Peer review history of the paper *When Does the Story Matter? No Evidence for the Foregrounding Hypothesis in Math Story Problems* by Sabrina M. Di Lonardo Burr, Jill Turner, Jesse Nietmann, & Jo-Anne LeFevre published in Special issue *Direct and Conceptual Replication in Numerical Cognition* in *Journal of Numerical Cognition* (vol 7, 3). https://doi/org/10.5964/jnc.6053

Table of Contents

Initial decision letter

Dear Ms Di Lonardo,

I have now received two reviews of your submission to Journal of Numerical Cognition, "Does the story really matter? No evidence for an effect of the situation model in simple math word problems" from experts in the field. I have also read the ms myself. I wish to thank the reviewers for volunteering their time. As you can see, the reviewers provided numerous and constructive comments. Even though Reviewer B recommended acceptance of the ms as it is, I encourage you to read this review carefully and consider implementing comments, you consider important in the context of your replication effort.

As you might notice, the reviewers differed considerably in their evaluation of your work. Reviewer A was much more skeptical, providing you detailed feedback, and list of concerns, which have to be addressed in the revision. The reviews are clear and I see no reason to recapitulate them.

When reading the ms myself I have noticed that you are reporting the results of the Bayesian analysis in quite an extensive (and redundant) way. As Bayesian methods are gaining popularity, BF values are understood quite well. I would suggest that you provide the BF values. For the readers, who are not yet familiar with them, you may provide a descriptive footnote when you introduce them for the first time. It can be done in a manner you are reporting all the results. As for now, the report with 26-digit number (Results section of Experiment 2) looks a bit awkward to me.

My decision is to invite major revisions. To streamline the review process, please prepare a detailed response to the reviews and mark the changes made to the manuscript.

Reviewer A:

Comments for author(s):

Across two studies, the authors attempted to conceptually replicate and extend findings reported by Mattarella-Micke and Beilock (2010) and Jarosz and Jaeger (2019). Whereas the first authors found that multiplication word problems in which an irrelevant number was associated with the protagonist of the problem (i.e., foregrounded in the text) were solved less accurately than problems in other conditions, Jarosz and Jaeger used similar materials but tested the inconsistent-operations hypothesis that association with the protagonist would interfere with multiplication whereas dissociation would interfere with division. In the present research, the authors conducted two studies, involving quite large samples, in which they similarly manipulated whether irrelevant content was associated with or dissociated from the story protagonist. They did not find support for either the foregrounding or inconsistent-operations hypotheses. According to the authors, their more careful implementation of these manipulations and their much greater power to detect effects, suggest that the two manipulation do not influence adults' performance on simple math story problems.

This study is well-written and well-structured. Moreover, it provides a very clear and accurate review of the directly relevant studies, a very clear report of the rationale, design, analysis and results of the two closely related studies, and a substantial discussion an conclusion section, which includes the necessary elements (summary and discussion of main results, limitations, educational implications…).

Our field needs well-executed replication studies, which pay careful attention to the details of the design of the original and new studies and which try to improve some specific problematic elements in the design or the analysis. As such this study is a good illustration of this kind of research.

However, personally, being a researcher with more affinity with (pychology of) mathematics education, I found the study not very interesting, because of the lack of strong (conceptual) links between these specific studies and the "broader picture" namely the very well established theoretical and empirical research on word problem solving and the role of various kinds of wording effects in the field of (psychology of) mathematics education, and because of the quite trivial nature of the results (I would never expect that the manipulated factors in this very simple word problems would have an effect on the problem solving performance of well educated adults). Stated differently, I found the contribution of this research to the research field of word problem solving quite meagre and I evaluate it quite low in terms of ecological validity.

However, I think that my personal scientific background and personal feelings (as a math educator) about the quality and importance of these replications studies should not play a decisive role in my judgment. Therefore, I think that the study can be accepted for publication in its current state.

Reviewer B:

Comments for author(s):
This research consists in an attempt of replication of previous findings. Adult participants were asked to solve multiplication and division word problems, either in

associative or dissociative contexts. Following the situation model theory, participants' performance should have been influenced by the context in which story problem were embedded. However, the authors did not replicate previous results of the literature

Within the current replicability crisis, attempts of replications are always very welcome. However, as I will detail below, I do not think that this research convincingly questions previous conclusions of the literature.

