Infant Behavior and Development 54 (2019) 13–21

Contents lists available at ScienceDirect

Infant Behavior and Development

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

Full length article

Reliability and generalizability of an acted-out false belief task in 3-year-olds

Sebastian Dörrenberg[a,b,*,1], Lisa Wenzel[a,**,1], Marina Proft[a], Hannes Rakoczy[a], Ulf Liszkowski[b]

[a] Developmental Psychology, University of Göttingen, Germany
[b] Developmental Psychology, University of Hamburg, Germany

ARTICLE INFO

Keywords:
Theory of Mind
Duplo task
Replication
Intensionality
Matched conditions

ABSTRACT

The current study tested the reliability and generalizability of a narrative act-out false belief task held to reveal Theory of Mind (ToM) competence at 3 years of age, before children pass verbal standard false belief tasks (the "Duplo task"; Rubio-Fernández & Geurts, 2013, Psychological Science). We conducted the task across two labs with methodologically improved matched control conditions. Further, we administered an analogue intensionality version to assess the scope of ToM competence in the Duplo task. 72 3-year-olds participated in a Duplo change-of-location task, a Duplo intensionality task, and half of them in a matched verbal standard change-of-location task, receiving either false belief or matched true belief scenarios. Children performed at chance in the false belief Duplo location change and intensionality tasks as well as in the standard false belief task. There were no differences to the standard task, and performance correlated across all three false belief tasks, revealing a rather unified competence and no task advantage. In the true belief conditions of both Duplo tasks, children performed at ceiling and significantly different from the false belief conditions, while they were at chance in the true belief condition of the standard task. The latter indicates that a pragmatic advantage of the Duplo task compared to the standard task holds only for the true belief scenarios. Our study shows that the Duplo task measures the same ToM competence as the standard task and rejects a notion of earlier false belief understanding on the group level in 3-year-old children.

1. Introduction

Theory of Mind (ToM), the ability to attribute mental states to others, is typically tested with false belief (FB) tasks that require the ascription of others' subjective representations of reality that can be false (Wimmer & Perner, 1983). In change-of-location tasks, for example, children see a protagonist put an object in one of two boxes. In the protagonist's absence, the object is then transferred to another box and children are asked to predict where she will look for it. These standard verbal tasks are mastered from age four while young 3-year-olds typically fail (by predicting that the protagonist will look for her object where it really is) and group performance at three-and-a-half years is typically at chance (Wellman, Cross, & Watson, 2001). Since these tasks require advanced pragmatic understanding of language and test questions, several studies have lowered these demands and found enhanced performance at

* Corresponding author at: Department of Developmental Psychology, University of Hamburg, Von-Melle-Park 5, 20146, Hamburg, Germany.
** Corresponding author at: Department of Developmental Psychology, University of Göttingen, Waldweg 26, 37073, Göttingen, Germany.
E-mail addresses: sebastian.doerrenberg@uni-hamburg.de (S. Dörrenberg), lisa.wenzel@uni-goettingen.de (L. Wenzel).
[1] Shared first authorship.

https://doi.org/10.1016/j.infbeh.2018.11.005
Received 29 March 2018; Received in revised form 21 September 2018; Accepted 19 November 2018

slightly younger ages (e.g., Mitchell & Lacohée, 1991; Psouni et al., 2018; Rhodes & Brandone, 2014; Sullivan & Winner, 1993; for a meta-analytic finding, see Wellman et al., 2001). A recent study employed the "Duplo task" (Rubio-Fernández & Geurts, 2013), and reduced demands by minimizing disruptions during the perspective tracking process. Using a narrative version of the change-of-location FB paradigm, children were prompted to act out the protagonist's action, instead of answering to an experimenter's explicit test question about the protagonist's belief. Furthermore, the protagonist remained visible throughout the narrative to facilitate keeping track of her perspective. With these variations, 3-year-olds performed in the Duplo task above chance while still failing a FB task with an explicit test question. The current study sought to test (i) the robustness of the finding through a multi-lab replication approach and by implementing methodological improvements, and (ii) the scope and unity of early ToM competence by administering an intensionality version of the Duplo task, and testing for correlated performance across tasks (see Rakoczy, Bergfeld, Schwarz, & Fizke, 2015).

In the original Duplo task, Rubio-Fernández and Geurts (2013) introduced two main modifications to the standard false belief (SFB) task to facilitate children's perspective tracking. First, children could undisturbedly keep track of the agent's knowledge access during the whole story: The protagonist did not leave the scene at all, but turned her back towards the scenery, so that she was unaware about the object's transfer, but still visible to the child throughout this procedure. Additionally, two prompts about her knowledge access (e.g. "She hasn't seen what I did, did she?") should help children to keep track of the protagonist's perspective. Second, they used a narrative story structure in which children were not asked explicitly where the protagonist would search for the object, but they were involved actively and encouraged to act out the story ("What happens next? You can take the girl yourself if you want… What is she going to do now?"). While 80% of children passed the novel Duplo task by acting out the belief-congruent ending (moved the Duplo girl to the container where she previously left her banana), the same proportion of children failed a standard unexpected-content task (Hogrefe, Wimmer, & Perner, 1986). In two follow-up experiments, Rubio-Fernández and Geurts (2013) showed that these two modifications crucially led to the increased performance in 3-year-old children. If the protagonist disappeared from the scenery during the location change, or if children were asked an explicit test question that mentioned the desired object instead of being given the actively engaging prompts, children's performance decreased to below chance. In another study (Rubio-Fernández & Geurts, 2016), the same authors investigated the impact of two further task modifications. A modification of the test question ("Where will Lola go now?" instead of "What happens next?"), which may highlight the binary choice between the two locations and thus increase attention towards the box containing the object, had no effect on children's increased performance in the Duplo FB task. A stressed salience of the target object, however, decreased children's performance: When mentioning the object, either in a control question after the transfer ("Where are the *bananas* now?"), or right before children were prompted to take the lead ("Now Lola is very hungry and wants a *banana*."), 3-year-old children failed to solve the Duplo FB task. Given the set of modifications implemented in the Duplo task and their implications on children's performance to pass or fail FB tasks, it becomes clear that 3-year-olds' perspective tracking skills are still fragile and can be both enhanced and disrupted quite easily by subtle but crucial factors.

