Cognitive Development 48 (2018) 316–320

Contents lists available at ScienceDirect

Cognitive Development

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

Letter to the Editor

What do failed (and successful) replications with the Duplo task show?

1. Introduction

In the classic Sally-Anne task (Baron-Cohen, Leslie, & Frith, 1985), Sally put s a marble in a box before going out to play and, in her absence, Anne moves the marble to a basket. The child is then asked the standard false-belief question: 'When Sally comes back, where will she look for her marble?' Hundreds of Theory of Mind studies in the last 30 years have shown that children are not able to pass change-of-location tasks before their 4th birthday, with younger children predicting that Sally will look for her marble in its current location rather than in the container where she left it (Wellman, Cross, & Watson, 2001).

The Duplo task is a variation on the Sally-Anne task that introduces two sets of modifications to the original paradigm, both intended to help the child stay focused on the protagonist throughout the false-belief narrative (Rubio-Fernández & Geurts, 2013). First, it is ensured that the child can see the protagonist throughout the displacement phase of the task. Rather than making the protagonist disappear, as is standardly done in this type of false-belief task, the experimenter makes the Duplo girl walk in the direction of the child and turn her back on the scene. As the experimenter transfers the target object from one container to the other, she keeps checking with the child that the protagonist cannot see what she is doing – another difference with the standard Sally-Anne task.

The second set of modifications is introduced in the test phase. When the experimenter returns the Duplo girl back to the center of the scene, rather than asking the child the standard false-belief question, she places the figure in front of the two cupboards and asks the child whether he would like to continue the story. The experimenter then encourages the child to take the lead by asking: 'What happens next? What is the girl going to do now?'

Three-year old children were able to pass the Duplo task (Rubio-Fernández & Geurts, 2013), performing significantly above chance level rather than below (as they normally do at that age). Moreover, the children in this study performed significantly better in the Duplo task than in the Smarties task (another classic false-belief task involving unexpected contents; Hogrefe, Wimmer, & Perner, 1986), showing success rates of 80% and 23%, respectively.

Kammermeier and Paulus (2018) have recently failed to replicate the above-chance performance observed by Rubio-Fernández and Geurts (2013) with 3-year-olds. In their first experiment, 4-year-old and 5-year-old children performed above chance in the Duplo task (with success rates of 85% and 100%, respectively), while 3-year-olds performed only at chance (50% success rate) and significantly worse than the other two age groups. In their second experiment, 3-year-old children performed again at chance in the Duplo task (48% success rate), while 4-year-olds performed significantly better, reaching an above-chance level (90% success rate). Both age groups, however, performed significantly better in the Duplo task than in two standard false-belief tasks (one change-of-location and one unexpected-contents task), with 3-year-olds performing below-chance (20% success rate) and 4-year-olds performing at chance (57% success rate) in both standard tasks.

In what follows, I will first discuss possible reasons why Kammermeier and Paulus may not have been able to replicate the above-chance performance observed by Rubio-Fernández and Geurts (2013) and extend the discussion to other replications – failed and successful. Then, I will argue that Kammermeier and Paulus built a strawman in their discussion of Rubio-Fernández and Geurts (2013), who never claimed that 3-year-old children pass the Duplo task by attributing the protagonist a false belief. An accurate description of the position defended by Rubio-Fernández and Geurts shows how the better performance observed by Kammermeier and Paulus with the Duplo task, relative to two standard false-belief tasks, supports in fact their original hypothesis. Finally, I will argue that Kammermeier and Paulus's analysis of the Duplo task sets the bar too high for false-belief tasks, running the risk of underestimating young children's cognitive abilities, which is as detrimental to the field as overestimating them.