First, the title of the paper is extremely misleading because the authors question the conclusion of one paper supposedly revealing the mental construction of a situation model during word problem solving and not the existence or the possibility of constructing situation models during this activity. Hundreds and hundreds paper have demonstrated that individuals construct situation models and not only mathematical models when they solve arithmetic word problems and, again, this research does not demonstrate that it is not the case. As a matter of fact, the authors never mention a questioning at a theoretical level in their manuscript, except in the title.

Under the inconsistent operation hypothesis, whereas the idea that addition can interfere with multiplication is straightforward, I have trouble in buying the fact that subtraction could interfere with division. The authors cite two studies that could support this prediction because subtraction might interfere with division when repeated subtraction are used to find the quotient. Nevertheless, I doubt that this strategy is often used by university students to solve problems such as 63 / 9, who should perform :63 -9 = 54, 54 -9 = 45, 45 – 9 = 36, 36 – 9 = 27, 27 – 9 = 18, 18 –9 = 9, 9-9 = 0 and count the number of steps that they needed to reach the answers : 7. If this strategy is not used, the prediction that dissociative problems should lead to interference when a division has to be performed does not make sense. Therefore, if the authors have to examine the inconsistent and foregrounding hypotheses again in the future, they need to record the strategies used by participants to solve the problems.

Anyway, and as stated by the authors, neither the foregrounding or the inconsistent operation hypotheses has already found strong support in the literature. The aim of the authors is therefore to replicate previous results with a better controlled material. Therefore, we expect that division and multiplication problems will be presented to participants both in stories involving dissociating and associating scenarios and both in highly-interfering and less interfering conditions. Unfortunately, this is not the case because the authors do not have high and less interfering conditions for division and, as a consequence, cannot properly examine the inconsistent operation hypothesis. They justify this point by explaining that they were not aware of the study by Jarosz and Jaeger (2019) when they conceived their material. In this case, I wonder whether this publication really has to play such a central role in the authors' rationale and questioning. If yes, I think that the experiment has to be redesigned and conducted with all the relevant conditions. (Incidentally, Table 1 note is wrong because interference is not manipulated for division. I wonder however what was the rationale of the authors for using either a number (15 in their example) or an indefinite determiner (some) in their division problems? Did the authors make a mistake and presented story problems presented exclusively in Experiment 2?).

As stated in my previous comment, the authors explain that the originality of their second study was to use non-numeric word in the text of their problems. However, it

seems to be already the case in Study 1, at least for division problems. It is very confusing.

Finally, the authors decided to present the text of the problem in its whole rather than in a segmented manner as Mattarella-Micke and Beilock (2010) did. One interpretation for the fact that they do not replicate their results is that the higher working memory demand of the task in the original study made them appear. Stated differently, it might be because participants in Mattarella-Micke and Beilock's study were put under cognitive pressure that reasoning or mental process biases were revealed. More generally, I do not think that the authors can really conclude that they do not replicate previous results of the literature when they have not strictly reused (or replicated) the same design as in the original study.

Authors' response

Manuscript: *When does the story matter? No evidence for the foregrounding hypothesis in math story problems*

Below are our responses to comments from the reviews. Comments from the reviewers appear in black font and our responses appear in blue font. Similarly, in the revised manuscript all changes made in response to the reviewers' comments appear in blue font.

We greatly appreciate the suggestions provided. We have carefully considered these suggestions and incorporated them into our manuscript, when appropriate. We believe that these revisions have improved the manuscript.

Editor:

As per your suggestion, we have simplified the reports of the Bayesian analyses. We have also more clearly stated that we are not directly testing the situation model hypothesis or implying that the situation model theory is not supported; rather that we found no evidence for any effects of the specific text and interference manipulations used by Mattarella-Micke and Beilock (2010) and Jarosz and Jaeger (2019).

Reviewer A:

Reviewer A's only concern was that the research was too narrowly focused. Specifically:

However, personally, being a researcher with more affinity with (psychology of) mathematics education, I found the study not very interesting, because of the lack of strong (conceptual) links between these specific studies and the "broader picture" namely the very well established theoretical and empirical research on word problem solving and the role of various kinds of wording effects in the field of (psychology of) mathematics education, and because of the quite trivial nature of the results (I would never expect that the manipulated factors in this very simple word problems would have an effect on the problem solving performance of well-educated adults). Stated

differently, I found the contribution of this research to the research field of word problem solving quite meagre and I evaluate it quite low in terms of ecological validity.

However, I think that my personal scientific background and personal feelings (as a math educator) about the quality and importance of these replications studies should not play a decisive role in my judgment. Therefore, I think that the study can be accepted for publication in its current state.