Findings of 3-year-old children passing simplified FB tasks have been taken to support early competence accounts and to suggest that extraneous task demands mask false belief understanding in younger children (Carruthers, 2013; Leslie, Friedman, & German, 2004; Scott & Baillargeon, 2017). While the interpretation has far-reaching implications for the origins and nature of ToM competence, it is important to first assess the validity and robustness of the empirical findings by replicating the original studies. Concerning the Duplo task, to date, there are two published studies that conducted conceptual replications with several (some substantial) modifications of the original protocol, one of which comes from an independent lab (Białecka-Pikul, Kosno, Białek, & Szpak, 2018; Rubio-Fernández & Geurts, 2016). These studies found enhancing effects and above chance performance in their Duplo task versions compared to standard verbal FB tasks. The task by Białecka-Pikul et al. (2018) retained only some minor elements of the Duplo task (e.g. the knowledge access prompts), but introduced a more interactive 'we-mode' (child and experimenter jointly tricked the agent). In contrast to the original version, children were asked an explicit test question and the target object was mentioned. These are both factors which should have decreased performance on the task according to Rubio-Fernández and Geurts (2013, 2016), thus questioning the validity of the original task manipulation. A more direct replication of the Duplo task (Kammermeier & Paulus, 2018) revealed a facilitating effect of the Duplo task compared to a verbal FB task, but in contrast to the original findings, 3-year-olds performed only at chance in the Duplo replication task. This finding questions the reliability of the original above chance finding in the Duplo task and calls for clarification.

Unfortunately, there were also some methodological limitations to the original Duplo task itself. First, the true belief (TB) control condition was not matched to the FB condition. In contrast to the FB condition, in the TB condition the Duplo girl moved the banana herself, and was never distracted from the events, which makes the TB demands easier and prone to a leaner interpretation like a simple agent-object association (e.g., Josef Perner & Ruffman, 2005). Second, for the benchmark comparison to the explicit verbal ToM competence, the authors used an unexpected-content task which does not match the change-of-location structure of the Duplo task (this was the same in Białecka-Pikul et al., 2018). While key manipulations of the Duplo task concern perspective tracking and mentioning of the target object, unexpected-content tasks actually do not involve comparable perspective tracking (there is no agent), and do not bias children to the wrong container via test question (the target object is not mentioned in the test question, "What will XY think is in the box?"). This makes the task less ideal for the investigation of these manipulations. Further, from a conceptual point of view, although performance in change-of-location and unexpected-content tasks correlate (Gopnik & Astington, 1988), the cognitive demands differ between tasks. For example, unexpected content tasks may be more difficult than location-tasks, because they are even more language-dependent, provide less supportive story context, and require transferring one's own FB to another person. From an empirical point of view, there is indeed evidence that unexpected-content tasks are more difficult than change-of-location tasks (Gopnik & Astington, 1988; Holmes, Black, & Miller, 1996).

Similarly, the direct replication study of the Duplo task (Kammermeier & Paulus, 2018) did not use an adequately matched

change-of-location FB task for performance comparison. In their task, children were only told that the protagonist thought his mittens were in a closet even though they were in his backpack, but they did not see the actual location change (Wellman & Liu, 2004). Meta-analytic findings have shown that at three-and-a-half years (which is the mean age group of participants in the original Duplo task study and in all replication studies), children perform at chance (with about 50% passing rate) in verbal FB tasks (Wellman et al., 2001). Compared to this, children performed rather poorly in the verbal location FB task in the replication study by Kammermeier and Paulus (at age 3: below chance, at age 4: at chance), rendering it possible that the facilitating effect of the Duplo task was rather an artefact of the poor performance on the non-matched SFB task.

Taken together, methodological differences make it thus difficult to interpret the performance in the Duplo FB task relative the TB condition and the employed verbal FB tasks. It remains to be tested, how the findings from the Duplo task compare to more closely matched control conditions. A recent study on FB understanding in a word learning context addressed the issue of miss-matched conditions (Papafragou, Fairchild, Cohen, & Friedberg, 2017). When the communicative and the non-communicative FB conditions were matched exactly, earlier findings of an advantage of word learning tasks (Carpenter, Call, & Tomasello, 2002; Happe & Loth, 2002) could not be reproduced. This highlights the importance of comparable controls.

Another important question is what the Duplo task actually measures. After the age of four, children master a variety of explicit ToM tasks in converging fashion. They not only solve classical change-of-location FB paradigms, but also solve unexpected-content, appearance-reality and intensionality tasks (measuring the understanding that someone has a false belief about different aspects of an object), and there is inter-task coherence (Gopnik & Astington, 1988; Josef Perner & Roessler, 2012; Rakoczy et al., 2015). Thus, 4-year-olds seem to have a fully-fledged, unified and flexible ToM competence. This convergence does not apply to implicit ToM tasks that use non-verbal measures. Here, different tasks are not equally solved, but show a dissociation, i.e. children master implicit change-of-location tasks, but fail implicit intensionality tasks (Fizke, Butterfill, van de Loo, Reindl, & Rakoczy, 2017; Low & Watts, 2013; Oktay-Gür, Schulz, & Rakoczy, 2018). Implicit ToM tasks may therefore tap into an earlier developing and efficient mind-reading system, which might be the developmental basis for later flexible ToM, as proposed by two-systems accounts (Apperly & Butterfill, 2009; Low, Apperly, Butterfill, & Rakoczy, 2016). This early ToM system may, according to two-systems views, only be capable of tracking belief-like states or simpler perceptual registrations (what someone saw or did not see), and exhibit specific signature limits, such as the ascription of false beliefs about aspects of objects. For mastering the Duplo task, however, it is unclear whether 3-year-olds utilize a fully-fledged and unified ToM competence that is camouflaged in classical tasks, or whether their ToM competence is immature and exhibits limitations.

