, and with structurally analogous false photo and true photo tasks (following the false photo task by Zaitchik, 1990). If the triviality of test questions as such was the crucial factor, then children should perform equally poorly in the equally trivial TB and true photo tasks (and equally well in the FB and false photo tasks). However, if it is triviality plus reference to beliefs, then children should perform successfully in FB and false photo tasks, fail TB tasks, but have no difficult with the matched true photo task that involves no reference to beliefs. Preliminary results (with 30 children) clearly speak for the latter pattern.

recursion at some point so that they become able, as adults clearly are, to understand the complex network of nested communicative intentions in test situations in which one can and often does engage in discourse stripped of its normal pragmatic point (see, e.g. Happé (1994) and Sullivan, Zaitchik, and Tager-Flusberg (1994), for similarly protracted development in understanding other types of complex, recursively nested speech acts).

6. Conclusion and future directions

The present studies, to the best of our knowledge, present the most comprehensive and systematic test of children's performance on different kinds of TB tasks to date. The results document clear and consistent developmental patterns (analogous and consistent performance in different kinds of TB tasks; U-shaped curve, negative correlations with FB performance; disappearance of these patterns when performance factors are reduced) that speak very much against existing competence limitation accounts and in favor of a sort of performance limitation account.

Taken by themselves, the present findings are compatible with different versions of performance limitation accounts, including explanations in terms of salience/relevance, and/or in terms of pragmatic performance factors. Which of these factors, by themselves or in combination, is crucial, and thus how exactly the best performance limitation account is to be spelled out in detail will need to be addressed by future research. What will be needed are systematic studies that pit relevance/salience and pragmatic factors against each other. For example, non-verbal FB and TB tasks (see above) could be devised that vary in relevance and salience. Pure pragmatic performance limitation accounts would predict that none of these TB tasks should be difficult (lacking questions they do not invite pragmatic confusion), whereas relevance/salience accounts would predict that difficult still depends on relevance and salience.

Acknowledgements

We would like to thank Lisa Wenzel, Mareike Spengler, Alexander Dieball, Frederike Wehr, Carina Neumann and Sandra Friedl for help with testing and coding. Thank you very much to Marlen Kaufmann and Konstanze Schirmer for the organization of the studies.

Appendix A. Control analyses for Study 3

In this control analysis, only the data of those children (N = 63, M = 68 months, 32 female) were included who answered all control questions correctly. The consistency in performance of children over trials 1 and 2 of the same type of task and belief was high for all conditions (Φs > 0.42). Therefore, again sum scores of trials answered correctly per condition [0–2] were used for further analyses.

The mean sum scores of trials in which children answered TB questions correctly and the sum scores of trials in which they answered FB questions correctly as a function of task type are depicted in Fig. A1.

A 2 (belief type: TB/FB) × 2 (task type: standard change-of-location/aspectuality) repeated measures ANOVA on these mean sum scores of correct trials yielded a main effect of task type ($F(1,62) = 4.84$, $p < 0.05$, $\eta\ p^2 = 0.08$), no main-effect of belief type ($F(1,62) = 1.43$, $p = 0.24$) and no interaction ($F(1,62) = 0.356$, $p = 0.55$). To test for children's competence as a function of task type separate planned comparisons against chance were conducted. These analyses revealed that these children performed significantly above chance on all TBs (standard TB, M = 1.63, $t(62) = 8.74$, $p < 0.001$, $d = 1.09$ and aspectuality TB, M = 1.46, $t(62) = 4.69$, $p < 0.001$, $d = 0.59$) and FBs (standard FB, M = 1.73, $t(62) = 12.02$, $p < 0.001$, $d = 1.52$ and aspectuality FB, M = 1.51, $t(62) = 5.31$, $p < 0.001$, $d = 0.67$). Furthermore FB and TB performance overall was highly correlated for both tasks (standard, $r = 0.74$, controlled for age and language ability, $r = 0.72$ and aspectuality, $r = 0.58$, controlled for age and language ability, $r = 55$

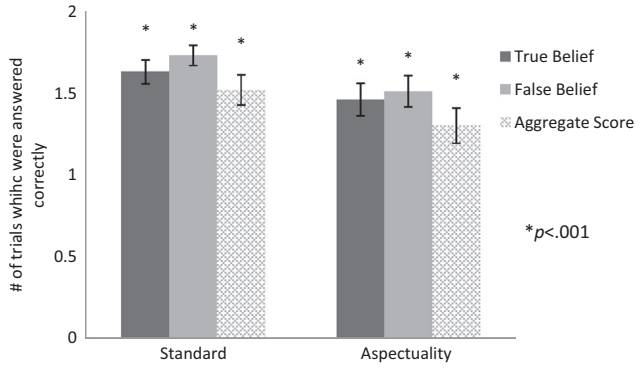

Fig. A1. Mean number of trials in which true and FB questions were answered correctly and aggregate scores as a function of task type. [note that the chance level of guessing correctly differed between TB/FB (chance level = 50%, i.e. 1) and the aggregate score combining both measure (chance level = 25%, i.e. 0.5)].

We computed aggregate scores that took into account whether children solved both TB and FB within a given trial. The sum aggregate scores (of trials in which children answered both TB and FB questions within a given trial) as a function of condition are depicted in Fig. A1 as well. An ANOVA for task type (standard/aspectuality) on these mean aggregate scores revealed that there was a main effect of task type ($F(1,62) = 6.13$, $p < 0.05$, $\eta\ p^2 = 0.09$): children performed better on the standard task than on the aspectuality task.