Thanks for this feedback. We similarly had doubts about the ecological validity of the manipulations in the previous papers which is why we attempted to replicate them. If replicable, such findings support a view in which almost any aspect of a word problem could influence performance, even when the text was irrelevant to the required computation. In that sense, we felt that a replication of this strong conclusion was important to establish boundary conditions on the foregrounding hypothesis. To address this issue more directly in the paper, we have stated that our results suggest that there are limitations on the extent to which textual manipulations might influence problem solving performance, especially in skilled adults. We have tried to make this focus/issue clearer throughout the paper (e.g., see the added text on pages 4, 5 and 7), and especially in the final section (see page 30-31). We also changed the title of the manuscript to reflect the scope of the research more accurately (see Reviewer B's comments).

Given that we did not find any effects of associative/dissociative language, we attempted to add value to the work by exploring the actual errors that participants made. These analyses showed that errors were consistent with the literature on arithmetic performance – and thus that performance was essentially independent of the textual manipulations. In our view, the value of this paper is to refute the strong claims, in particular of Mattarella-Micke and Beilock, and show that the errors that people made on these problems has little to do with these specific textual manipulations.

Reviewer B:

This research consists in an attempt of replication of previous findings. Adult participants were asked to solve multiplication and division word problems, either in associative or dissociative contexts. Following the situation model theory, participants' performance should have been
influenced by the context in which story problem were embedded. However, the authors did not replicate previous results of the literature.

Within the current replicability crisis, attempts of replications are always very welcome. However, as I will detail below, I do not think that this research convincingly questions previous conclusions of the literature.

First, the title of the paper is extremely misleading because the authors question the conclusion of one paper supposedly revealing the mental construction of a situation model during word problem solving and not the existence or the possibility of constructing situation models during this activity. Hundreds and hundreds paper have demonstrated that individuals construct situation models and not only mathematical models when they solve arithmetic word problems and, again, this research does not

demonstrate that it is not the case. As a matter of fact, the authors never mention a questioning at a theoretical level in their manuscript, except in the title.

Thank you for your feedback. We agree, we are not trying to question the situation model theory in general and so we have modified the title and checked that we did not make that claim in the paper. The general claim that wording can affect math problem solving is not in doubt, however, the specific manipulations and hypotheses involved in the present research seemed weak and unlikely to strongly influence adults' solutions (see also our response to Reviewer A). The idea that foregrounding in relation to a story protagonist could influence the construction of a situation model, as proposed by Mattarella-Micke & Beilock, was very interesting. When we read their paper, however, we felt that there may be some boundary conditions on the extent to which textual manipulations can be expected to influence problem solving performance, especially for skilled adults.

We avoided any extensive theoretical discussion in this paper because the instructions for the special issue emphasized the empirical replication. Thus, we only outlined/explained the theories/hypotheses from the previous research and focused on whether we could reproduce the results of the target papers.

Under the inconsistent operation hypothesis, whereas the idea that addition can interfere with multiplication is straightforward, I have trouble in buying the fact that subtraction could interfere with division. The authors cite two studies that could support this prediction because subtraction
might interfere with division when repeated subtraction are used to find the quotient. Nevertheless, I doubt that this strategy is often used by university students to solve problems such as 63 / 9, who should perform : 63 -9 = 54, 54 -9 = 45, 45 – 9 = 36, 36 – 9 = 27, 27 – 9 = 18, 18 –9 = 9, 9-9 = 0 and count the number of steps that they needed to reach the
answers : 7. If this strategy is not used, the prediction that dissociative problems should lead to interference when a division has to be performed does not make sense. Therefore, if the authors have to examine the inconsistent and foregrounding hypotheses again in the future, they need to record the strategies used by participants to solve the problems.

We agree. It seemed unlikely that skilled adults would use repeated subtraction to solve these problems. Adults are more likely to solve division problems (if they don't retrieve the answers) by reframing as a multiplication problem (e.g., 6 x __ = 54) and then using their multiplication knowledge to evaluate possible answers. Occasionally they may use repeated subtraction, but it is highly inefficient and error prone. Repeated addition would be slightly more likely (i.e., skip counting 9, 18, 27, 36, 45, 54 to figure out that there were 6 nines in 54). In the introduction, we provide information about the source of the inconsistent operations hypothesis as proposed by Jarosz and Jaeger (and tested with university students). However, we are not claiming that we expected those results. In fact, because we designed the study before the publication of Jarosz and Jaeger, we did not have a highly-interfering division condition. We chose to use the interfering number in the division problems in Study 1 because Mattarella-Micke and Beilock did not describe their division stimuli. In Study 2, we were attempting to make the interfering number and the foregrounding condition

as salient as possible for multiplication, and so the division problems were matched to the non-interfering word condition.