Against this background, the rationale of the present study was to test the robustness and reliability of the performance enhancing factors implemented in the Duplo task, and the scope of the underlying ToM capacity. In order to check for robustness and reliability, we tested 3-year-olds in a change-of-location version of the Duplo task and implemented further methodological changes. In contrast to the original study, we conducted a TB condition that closely matched the FB condition and was equated in terms of performance factors. Therefore, both conditions equally required children to track the protagonist's mental state. The only difference between conditions was that in the FB condition the protagonist turned around somewhat later and thus did not witness the crucial event (location change). In addition, we implemented the feature of joint trickery, in order to increase children's involvement throughout the story (Białecka-Pikul et al., 2018; Sullivan & Winner, 1993). Instead of an unexpected-content task as benchmark comparison, we administered a change-of-location task, which matched the features of the Duplo task, but was different in two crucial respects. First, we used an explicit test question as dependent measure (mentioning the desired object), instead of encouraging children to act out the story. Second, rather than turning his back at the scenery during the location change, the protagonist left the scene, as in standard false belief tasks. Following the Duplo task logic, and assuming the task to be robust and reliable, we should find correct above chance performance in the FB condition of the Duplo task, and a differential choice pattern between TB and FB conditions. Further, children should perform differently and less competently in the SFB compared to the Duplo task FB condition.

To find out whether there are limits in 3-year-olds' Duplo task performance, or whether it measures a fully-fledged ToM competence, we designed an intensionality version of the Duplo task that was identical to the change-of-location version with one exception: Rather than failing to witness the location change in the FB condition, the protagonist failed to witness how the experimenter revealed that the object (a pen) also had a second identity (is also a rattle). The protagonist then saw how the object, under this second identity (as a rattle), was transferred to the other container (Rakoczy et al., 2015). If there is unity, children should solve the intensionality version on the same level as the change-of-location version, and there should be convergence and correlation between them. We also tested for correlations between the Duplo task versions and the SFB task, which has not been done yet, in order to clarify if they measure the same ToM capacity.

2. Methods

2.1. Participants

The final sample of the study included 72 3-year-old children (36-47 months, $M = 41.6$, $SD = 3.1$; 31 boys) from mixed socio-economic background. The data collection was conducted in two different labs (first in Hamburg ($n = 32$), then in Göttingen ($n = 40$)) each by a female experimenter. In Hamburg, children were recruited and tested in seven different nurseries with written consent of their parents. In Göttingen, data collection was conducted either in children's nursery or in the lab and were recruited from databanks of children whose parents had previously agreed to participate in child studies. Four children were excluded because they did not cooperate, and testing sessions had to be interrupted.

We conducted statistical power analyses for sample size estimation (using G*Power 3.1.9.2; Faul, Erdfelder, Lang, & Buchner,

2007), aiming at alpha = .05 and power = 0.95, based on data from the original Duplo task study (N = 25; Rubio-Fernández & Geurts, 2013). The effect size in the original study for the above chance performance in the Duplo task was 0.3. The projected sample size needed with this effect size is approximately N = 35. For the performance improvement compared to the verbal FB task, the effect size in the original study was 1.15. The projected sample size needed with this effect size is approximately N = 15.

2.2. Design and Procedure

We presented each child with two versions of the Duplo task: A change-of-location task and an analogue intensionality task (within-subject; modeled after Rakoczy et al., 2015). Children received either TB or FB scenarios (between-subject). The tasks were presented in counterbalanced order. The Göttinger children additionally received a standard change-of-location task in the same condition (TB/FB) as the Duplo task versions (Wimmer & Perner, 1983). Here, the order of task type (standard and Duplo versions) was also counterbalanced. Children were tested individually in a quiet room. Each session started with a warm-up phase, in which the child and the experimenter (E) played with a cuddly toy and some objects. Already in this phase, children were encouraged to slip into the role of the puppet and act on her behalf.

2.2.1. Duplo tasks

The procedure of the Duplo change-of-location task was adopted from Rubio-Fernández and Geurts (2013) and analogously adapted for the intensionality task (after Rakoczy et al., 2015). Since the original intensionality task included explicit trickery together with the child (see Rakoczy et al., 2015, Appendix A), a feature that increased performance in other FB tasks (e.g., Białecka-Pikul et al., 2018; Sullivan & Winner, 1993), we implemented this feature in all of our tasks. Note that also in the original Duplo task the experimenter acted as if she would play a trick on the protagonist. During creation and piloting of the experimental protocol, we were in regular contact with the first author, P. Rubio-Fernández.

2.2.1.1. Duplo change-of-location task. E sat next to the child at a table and acted a puppet story. Before she started, E noted that at the end of the story she would need some help from the child. First, two boxes (representing fridges; materials are depicted in Fig. 1) and two objects (an apple and a banana) were shown to the child. The second object, the apple, was not used in the original study. However, in our design, it served to equate the demands of the two task versions, since the intensionality version included an object with two identities (see e.g., Fizke et al., 2017). Next, the protagonist, a cuddly toy ape (either the chimp Freddy, or the orangutan Klaus, counterbalanced across trials), was introduced and expressed his obsession with bananas (E mimicked the ape's voice). After he joyfully discovered the banana next to the fridges, he declared his plan to eat it right after his return, in order to stress his intention for the end of the story, and placed it together with the apple in one of the two fridges (side counterbalanced). He then walked across the table, passed the child and sat with his back turned to the fridges beside the scenery. Different to the original study, the ape put on little headphones when he turned away, to make sure that he could not hear what would happen. This was especially important for the novel intensionality version, which included a toy that made noise (rattling). In both conditions, when the ape was beside the scenery, E proposed to play a trick on the protagonist together in a sneaky manner ("Do we want to play a trick on Freddy/Klaus?"). In the FB condition, E then asked whether the ape was able to see or hear what the child and E were doing, in order to draw the participant's attention to the ape's perceptual state. In the TB condition, on the contrary, the ape turned back too early, and E asked about his perceptual access while he was already sitting in front of the two fridges. E reacted always in a confirmative (for correct answers) or a corrective (for incorrect answers) manner, raising the awareness of the ape's attention to the scene, which was either