2. Failed and successful replications with the Duplo task

Rubio-Fernández and Geurts (2016) successfully replicated their original findings in a second study. However, Kammermeier and Paulus (2018) do not refer to this follow-up study in their paper. More recently, Marta Białecka-Pikul, Kosno, Białek, and Szpak, 2018a,b, in press) tested a group of 210 children in their third year (131 more than the 3-year-olds tested by Kammermeier and

https://doi.org/10.1016/j.cogdev.2018.07.004
Received 22 January 2018
Available online 27 August 2018

Paulus, and with large effect sizes of 4:0) and observed that the same group of children went from below-chance performance in the Duplo task at age 3;0 to above-chance performance at age 3;6 (which was the mean age in the original study).

However, the Duplo task seems to be what is sometimes called a 'fragile paradigm' in developmental psychology, since it is not always possible to replicate the above-chance performance originally observed with young children. To this date, I am aware of another two research groups, one led by Josef Perner and Beate Priewasser and another one by Ulf Liszkowski and Hannes Rakoczy, who have also failed to replicate our original results with 3-year-olds (the former in an undergraduate thesis and the latter in a graduate research project). These researchers engaged with me in detailed discussion of their protocols and shared videos of their implementations of the task, which allowed me to identify differences in the protocols.

Kammermeier and Paulus, on the other hand, never shared with us any written description or footage of their own protocol. This is important because these authors claim in their paper to have kept 'close contact with the study's first author to ensure an exact replication protocol' (p. 37). However, while I did volunteer a few pointers when I first learned of their failed replication, Kammermeier and Paulus never confirmed whether they had implemented my suggestions, nor did they describe or show me how they had actually implemented the task in the first place. I therefore cannot confirm that Kammermeier and Paulus have used 'the exact same procedure' that we used in our study (p. 32), as they repeatedly claim in their paper.

The results of Rubio-Fernández and Geurts (2013, 2016) show that subtle changes to the original paradigm can make 3-year-olds go from above-chance performance to below-chance performance. It is therefore crucial that the original paradigm is accurately reproduced in attempted replications of these findings. In the case of Liszkowski and Rakoczy's failed replication, their experimenter mentioned the target object prior to the test phase ('Did Klaus see that we moved the banana?'), which is not only different from the original Duplo task (where children's attention was never drawn to the object) but has actually been shown to compromise young children's performance (see Rubio-Fernández & Geurts, 2016).[1]

Another feature of the Duplo task that Kammermeier and Paulus may not have achieved in their study is the appropriate level of engagement with the kids in the task. Since children in the Duplo task are supposed to engage with the experimenter in a plot to deceive the protagonist, the experimenter should try to engage the children in the story, especially the younger ones who are more susceptible to distraction. The results by Kammermeier and Paulus show that around 20% of their 3-year-old children failed to produce a relevant response to the test question in the Duplo task, whereas uncooperative/irrelevant responses were observed in 4% of participants in Rubio-Fernández and Geurts (2013) and 5% in Rubio-Fernández and Geurts (2016). The relatively large proportion of children who failed to engage in the false-belief narrative in the study by Kammermeier and Paulus suggests that the experimenters who ran their task might not have sufficiently engaged with the children.

Białecka-Pikul et al.'s (2018a,b in press) replication of the results of Rubio-Fernández and Geurts (2013, 2016) suggests that engaging the kids in the deception may be critical for their successful performance, since their critical manipulation was to engage the children in the deception by inviting them to trick the protagonist before the experimenter transferred the object (for further evidence of the importance of engaging with 3-year-olds in a false-belief narrative, see Psouni et al., 2018).