Collecting strategy data is always interesting. However, the error analyses that we presented suggest that errors are related to the arithmetic calculations, not to influences of the associative story content. If people had added instead of multiplied or subtracted instead of divided, very specific errors would have been observed. The errors we observed are consistent with effects of operation-confusion and operand-interference, effects that are well established in the literature on simple arithmetic problems. Strategy data are unlikely to have changed the conclusions.

Anyway, and as stated by the authors, neither the foregrounding or the inconsistent operation hypotheses has already found strong support in the literature. The aim of the authors is therefore to replicate previous results with a better controlled material. Therefore, we expect that
division and multiplication problems will be presented to participants both in stories involving dissociating and associating scenarios and both in highly-interfering and less interfering conditions. Unfortunately, this is not the case because the authors do not have high and less interfering conditions for division and, as a consequence, cannot properly examine the
inconsistent operation hypothesis. They justify this point by explaining that they were not aware of the study by Jarosz and Jaeger (2019) when they conceived their material. In this case, I wonder whether this publication really has to play such a central role in the authors' rationale and questioning. If yes, I think that the experiment has to be redesigned and conducted with all the relevant conditions. (Incidentally, Table 1 note is wrong because interference is not manipulated for division. I wonder however what was the rationale of the authors for using either a number (15 in their example) or an indefinite determiner (some) in their division problems? Did the authors make a mistake and presented story problems presented exclusively in Experiment 2?).

We changed the Table 1 note content to more accurately reflect the division manipulation (p. 10).

The focus of our replication was on the original Mattarella-Micke and Beilock (2010) studies, but because Jarosz and Jaeger (2019) also tried to replicate and extend these studies it does not seem right to ignore or minimize their studies in our manuscript. Instead, we have tried to be more cautious about what aspects of their work that overlap with ours. Notably, Jarosz and Jaeger did not find any effects of an interfering number on multiplication, nor were their findings for division dependent on any number at all being present in the associative/dissociative component of the story problem (see their Study 3). Thus, repeating the study to include highly-interfering and less-interfering conditions for division seems unlikely to change the results. We also stress that our error analysis did not support the view that people used addition on multiplication problems or subtraction on division problems, as Jarosz and Jaeger propose in their inconsistent operations hypothesis.

As mentioned, the Jarosz and Jaeger paper had not been published when we designed our stimuli. For Study 1, our rationale was that division problems should not just be treated as fillers (as they were in Mattarella-Micke and Beilock), but rather we should test to see if there is an effect of association/dissociation. To be consistent with

the formatting of the problems, a number was included in division problems, but it was neither highly- nor less-interfering; it was simply extraneous numeric information. For Study 2, we matched the division text to that of the non-interfering multiplication condition.

As stated in my previous comment, the authors explain that the originality of their second study was to use non-numeric word in the text of their problems. However, it seems to be already the case in Study 1, at least for division problems. It is very confusing.

As mentioned in the previous comment, Study 1 did not have "high" and "less" numeric interference for division problems. Mattarella-Micke and Beilock did not provide any information about the composition of their division problems and so we decided to match the "interfering" division number to the highly-interfering multiplication number. For example, if the original division problem was 6 x 9, the highly-interfering number 15 was used for the division problem 54 / 9. This decision is explained on page 9 in the manuscript.

In Study 2, the associative/dissociative portion of division problems matched the non-numeric conditions for multiplication and thus was very similar to the manipulation used by Jarosz and Jaeger in their Study 3. We have attempted to clarify this in the manuscript (e.g., see page 16-17).

Finally, the authors decided to present the text of the problem in its whole rather than in a segmented manner as Mattarella-Micke and Beilock (2010) did. One interpretation for the fact that they do not replicate their results is that the higher working memory demand of the task in the original study made them appear. Stated differently, it might be because participants in Mattarella-Micke and Beilock's study were put under cognitive pressure that reasoning or mental process biases were revealed. More generally, I do not think that the authors can really conclude that they do not replicate previous results of the literature when they have not strictly reused (or replicated) the same design as in the original study.