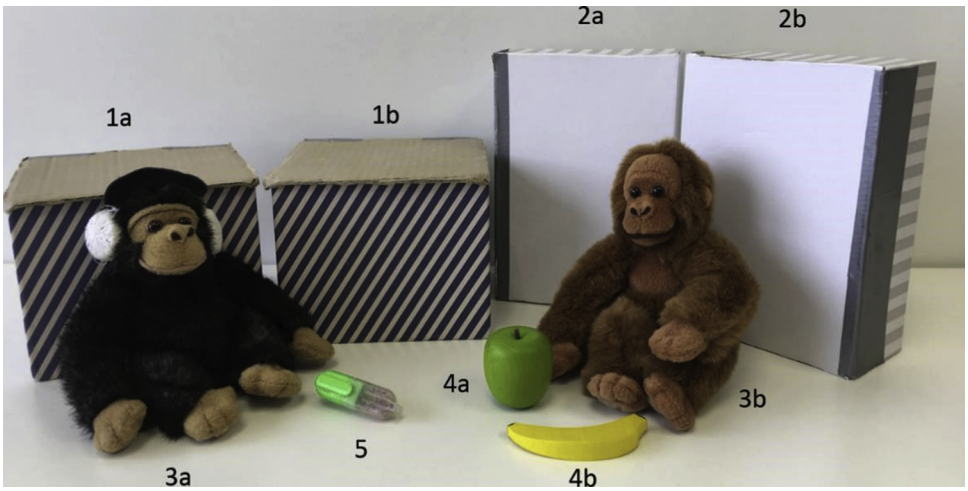

Fig. 1. Materials used in Duplo tasks: (1a, b) toy boxes and (5) the rattling pen used in the Duplo intensionality task, (2a, b) fridges and (4a, b) objects used in the Duplo change-of-location task, as well as both protagonists, (3a) the chimp Freddy wearing the headphones and (3b) the orangutan Klaus.

absent (FB) or present (TB). E then moved the object from one box to the other, and said, “Look, I put the banana in this fridge.” After the transfer, E asked whether the monkey witnessed the object’s location change (“Did Freddy/Klaus see that we put the banana in the other fridge?”). Note that, unfortunately, our phrasing of this prompt deviated from that in the original study (“She hasn’t seen what I did, has she?”). E reacted in either a confirmative or a corrective manner to the child’s answer, pointing to the knowledge (TB) or the ignorance (FB) about the transfer (e.g., “No, he hasn’t seen what happened.”). In contrast to the original procedure, the story was re-acted from the beginning if children answered incorrectly, to make sure they followed the narrative (this modification was conducted in the Göttinger sample only, as an improvement due to a few children’s failure to answer this question correctly in the Hamburger sample). While the protagonist was already present during the transfer in the TB condition, the ape returned at this point of the story in the FB condition and was positioned in front of the two fridges. In order to actively engage the participant in taking over the ape’s action, E uttered the question, “Humph, what happens next? Can you help me? Can you now take over Freddy/Klaus and continue playing the story?” If children did not cooperate in the first place, E prompted further, “What is Freddy/Klaus going to do?” If the prompt was also insufficient, she asked, “Will he approach one of the fridges?”

2.2.1.2. Duplo intensionality task. The intensionality task followed the very same procedure as the Duplo change-of-location task and was modeled after Rakoczy et al. (2015). The protagonist was introduced and stated his obsession with painting. Joyfully discovering a new pen he never saw before, he placed it into one of two toy boxes and stated his intention to paint when he comes back, at the same time showing that the box was otherwise empty. He then walked across the table, passed the child, put on the headphones, and sat with his back turned to the toy boxes beside the scenery. E then asked the child if they would want to play a trick on the protagonist and stated that she would share a secret with the child (acting in sneaky manner), namely that the pen had a second non-obvious identity as a rattle when shaken (“Look, the pen is also a rattle!”). In the TB condition, the ape turned around too early, so the secret was also shared with him. Thus, in contrast to the FB condition, he knew about the object’s second identity. The child’s awareness about the protagonist’s states of perception and knowledge were raised with questions similar as in the Duplo change-of-location task described above. In the presence of the protagonist, E transferred the object into the other toy box under the second identity: she covered the object with her hand (so that it was not visible), shook it during the transfer (so that its rattle-identity became perceivable), and said, “Look, I put the rattle in this box.” After the transfer, children were engaged in acting out the story as descripted in the Duplo change-of-location version.

2.2.2. Standard change-of-location task

The standard change-of-location task was adapted to the narrative puppet play of the Duplo task, but crucial factors responsible for disrupted perspective tracking (e.g. the visual absence of the protagonist during the swap, or mentioning the target object in an explicit question; Rubio-Fernández & Geurts, 2013, 2016) were included. A cuddly toy lynx (Luchsi) put his car into one of two containers and left the scene, so he was not visible for the child. E then proposed to play a trick together on the protagonist in the same way as in the Duplo tasks. In his absence (FB), or after his return (TB), E swapped the car to the other container in a sneaky manner. Children got the same prompts of the protagonist’s perceptual and knowledge states as in the Duplo task. However, instead of actively engaging the children to act out the end of the story, an explicit test question mentioning the target object was asked (“Where will Luchsi look for his car first?”).