These studies suggest that the original results with the Duplo task are replicable (not only by the authors of the first study but also by independent researchers), although this false-belief task is certainly a fragile paradigm with 3-year-olds given the difficulties that various research groups have had to replicate the original results. The question is what to make of these mixed findings. My personal response is in line with Rubio-Fernández and Geurts's original position, which unfortunately was not discussed by Kammermeier and Paulus: the aim of the Duplo task was not to show that children under 4 are able to attribute false beliefs to others, but to see whether task manipulations could help improve young children's performance in these tasks. In particular, Rubio-Fernández and Geurts (2013) hypothesized that the infant studies that had reported significant passing rates in false-belief tasks (e.g., Onishi & Baillargeon, 2005; Southgate, Senju, & Csibra, 2007) had allowed babies to focus on the protagonist throughout the narrative, not drawing their attention to the incorrect response; by contrast, the standard Sally-Anne task (and similar paradigms) did not preserve the children's focus of attention on the protagonist and increased the salience of the wrong response in the test phase by mentioning the target object ('Where will Sally look for her marble?').

The different designs of these false-belief tasks leave open the question of whether the modified tasks that have revealed improved performance with infants and toddlers are tapping the same Theory of Mind abilities as the standard tasks that children under 4 tend to fail. The way in which Rubio-Fernández and Geurts addressed this question was to further investigate the limits of toddlers' capacity to pass the Duplo task by manipulating the original paradigm. In other words, what were the features of task design that allowed 3-year-olds to pass the Duplo task while failing standard false-belief tasks? The results of Rubio-Fernández and Geurts (2013, 2016) show that 3-year-olds need to focus on the protagonist throughout the false-belief narrative in order to predict the correct outcome, and that drawing their attention to the target object compromises their performance as it increases the salience of the wrong response. Rubio-Fernández and Geurts (2016); see also Rubio-Fernández, 2018a, forthcoming) further discuss that those features of the Duplo task that allow 3-year-old to form the correct expectations about the protagonist's actions may also allow for low-level responses rather than false-belief attribution. For example, by not drawing children's attention to the current location of the object, 3-year-olds may rely on a process of 'spatial indexing' (Richardson & Kirkham, 2004) whereby they keep track of the location where the agent had left the object before it was moved.

When interpreting the results of failed and successful replications with the Duplo task, it must therefore be noted that the original

[1] Regarding Josef Perner and Beate Priewasser's failed replication, I did not observe any obvious differences in the way they implemented the Duplo task, but Josef Perner explained (email communication) that there were a number of reasons why their undergraduate student may not have replicated the original findings, including the order of presentation of the tasks they used.

results by Rubio-Fernández and Geurts (2013, 2016) were actually mixed, with 3-year-olds being able to pass two versions of the Duplo task, while failing four. Given the subtle differences in the design of these different versions of the Duplo task, the results of the original studies (like the attempted replications by other groups) suggest that 3-year-olds' capacity to pass the Duplo task is rather fragile and dependent on very specific experimental factors. The limits of this capacity, however, should be as informative about 3-year-olds' developing Theory of Mind as their success. At least, that was the rationale for Rubio-Fernández and Geurts's research program.

3. Baseline performance and control condition in the Duplo task

In this section, I will respond to some specific criticisms raised by Kammermeier and Paulus (2018) against the Duplo task and its interpretability. Kammermeier and Paulus (2018) observed that their 3-year-old children performed significantly better in the Duplo task than in two standard false-belief tasks: whereas in the standard tasks they showed a reliable preference for the incorrect response, these children were evenly split between the two responses in the Duplo task. A significant improvement was also observed with their 4-year-olds, who went from being divided between the two responses in the standard false-belief tasks to showing a reliable preference for the correct response in the Duplo task. This significant improvement is a direct replication with the same age group and an extension with older children of the significant difference observed by Rubio-Fernández and Geurts (2013) between 3-year-olds' performance in the Duplo and the Smarties tasks – a significant difference that was also replicated by Białecka-Pikul et al. (2018a,b in press). However, Kammermeier and Paulus do not acknowledge this partial replication and extension in their paper. In fact, the better performance observed with the Duplo task is not even mentioned in the Abstract, despite being observed in the two age groups. Instead, Kammermeier and Paulus argue that the Duplo task may not be directly comparable to other false-belief tasks because it is unclear what the baseline performance should be (below-chance or chance level) given the open test questions ('What happens next? What is the girl going to do now?').