Although Mattarella-Micke and Beilock did present the initial scenarios first, it is not clear whether those scenarios disappeared when the participant pressed the space bar to "continue the problem" (see their page 108). Their wording is ambiguous. When we decided to try and replicate their work, in the absence of any communication from them, we had to decide on a procedure. We chose to maximize ecological validity by presenting the whole story problem simultaneously. If anything, this procedure should have increased the cognitive demand required by *not* separating the irrelevant text from the main content of the story problem while the participant was coming up with a solution. Moreover, Jarosz and Jaeger also did not replicate the multiplication results of Mattarella-Micke and Beilock, even though the introductory material disappeared after it was read in their studies.

Overall, the working memory demands were the same between our studies and those of Mattarella-Micke and Beilock. The *relevant* problem text was visible until the participant responded. After they responded and the problem text disappeared, participants either rated the clarity of the texts (in Mattarella-Micke and Beilock, see their page 108) or they answered questions about the texts (in our studies). So,

although we did not provide an exact replication of their procedure, we feel that it is a solid conceptual replication, such that this minor procedural detail was unlikely to have caused or even influenced the results they obtained. We have explained this on pages 26-27.

But more generally, it is important to note there is no evidence that the cognitive load was greater in the two previous papers than in the current research – note how similar the overall error percentages are to ours, for example. Cognitive load was not manipulated in any of the papers. Thus, potential differences in cognitive load is not a plausible explanation of the lack of replication.

To be thorough, we have also included a new section about the working memory results from the other papers, even though we did not explore individual differences in working memory as a potential factor. The previous papers showed that, in some conditions or for some groups, individual differences in working memory capacity were related to performance (see summary on pages 28-29). When we conceived this work, we wanted to first establish that the interfering-number and the associative-dissociative manipulations would give robust results and so we focused on that goal. Any studies on the effects of individual differences in working memory capacity would depend upon having strong materials with well-established effects.

Second decision letter

Dear Sabrina Michelle Di Lonardo Burr,

Thank you for your careful revision. Your article entitled "Does the story really matter? No evidence for an effect of the situation model in simple math word problems" has now been accepted for publication in the Journal of Numerical Cognition (JNC) – congratulations!

At the very bottom of this email you can see the final thoughts of the reviewers. As you can see, Reviewer B remained not convinced to your paper. At the same time they acknowledged that their comments have been properly addressed and that the changes in theoretical framing of the manuscript make it way less problematic than it was before.

Also reviewer A commented on how you dealt with the feedback of reviewer B, which again assured me that the revision was careful and adequate.

Problems of not following Mattarella-Micke and Beilock procedure remain valid, and they are acknowledged properly, which in this situation, according to my judgment is enough. What we as the field with huge potential of educational implications need to assure is that findings remain robust even if specific differences in procedures / interventions appear. This is because one can hardly expect that the interventions moved to the classroom will follow the exact protocol of a single experiment, which proved their efficacy. Therefore, in my view the replication effort should be published as it at least points that the effect under scrutiny is not as robust as could be inferred from the existing literature, and hopefully in the long run contributes to reduction in publication bias.

Reviewer A:

Comments for author(s):
I recommended already to accept the paper in the previous round. I accept the authors' reaction to my remaining concerns. I have read, with great interest, the comments of the other reviewer, who raised several important and critical issues both about the internal validity of the study and about the way in which the research is being framed in relation to previous work on which it is based. My personal opinion is that all these critical comments of this reviewer make a lot of sense, but at the same time I also think that the authors did their very best to defend in their response letter and to explain in the revised manuscript what they intended to do with the study, why they gave ample attention in the manuscript to the recently published very relevant paper they were not aware of the moment they designed their study, and why they made certain methodological choices in the absence of clear information in the original study they intended to replicate. Not all responses are completely satisfying but I think the authors did a good job in defending their study and in improving the clarity of the argumentation and explanation through this revision. Therefore, I think the revised paper deserves to be accepted.

Reviewer B:

Comments for author(s):
This is my second evaluation of the manuscript. I've carefully read the answers given by the authors to my comment and I think that they are overall satisfactory.
Noticeably, having changed the title and consequently the main message of the manuscript is wise because the present research does not question at all the relevance of the situation model framework. Nevertheless, my remaining concern is related to the fact that the authors did not replicate exactly the methodology used by Mattarella-Micke and Beilock. I'm aware that it is now addressed explicitly in the Discussion but still, I do not know whether it is sufficient to minimize the detrimental effect it has on the quality of the research.