3. Results

3.1. The Duplo tasks

In the location change version, six of the 72 participating children gave ambiguous or no answer and were thus excluded from the main analysis. Fig. 2 depicts the percentage of children’s answers as a function of task and condition. In the TB condition, 94% of the children acted out the belief-congruent ending of the story and placed the protagonist in front of the box containing the target object (binomial test, test value = 0.5, 29 out of 31 correct, $p < .001$, two-tailed; see Table 1). In the FB condition, 51% chose the belief-congruent box that did not contain the target object (further referred to as empty box) for the story ending (binomial test, test value = 0.5, 18 out of 35 correct, $p = 1$, two-tailed). The difference between the conditions in selecting either the full box or the empty box was significant ($\chi^2(1, N = 66) = 15.75$, $p < .001$, two-tailed).

In the intensionality version, four children had to be excluded due to ambiguous or no answers. In the TB condition, 85% of the children chose the belief-congruent box containing the object (binomial test, test value = 0.5, 29 out of 34 correct, $p < .001$, two-tailed). In the FB condition, 53% of the children chose the belief-congruent and thus empty box (binomial test, test value = 0.5, 18 out of 34 correct, $p = .864$, two-tailed). The difference between the conditions in selecting either the full box or the empty box was significant ($\chi^2(1, N = 68) = 11.1$, $p = .002$, two-tailed).

A comparison of the performances between the location change and intensionality versions revealed no significant differences for the FB conditions ($p = 1$, McNemar test, two-tailed, $n = 33$) and TB conditions ($p = .375$, McNemar test, two-tailed, $n = 29$). The FB conditions of the location change and the intensionality version correlated significantly with each other ($\phi(33) = .393$, $p = .024$). The TB conditions did not correlate with each other, due to ceiling effects in each task ($\phi(29) = .236$, $p = .204$).

3.2. Standard change-of-location task and relations to Duplo tasks

None of the 40 children who received a standard change-of-location task had to be excluded. In the SFB condition, children

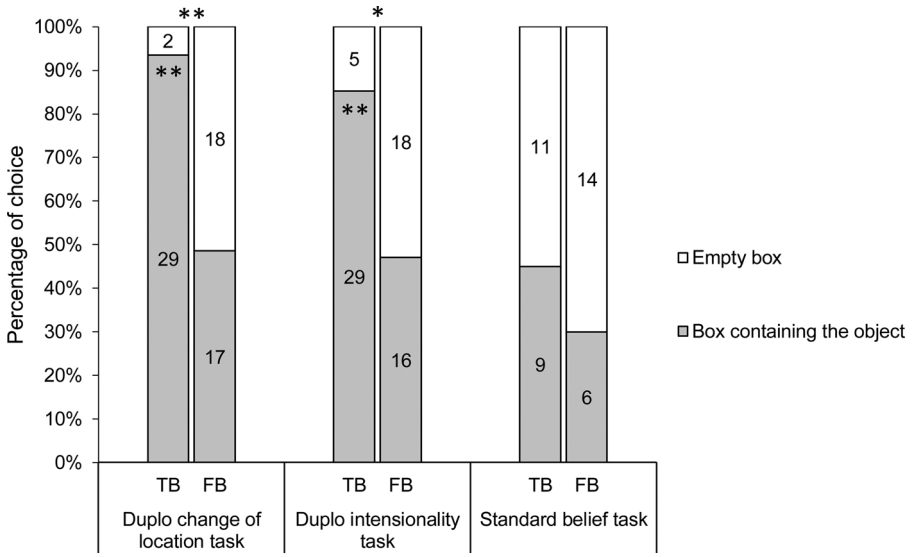

Fig. 2. Percentage of chosen box in all three tasks for both true belief (TB) and false belief (FB) conditions. Numbers in bars show number of children. *$p < .01$, **$p < .001$.

Table 1
Contingency pattern of given answers in the different tasks for false belief (FB) and true belief (TB) conditions.

		Standard change-of-location task		Duplo intensionality task	
		Empty box	Full box	Empty box	Full box
	Duplo change-of-location task				
FB	Empty box	10	0	12	5
	Full box	4	6	5	11
TB	Empty box	1	0	1	1
	Full box	9	8	4	23

performed at chance level (binomial test, test value = 0.5, 14 out of 20 correct, $p = .115$, two-tailed). A comparison between the Duplo location change FB and the verbal SFB yielded no differences ($p = .125$, McNemar test, two-tailed, $n = 20$). Further, the Duplo location change FB and the verbal SFB tasks correlated significantly with each other ($\phi(20) = .655$, $p = .003$). Similarly, a comparison between the Duplo intensionality FB and the verbal SFB revealed no significant differences ($p = .625$, McNemar test, two-tailed, $n = 18$) and a significant correlation ($\phi(18) = .553$, $p = .019$).

In the TB condition of the standard location change task, children performed at chance (binomial test, test value = 0.5, 9 out of 20 correct, $p = .824$, two-tailed) and there were no significant differences between the TB and FB conditions ($\chi^2(1, N = 40) = 2.56$, $p = .200$, two-tailed). Children performed significantly better in the TB conditions of both the Duplo location change task and the intensionality task compared to the TB condition of the verbal standard task (respectively: $p = .004$, McNemar test, $n = 18$; $p = .039$, McNemar test, $n = 19$; both two-tailed), and performance in the TB Duplo tasks did not correlate with the standard TB task, again likely due to the ceiling effects in the Duplo task (respectively: $\phi(18) = .217$, $p = .357$; $\phi(19) = .179$, $p = .435$).

3.3. Comparisons between the two labs

Preliminary analyses revealed no differences between the two labs in sex ($\chi^2(1, N = 72) = 0.01$, $p = 1$, two-tailed), age ($t(70) = -0.81$, $p = .432$, two-tailed) or overall performance per task (Duplo change-of-location: $\chi^2(1, N = 66) = 0.001$, $p = .973$; Duplo intensionality: $\chi^2(1, N = 68) = 0.09$, $p = .762$; both two-tailed). We therefore collapsed the data sets for the main analyses. However, we also found the same pattern of results as reported above in both labs separately, except for the significant correlation between the two Duplo tasks, which was found only in the Hamburger sample. However, with the increased power of both samples the correlation of performance in the two Duplo tasks was affirmed.