Post-hoc tests against chance showed that children performed above chance on both of the tasks (standard, M = 1.60, $t(62) = 14.35$, $p < 0.001$, $d = 1.80$ and aspectuality, M = 1.33, $t(62) = 7.85$, $p < 0.001$, $d = 0.99$) (see Fig. A1).

Furthermore, aggregate scores for the standard and the aspectuality tasks were correlated ($r = 0.32$, $p < 0.05$; controlled for age and verbal ability $r = 0.24$, $p = 0.06$).

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.05.002.

References

Apperly, I. A., & Robinson, E. J. (1998). Children's mental representation of referential relations. *Cognition, 67*(3), 287–309.

Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K., . . ., & Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer, P. R. Shaver (Eds), E. Borgida, & J. A. Bargh (Assoc. Eds.), *APA handbook of personality and social psychology, Attitudes and social cognition* (Vol. 1, pp. 79–150). Washington, DC: American Psychological Association.

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology, 67*, 159–186.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110–118. http://dx.doi.org/10.1016/j.tics.2009.12.006.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition, 112*(2), 337–342.

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

Carey, S. (2009). *The origin of concepts*. New York, NY, US: Oxford University Press.

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. http://dx.doi.org/10.1037/a0017648.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science, 13*(2), 150–156.

Frege, G. (1980 [1892]). Über Sinn und Bedeutung. In G. Patzig (Ed.), Funktion, Begriff, Bedeutung (pp. 40–65). Göttingen: Vandenhoeck & Ruprecht.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*, 121–123.

Happé, F. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition, 48*(2), 101–119.

Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129–154. http://dx.doi.org/10.1007/bf02172093.

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. http://dx.doi.org/10.1007/s13164-011-0069-9.

Helming, K., Strickland, B., & Jacob, P. (2016). A pragmatic approach to the puzzle about early belief ascription. *Mind & Language, 31*(4), 438–469.

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*(4), 167–170.

Kamawar, D., & Olson, D. R. (1999). Children's representational theory of language: The problem of opaque contexts. *Cognitive Development, 14*(4), 531–548. http://dx.doi.org/10.1016/s0885-2014(99)00018-0.

Kamawar, D., & Olson, D. R. (2009). Children's understanding of referentially opaque contexts: The role of metarepresentational and metalinguistic ability. *Journal of Cognition and Development, 10*(4), 285–305. http://dx.doi.org/10.1080/15248370903389499.

Kamawar, D., & Olson, D. R. (2011). Thinking about representations: The case of opaque contexts. *Journal of Experimental Child Psychology, 108*(4), 734–746. http://dx.doi.org/10.1016/j.jecp.2010.10.005.

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition, 109*(2), 224–234.

Kaufman, A., & Kaufman, N. (2001). Kaufman Assessment Battery for Children (4th ed.). Frankfurt am Main, Germany.

Lewis, S., Hacquard, V., & Lidz, J. (2012). The semantics and pragmatics of belief reports in preschoolers. *Proceedings of SALT, 22*, 247–267.

McKay, T., & Nelson, M. (2014). Propositional Attitude Reports (Spring 2014 Edition). In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258. http://dx.doi.org/10.1126/science.1107621.

Papafragou, A., Cassidy, K., & Gleitman, L. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition, 105*, 125–165.

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

Perner, J., Huemer, M., & Leahy, B. (2015). Mental files and belief: A cognitive theory of how children represent belief and its intensionality. *Cognition, 145*, 77–88. http://dx.doi.org/10.1016/j.cognition.2015.08.006.

Perner, J., & Leahy, B. (2015). Mental files in development: Dual naming, false belief, identity, and intensionality. *Review of Philosophy and Psychology, 6*. http://dx.doi.org/10.1007/s13164-015-0235-6.

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. *Science, 333*(6041), 474–477. http://dx.doi.org/10.1126/science.1201216.

Pham, K., Bonawitz, E., & Gopnik, A. (2012). Seeing who sees: Contrastive access helps children rason about other minds. In *Proceedings of the thirty-fourth cognitive science society*. .

Rakoczy, H. (2015). In defense of a developmental dogma: Children acquire propositional attitude folk psychology around age 4. *Synthese*, 1–19. http://dx.doi.org/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. http://dx.doi.org/10.1111/cdev.12311.

Recanati, F. (2012). *Mental files*. Oxford University Press.

Russell, J. (1987). 'Can we say Ellipsis?' Children's understanding of intensionality. *Cognition, 25*, 289–308. http://dx.doi.org/10.1016/S0010-0277(87)80007-0.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition, 38*, 1–12.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others referential communication. *Developmental Science, 13*.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.

Sprung, M., Perner, J., & Mitchell, P. (2007). Opacity and discourse referents: Object identity and object properties. *Mind & Language, 22*(3), 215–245. http://dx.doi.org/10.1111/j.1468-0017.2007.00307.x.

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

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(7), 580–586. http://dx.doi.org/10.1111/j.1467-9280.2007.01943.x.

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. http://dx.doi.org/10.1111/1467-8624.00304.

Westra, E. (2016). *Talking about minds: Social Experience, Pragmatic Development, and the False Belief Task*. Unpublished manuscript.

Westra, E., & Carruthers, P. (2016). *The theory-of-mind scale: A pragmatic approach*. Unpublished manuscript.

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

Winner, E., & Gardner, H. (1993). Metaphor and irony: Two levels of understanding. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed.. Cambridge: Cambridge University Press.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition, 35*, 41–68.