I have two points in response to this counterargument. First, the reason why standard false-belief tasks are considered to have a below-chance baseline is because one of the containers hides the target object and children need to overcome a 'pull of the real' or 'true-belief bias' in order to pass the task. In the Duplo task, the target object remains inside one of the containers during the test phase, rather than being removed from the scene, as in some false-belief tasks for infants (e.g., Southgate et al., 2007). Therefore, children in the Duplo task should be subject to the pull of the real just as much as they are in standard versions of the false-belief task, with all these tasks arguably having a below-chance baseline.

Second, our own results support the view that the Duplo task, like standard versions of the false-belief task, has a below-chance level baseline: in those versions of the Duplo task that 3-year-olds failed (Rubio-Fernández & Geurts, 2013, Experiments 2a and 2b; 2016, Experiment 2), children performed significantly below chance, rather than at chance level – as Kammermeier and Paulus would expect. The results of those experiments support the hypothesis that young children need to be continuously focused on the protagonist of a false-belief narrative in order to predict that she will return to the empty container. Making the girl figure disappear before the transfer of the object (as is normally done in the Sally-Anne task) is sufficient for young children to perform below chance in the Duplo task. Likewise, drawing children's attention to the target object before they have to continue the story (e.g., by asking the child 'Where are the bananas now?' or 'Now the Duplo girl is hungry and wants a banana'), or simply mentioning the target object in the test question (e.g., 'Where will the Duplo girl look for her bananas?') also disrupt young children's focus on the protagonist and lead them to predict the wrong outcome.

It must be noted that Kammermeier and Paulus (2018) do not make any reference to those versions of the Duplo task that 3-year-olds were unable to pass, even though they actually failed most versions of the task (more precisely, four out of six versions; see Rubio-Fernández & Geurts, 2013, 2016). As we acknowledge in those studies, 3-year-old children's ability to pass the Duplo task is rather fragile, with minimal changes to the original protocol drawing children's attention to the wrong response and resulting in below-chance performance.

Kammermeier and Paulus also argue that the open test questions used in the Duplo task make it impossible to interpret children's responses, especially since the task does not include control questions that could show whether children had understood the story correctly. First of all, Rubio-Fernández and Geurts (2016; Experiment 2) showed that the standard control questions that Kammermeier and Paulus favour (e.g., 'Where is the marble now? And where did Sally put the marble before she went away?') actually compromise young children's performance in the Duplo task: when 3-year-olds were asked 'Where are the bananas now?' before they were invited to continue the story, children showed a reliable preference for the actual location of the bananas when they moved the Duplo girl (whereas the reverse preference was observed without the control question).

Second, in order to interpret children's responses to the open test questions, Rubio-Fernández and Geurts (2013, 2016) introduced a true-belief condition in which the Duplo girl was right about the location of the bananas. Therefore, Rubio-Fernández and Geurts's positive interpretation of children's performance in the Duplo task was based not only on children's preference for the empty container in the false-belief condition, but also on children's preference for the actual location of the bananas in the true-belief control.

Kammermeier and Paulus acknowledged that Rubio-Fernández and Geurts (2013) had used a true-belief condition, but criticized it for including a low number of children. The reason for this is that one of the reviewers from *Psychological Science* asked that we ran a true-belief control (precisely, so that children's responses to the open test questions could be interpreted), and there were only 14 children in the preschool where we had run the study who had not yet been tested. We therefore had to keep the number of participants low if we wanted to test 3-year-old children from the same pool in the control task. However, Rubio-Fernández and Geurts (2016) ran equal numbers of 3-year-olds in their Duplo task and true-belief control (21 each) and observed reverse performance: children in the Duplo task continued the story by taking the protagonist to the empty container, whereas children in the true-

belief control took the protagonist to the actual location of the object.