4. Discussion

The aims of the present study were to clarify how robust and reliable earlier Duplo task findings are and whether the Duplo task measures fully-fledged ToM competence in 3-year-olds. The main findings were the following. First, we could not reproduce the above chance performance in the Duplo change-of-location version. Second, performance in the Duplo task FB versions was not

different from the SFB, and the tasks were correlated. Third, in the Duplo TB conditions, children performed competently, and there was a significant difference to the Duplo FB conditions. Fourth, performance in the Duplo intensionality version was on par with the change-of-location version, and both versions were correlated.

Now, how robust and reliable are the Duplo task findings? Earlier findings of an above chance performance in a narrative FB task could not be reliably reproduced in the present study, which is in line with another recentstudy (Kammermeier & Paulus, 2018). Why did we find different results? One possibility may be that methodological issues played a key role. Although we adopted the procedure of the Duplo task as closely as possible, and the original first author checked and confirmed the accuracy of our implementation, there were some differences. First, we introduced a second object (an apple) in the Duplo change-of-location task, to equate the demands to the intensionality task that included an object with two identities. Since the Duplo task uses an open response design, with no further instructions about the storyline's direction, this second object may have distracted or confused our participants. However, this seems very unlikely because the apple always remained in the first (and correct) fridge in the FB condition, and thus did not lead children to the incorrect one. Further, not a single child ever referred to the apple in the FB condition when choosing a fridge, thus, there was no indication of confusion. Additionally, in the TB condition, most children correctly chose the location containing the banana, indicating that they understood the protagonist's desire to obtain the focal object, thus ruling out that they were confused by the presence of the apple.

Second, we included one additional feature in half of the sample: we repeated the story when the control question was answered incorrectly. This is a common procedure in explicit ToM tasks, ensuring that children understand the story and reducing the amount of drop-outs (e.g., Clements & Perner, 1994). In our case, we only had to repeat the story for six children, an equal number in each condition (TB or FB). Since we implemented this modification in only one lab and found no differences between the two sample groups, this feature cannot explain the different findings.

Third, there are certainly differences in the interaction style of the experimenters, or in the sample compositions. However, the current study is a collaboration of two labs and testing thus took place in two different cities with two different experimenters. Since both labs found the same results, this seems to be an unlikely explanation. One informal observation was that children often needed a second prompt to act out the story, which might indicate procedural difficulties in obtaining the dependent measure. Unfortunately, we have no information on similar findings from the other Duplo task studies (Kammermeier & Paulus, 2018; Rubio-Fernández & Geurts, 2013, 2016).

Fourth, we used a slightly different phrasing for the last prompt, during the belief-induction phase ("Did Freddy/Klaus see that we put the banana in the other fridge?" instead of "She hasn't seen what I did, has she?"). This may have stressed the salience of the target object. Rubio-Fernández and Geurts (2016) showed that emphasizing during the test phase that the protagonist wants the object ("Now Lola is very hungry and wants a banana."), or asking for the current location of the object ("Where is the banana now?"), can disrupt perspective tracking. Note that our prompt during the belief-induction phase rather emphasized the ignorance of the protagonist, not his desire to get the banana or the current banana location. Note also that German word order buries 'banana' in the middle of the sentence rather than exposing it prominently at the end (as in English). Ultimately, it remains an empirical question whether the 'banana' in our prompt could explain the worse performance. Findings by Białecka-Pikul et al. (2018) show that mentioning the target object in the test phase has no detrimental effect on children's performance.

A more general limitation of the Duplo task could be the open response design, which allows for less straightforward predictions of children's ToM competences than explicit tasks. That is, negative findings do not necessarily speak for a lack of belief ascription abilities. In our Duplo intensionality task, for example, a possible alternative ending could also be that the ape wanted to get the rattle he just found out about; or, in the Duplo change-of-location task that he wanted to check the other fridge. However, as in the original study, we strongly stressed the monkeys' desire to get the banana/pen at the beginning of the narrative. Most importantly, our TB-control conditions show that children clearly knew how to act out the story coherent with the protagonist's desire.

While our findings of children's performance at chance level in the Duplo FB task are in line with the direct replication study by Kammermeier and Paulus (2018), the Duplo-study by Białecka-Pikul et al. (2018), which focused on the influence of interaction in FB tasks, found above chance performance in the FB condition in three-and-half-year-old children and an increased performance compared to a standard unexpected-content FB task. During that task, the experimenter made the child focus on the protagonist's state of knowledge via prompts just like those in the Duplo task. However, in contrast to the original Duplo task, an explicit test question mentioning the target object was asked, and the protagonist was covered by a cloth during the location change rather than being visible beside the scene. Curiously, then, none of the potentially facilitating main factors of the original Duplo task, namely not mentioning the target object in the test phase, and continuous visibility of the protagonist, were implemented by Białecka-Pikul et al. (2018). Instead, that study implemented an emphasis on a 'we-mode' in their procedure that putatively led to children's enhanced performance. That is, the experimenter actively engaged and involved the participating child by using a deceptive motive of joint trickery, saying, "Hey, let's surprise the mouse!" However, in the current study we also implemented the interactive motive of joint trickery in all our tasks ("Do we want to play a trick on Freddy/Klaus?"), in addition to the original key manipulations of the Duplo task, and still found no significant above chance performance. It should be noted, however, that mean performance in Białecka-Pikul et al. (2018) was not drastically different from ours (respectively, 57% correct vs. 51% correct) and certainly diverged strongly from the original finding of 80% correct (Rubio-Fernández & Geurts, 2013). However, Białecka-Pikul et al. (2018) used a six times larger N of 210 children, rendering a statistical significance perhaps less meaningful.

Our interpretation of the findings is then that there is no robust facilitating effect of the Duplo task after all. In the unexpected-content task that was originally used as SFB control (and also in Białecka-Pikul et al., 2018) and in the location FB task used by Kammermeier and Paulus (2018), children performed below chance, and thus, performance in the Duplo task was significantly better. However, we found no such difference in performance between the Duplo task and a SFB task when the latter was matched

accordingly. It is theoretically possible that the matching artificially enhanced performance in our SFB task, because we included the Duplo prompts during the belief-induction phase about the protagonist's epistemic state (e.g., "Did Luchsi see that we put the car in the other box?"). However, given that the prompts also mentioned the object, it is equally possible that they led to disrupted perspective tracking. Our at chance level findings, however, concur well with the meta-analytic findings (Wellman et al., 2001) suggesting at chance level performance at 3.5 years of age. Though beyond the scope of the current paper, we have recently run our current SFB task in a different study without prompts during the belief-induction phase. We found a very similar pattern of performance at chance with no significant differences to the current results. In addition, other findings show that when adding an explicit question to the Duplo task (Rubio-Fernández & Geurts, 2013, Experiment 2b), or stressing the target object in the Duplo task (Rubio-Fernández & Geurts, 2016, Experiment 2), performance drops dramatically below chance - despite the epistemic prompts. Thus, it is unlikely that the epistemic prompts alone led to an increase in performance on our SFB task.

Our study equated both kinds of FB tasks in terms of structure and performance factors and implemented an analogue standard change-of-location control that only differed from the Duplo task in the most important elements that disrupt perspective tracking (i.e., agent disappeared from scene, explicit test question mentioning the target object). This is reminiscent to findings of Papafragou et al. (2017) who could not replicate advantages of word learning tasks on FB understanding with matched controls. Further, it urges future researchers to adjust their control conditions carefully. The previously used SFB comparisons were cognitively more demanding than the Duplo task (Gopnik & Astington, 1988; Holmes et al., 1996), which led to the assumption of a facilitating effect. Our minimal contrast design, however, suggests that the Duplo task leads to no improvement and has no facilitating effect when compared to a matching SFB. Indeed, our findings reveal not just that the Duplo task has no facilitating effect – in addition, they reveal that performance on the Duplo task and the SFB task are based on a common capacity. We found evidence for converging performance, i.e. both Duplo task versions correlated significantly with the matching SFB task. Thus, the Duplo task may tap the same ToM competence as standard explicit tasks, and explicit tasks may not underestimate children's FB understanding.

The second aim of the present study was to investigate whether we could transfer the facilitating effect of the Duplo change-of-location task to other FB paradigms, and whether 3-year-olds possess a fully-fledged ToM competence or one that exhibits limitations. As described above, we did not find any facilitating effects in the Duplo change-of-location FB version in the first place. Similarly, performance in the Duplo intensionality FB version was at chance. However, performance of both task versions showed convergence and correlation. Thus, the Duplo task seems to measure a generalizable and not merely local phenomenon. In this respect, it behaves like standard explicit tasks that also reflect this unity and convergence (Perner & Roessler, 2012; Rakoczy et al., 2015), rather than implicit tasks that often show disunity and divergence (Fizke et al., 2017; Oktay-Gür et al., 2018). Performance on both Duplo task FB versions was significantly different from the TB controls. This makes it unlikely that children were totally random in their choice. Given the dichotomous data of the task, a possible interpretation of 3-year-old children's at-chance performance on the SFB task is that half of the children possess the competence (Lohmann, Carpenter, & Call, 2005). This interpretation is corroborated by our correlational findings which exclude lower level interpretations that children were just guessing or perseverating. Instead, the pattern of performance across all three FB tasks indicates that children either systematically passed or failed in answering in a belief-congruent way, suggesting that half our participants were competent FB passers. Presumably, these children already possessed a robust and unified ToM capacity that is comparable to that of older children. The other half might have failed, as at the age of three, FB understanding is still a fragile competence that is prone to overwhelming demands on pragmatic confusion and cognitive capacities (Helming, Strickland, & Jacob, 2014; Wellman et al., 2001). The findings thus show that performance in the Duplo task is much closer related to standard explicit tasks than previously assumed.

In line with the original study, we found very good performance in our new matching Duplo TB conditions, and we found significant condition differences. Thus, despite methodological differences, the original TB condition did not overestimate children's performance. In contrast, children performed at chance level in the TB condition of the standard task. On first glance, this finding might seem quite surprising. However, it nicely fits with recent studies that focus on children's TB performance, showing that the relation to FB performance seems more complex than previously assumed (Fabricius, Boyer, Weimer, & Carroll, 2010; Oktay-Gür & Rakoczy, 2017; Perner, Huemer, & Leahy, 2015). Oktay-Gür and Rakoczy (2017) found a U-shaped curve of TB development, i.e. while young 3-year-olds passed standard TBs, from three-and-a-half years on performances dropped and only until the age of ten years, children began to pass the TBs again. The FB performance, on the other hand, increased with age. Even more astonishing, the study found negative correlations between TB and FB conditions between age three and ten, i.e. those who passed one condition failed the other. An explanation of this paradoxical pattern of findings is that standard TB scenarios create an artificial situation in which everyone (experimenter, protagonist and child) has the same state of knowledge about locations or identities, so that it seems trivial to ask explicitly for a prediction about the protagonist's behavior. In line with this idea of pragmatic confusion rather than a competence limitation, Oktay-Gür and Rakoczy (2017) found that the U-shaped curve of TB performance disappeared once the scenario and the test situation were made less trivial. Applying this logic to the current pattern in the TB conditions suggests that the Duplo task might similarly decrease pragmatic confusion factors compared to standard tasks, leading to a better performance in the Duplo-TB than the standard TB. This pragmatic advantage does not extend to the FB conditions in the same way because children's competence in understanding false beliefs is still limited (at least in half of our 3-year-old sample). While several studies have found enhanced performance by manipulating different aspects of the standard FB task (e.g., Mitchell & Lacohée, 1991; Psouni et al., 2018; Rhodes & Brandone, 2014; Sullivan & Winner, 1993), meta-analytic findings converge to show that group-level performance remains at chance at 3.5 years of age (Wellman et al., 2001).