Kammermeier and Paulus also argue that associative mechanisms or behavioural rules may explain children's performance in the true-belief condition, but what behavioural rules would those be? If Kammermeier and Paulus's counterargument is to be taken seriously, they need to spell out what behavioural rules could explain the results of the true-belief conditions used by Rubio-Fernández and Geurts (2013, 2016); simply suggesting that there may be such an explanation is not good enough to discredit their findings.

One last point to consider regarding the interpretability of the Duplo task is that Kammermeier and Paulus observed chance performance with 3-year-olds but close to ceiling performance with 4- and 5-year-olds (with success rates of 85–90% and 100%, respectively). If the test questions used in the Duplo task were as ambiguous and open to random interpretation as these authors suggest, there should have been more variability in all of the children's responses, and not only in the younger group. Therefore, the fact that the older children gave almost consistently the correct response to the open test questions seems to suggest that they interpreted them as expected.

4. On building a strawman and the danger of expectancy biases

It is important to note that Rubio-Fernández and Geurts have never claimed, either in print or in conference presentations of their work, that 3-year-olds' success in the Duplo task is evidence of their false-belief understanding. Rubio-Fernández and Geurts (2016) even suggested possible low-level explanations of their findings, and my most recent research has indeed tried to investigate this possibility with interesting results (Rubio-Fernández, 2018a, forthcoming).

Our aim with the Duplo task was to address the debate on the so called 'false-belief paradox': the seemingly puzzling finding that infants and young children are able to pass non-verbal or indirect false-belief tasks (e.g., Clements & Perner, 1994; Onishi & Baillargeon, 2005), whereas children under 4 years fail standard tasks. Our analysis of these different false-belief tasks suggested that they make different pragmatic and attentional demands, with the indirect tasks, but not the standard tasks, allowing infants and young children to focus on the protagonist throughout the narrative. Standard false-belief tasks, on the other hand, draw children's attention to the target object with the test question (or even before, with the control questions), increasing the salience of the wrong response (see Rubio-Fernández (2018b) for eye-tracking evidence with 3- and 5-year-old children, and Rubio-Fernández (2013) for eye-tracking evidence with adults).

However, Kammermeier and Paulus (2018) attribute to Rubio-Fernández and Geurts a much stronger position than we have ever defended, claiming, for example, that the results of our first study were interpreted as 'strong support for theoretical claims on an early emerging understanding of other's false beliefs' (p. 32) instead of reporting our position. Likewise, in motivating their first experiment, they argue that 'if children indeed show an early understanding of false beliefs, we would expect above-chance performance already in 3-year-old children that should not be different from the other age groups' (p. 32). However, we have never claimed that children do not undergo any Theory of Mind development between ages 3 and 5, or that their performance in the Duplo task should be comparable across those ages.

Kammermeier and Paulus's building of a strawman suggests a possible expectancy bias in their negative interpretation of their own results. It is often argued that the rich interpretations of infants' and young children's success in modified false-belief tasks may be fueled by an expectancy bias from those research groups defending innate or early-emerging Theory of Mind abilities (for discussion, see Heyes, 2014; Ruffman, 2014). However, we should not forget that the reverse bias can also affect attempted replications of these studies.

Two points in Kammermeier and Paulus's (2018) study deserve attention in this respect. First, these authors used a double standard when interpreting their own findings. In motivating their first experiment, they argue that if children show an early understanding of false beliefs, 3-year-olds should perform above chance in the Duplo task. In line with that prediction, Kammermeier and Paulus interpreted 3-year-olds' chance performance in their two experiments as evidence that young children 'do not demonstrate early false-belief understanding in an action-based task' (p. 33 and p. 37). However, when they discuss 3- and 4-year-olds' better performance in the Duplo task relative to two standard false-belief tasks, Kammermeier and Paulus dismiss the Duplo task as a genuine test of false-belief understanding and conclude that they 'have good reasons to argue that we do not know what exactly is assessed by the Duplo task' (p. 38). If that is the case, how could these authors interpret 3-year-olds' chance performance in the Duplo task as a Theory of Mind failure? This double standard needs to be addressed if their results and conclusions are to be taken seriously, either way.