The current study failed to reproduce earlier facilitating effects of a simplified FB task, when using matching control conditions. The Duplo task rather seems to tap into the same cognitive system as standard explicit tasks, and make the same demands. Further, correlation and convergence across task types indicate the measurement of a unified ToM competence in 3-year-olds. Yet, the

different studies to date using the Duplo task have found very different patterns of results, which leaves open many questions on robustness, replicability and validity of the findings. What we need, in light of the growing amount of recent replication failures in our field, are systematic, large-scale, pre-registered and collaborative cross lab replication studies of measures of early ToM.

Acknowledgments

This study was supported by the German Research Foundation (LI 1989/3-1, RA 2155/4-1, Project: FOR 2253). We are grateful for the participation of all the parents, children and nurseries. Further, we want to thank Fanny Klein for help with data collection, as well as Marlen Kaufmann, Konstanze Schirmer and Jessica Schröter for lab coordination. We also want to thank Paula Rubio-Fernández for providing advice on her task.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*(4), 953–970. https://doi.org/10.1037/a0016923.

Białecka-Pikul, M., Kosno, M., Białek, A., & Szpak, M. (2019). Let's do it together! The role of interaction in false belief understanding. *Journal of Experimental Child Psychology*. https://doi.org/10.1016/J.JECP.2018.07.018.

Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology, 20*(3), 393–420. https://doi.org/10.1348/026151002320620316.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language, 28*(2), 141–172. https://doi.org/10.1111/mila.12014.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*(4), 377–395. https://doi.org/10.1016/0885-2014(94)90012-4.

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. https://doi.org/10.1037/a0017648.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146.

Fizke, E., Butterfill, S., van de Loo, L., Reindl, E., & Rakoczy, H. (2017). Are there signature limits in early Theory of Mind? *Journal of Experimental Child Psychology, 162*, 209–224. https://doi.org/10.1016/j.jecp.2017.05.005.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*(1), 26–37. https://doi.org/10.2307/1130386.

Happe, F., & Loth, E. (2002). "Theory of Mind" and tracking speakers' intentions. *Mind and Language, 17*(1&2), 24–36. https://doi.org/10.1111/1468-0017.00187.

Helming, K. A., Strickland, B., & Jacob, P. (2014). Making sense of early false-belief understanding. *Trends in Cognitive Sciences, 18*(4), 167–170. https://doi.org/10.1016/j.tics.2014.01.005.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development, 57*(3), 567–582. https://doi.org/10.2307/1130337.

Holmes, H. A., Black, C., & Miller, S. A. (1996). A cross-task comparison of false belief understanding in a head start population. *Journal of Experimental Child Psychology, 63*(2), 263–285. https://doi.org/10.1006/jecp.1996.0050.

Kammermeier, M., & Paulus, M. (2018). Do action-based tasks evidence false-belief understanding in young children? *Cognitive Development, 46*, 31–39. https://doi.org/10.1016/j.cogdev.2017.11.004.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in "Theory of Mind.". *Trends in Cognitive Sciences, 8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001.

Lohmann, H., Carpenter, M., & Call, J. (2005). Guessing versus choosing - and seeing versus believing - in false belief tasks. *British Journal of Developmental Psychology, 23*(3), 451–469. https://doi.org/10.1348/026151005X26877.

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives, 10*(3), 184–189. https://doi.org/10.1111/cdep.12183.

Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a signature blind spot in humans' efficient mind-reading system. *Psychological Science, 24*(3), 305–311. https://doi.org/10.1177/0956797612451469.

Mitchell, P., & Lacohée, H. (1991). Children's early understanding of false belief. *Cognition, 39*, 107–127.

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

Oktay-Gür, N., Schulz, A., & Rakoczy, H. (2018). Children exhibit different performance patterns in explicit and implicit Theory of Mind tasks. *Cognition, 173*, 60–74. https://doi.org/10.1016/j.cognition.2018.01.001.

Papafragou, A., Fairchild, S., Cohen, M. L., & Friedberg, C. (2017). Learning words from speakers with false beliefs. *Journal of Child Language, 44*(4), 905–923. https://doi.org/10.1017/S0305000916000301.

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

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences, 16*(10), 519–525. https://doi.org/10.1016/j.tics.2012.08.004.

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308*, 214–216. https://doi.org/10.1126/science.1111656.

Psouni, E., Falck, A., Boström, L., Persson, M., Sidén, L., & Wallin, M. (2018). Together I can! Joint attention boosts 3- to 4-year-olds' performance in a verbal false-belief test. *Child Development*. https://doi.org/10.1111/cdev.13075.

Rakoczy, H., Bergfeld, D., Schwarz, I., & Fizke, E. (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. https://doi.org/10.1111/cdev.12311.

Rhodes, M., & Brandone, A. C. (2014). Three-year-olds' theories of mind in actions and words. *Frontiers in Psychology, 5*(263), 1–8. https://doi.org/10.3389/fpsyg.2014.00263.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33. https://doi.org/10.1177/0956797612447819.

Rubio-Fernández, P., & Geurts, B. (2016). Don't mention the marble! The role of attentional processes in false-belief tasks. *Review of Philosophy and Psychology, 7*(4), 835–850. https://doi.org/10.1007/s13164-015-0290-z.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012.

Sullivan, K., & Winner, E. (1993). Three-year-olds' understanding of mental states: The influence of trickery. *Journal of Experimental Child Psychology, 56*, 135–148.

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

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind tasks. *Child Development, 75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x.

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