Finally, Kammermeier and Paulus argue that when children fail to produce a response in the Duplo task, or give a response that is irrelevant to the false-belief narrative, their responses should be counted as failing the task. This, however, is highly debatable. Consider the standard Sally-Anne task: if a child refused to respond to the question 'Where will Sally look for her marble?', would their (lack of a) response reveal the same failure to take Sally's perspective as that of a child who predicted the wrong outcome? Surely not. We simply have no evidence of whether a child who refuses to answer a false-belief question (or who continues the narrative in a completely unrelated direction in the Duplo task) is able or unable to track the protagonist's perspective. This is why we did not include those responses in our statistical analyses.

Counting children's uncooperative/irrelevant responses as Theory of Mind failures sets the bar too high for false-belief tasks, which may offer support to the lean views of early development generally favoured by Kammermeier and Paulus, but is not beneficial for the field by and large. Let us not forget that attributing young children cognitive abilities that they do not yet possess is as inaccurate as denying them those abilities that they may possess.

In conclusion, Kammermeier and Paulus (2018) observed lower passing rates than Rubio-Fernández and Geurts (2013, 2016),

which casts doubts on 3-year-olds' abilities to pass verbal false-belief tasks. Their results, however, support the conclusions of Rubio-Fernández and Geurts as to how young children can perform better in false-belief tasks where they are allowed to focus on the protagonist throughout the narrative – something they are not allowed to do in standard false-belief tasks. Given their need to stay focused on the protagonist without paying attention to the actual location of the object, future studies should try to establish what cognitive abilities young children rely on to pass modified false-belief tasks. Such an approach should also shed light on the variability of the results observed with the Duplo task.

Acknowledgements

This research was supported by a *Young Research Talent Grant* from the Research Council of Norway (Ref. 230718). The author gratefully acknowledges this funding and thanks Stuart Markovitch for the opportunity to write a Letter to the Editor, and Chris Cummins and two reviewers for valuable comments on the manuscript.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*(1), 37–46.
Białecka-Pikul, M., Kosno, M., Białek, M., & Szpak, M. (2018a). Let's do it together! The role of interaction in false belief understanding. *Poster presented at the 2018 Budapest CEU Conference on Cognitive Development (BCCCD18)* in press.
Białecka-Pikul, M., Kosno, M., Białek, M., & Szpak, M. (2018b). Let's do it together! The role of interaction in false belief understanding. *Journal of Experimental Child Psychology* in press.
Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*(4), 377–395.
Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science, 17*(5), 647–659.
Hogrefe, J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development, 57*(3), 567–582.
Kammermeier, M., & Paulus, M. (2018). Do action-based tasks evidence false-belief understanding in young children? *Cognitive Development, 46*(2), 31–39.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.
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. in press *Child Development*.
Richardson, D. C., & Kirkham, N. Z. (2004). Multi-modal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General, 133*(1), 46–62.
Rubio-Fernández, P. (2013). Perspective tracking in progress: Do not disturb. *Cognition, 129*(2), 264–272.
Rubio-Fernández, P. (2018a). *Associative and inferential processes in false-belief tasks: An investigation of the unexpected-contents paradigm.* forthcoming*Journal of Experimental Child Psychology*.
Rubio-Fernández, P. (2018b). Spatial indexing in false-belief tasks: A continuous eye-tracking study. *Poster presented at the 2018 Budapest CEU Conference on Cognitive Development (BCCCD18)* Manuscript in preparation.
Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science, 24*(1), 27–33.
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.
Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review, 34*(3), 265–293.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.
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.

Paula Rubio-Fernández
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA / Department of Philosophy, University of Oslo, Norway
E-mail address: paula.rubio-fernandez@ifikk.uio.no