**Peer Review and Communication History**

**MS Title:** Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills

**Author Names:** Lara M. van Peppen, Peter P. J. L. Verkoeijen, Anita E. G. Heijltjes, Eva M. Janssen, and Tamara van Gog

**Submitted:** Dec 6, 2020

**Editor First Decision: Revise & Resubmit**

March 30, 2021

Dear L.M. van Peppen,

I have now received all reviews of your manuscript, “Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills” from qualified researchers. I also independently read the manuscript before consulting these reviews.

The reviewers were particularly conflicted here, with two reviewers recommending acceptance as-is and the third reviewer recommending rejection. Here I am making the decision that you consider the critical feedback of the the third reviewer and submit a revised version and I will likely make a decision on that version without sending it out for further review–though it may still be either acceptance or rejection. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next two months. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Christopher Madan

# Reviewer 1

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I think this is well written, with a clear introduction/rationale, description of methods, and presentation of results. The authors have also pre-empted many questions/thoughts readers would have in their discussion section, particularly around potential weaknesses in their experimental design, number of participants, and reasons why certain results may have been found.  
Table 1 is a little confusing, could the italicised text that indicates procedures were not part of the study just be deleted?

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

# Reviewer 2

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Thank you for the opportunity to review the manuscript “Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills”.

First, I would like to appreciate that author(s) have pre-registered the study on OSF. The problem of repeated retrieval practice as a way to improve critical thinking skills is interesting. I want to congratulate the authors on well-prepared research and a well-written study. I do not have any questions (the authors addressed all my questions and comments before I asked them).

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

# Reviewer 3

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The study aim at investigating whether using retrieval practice (RP) in combination with distributed practice can enhance critical thinking, operationalized by the ability to avoid biases in reasoning and decision making.  
The authors argue that RP, as a learning strategy in combination distributed practice is a robust phenomenon, proven to be effective in educational contexts, across material, and test formats.  
The authors hypothesized that explicit CT-instructions combined with retrieval practice would be effective for learning: thus, they expected;

1. pretest to posttest performance gains on learning items in all conditions (Hypothesis 1).
2. that practicing twice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2a) and a higher posttest performance on transfer items (Hypothesis 3a) than practicing retrieval once.
3. We expected that retrieval thrice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2b) and a higher posttest performance on transfer items (Hypothesis 3b) than practicing retrieval twice.
4. However, as cognitive psychology research has revealed that practicing twice leads to a large learning gain with diminishing returns for more repetitions, we expected these differences to be smaller than the differences between practicing retrieval once and twice.  
   The authors posted their hypothesis in advance on the open since framework(OSF)  
   A mixed design with distributed practice ( once, twice triche) as between-subject factor and pre-post and within-subject

I really like attempts to broaden the vast literature on RP and trying to evaluate potential benefits when we should use retrieval practice and, as important, when we should consider NOT use retrieval practice. The conclusion from this study seems on the face of it to be: Don’t use retrieval practice in this context – the effects are almost nonexistent (although progress was seen within the practice sessions). With that said, I need to question how the study was designed and how RP was used.

Introduction  
I think it was an adequate introduction with respect to both CT and RP. Concerning RP (which I know better) there are many studies to choose from. Since this is a classroom-based study, perhaps RP review on classroom studies would be good to include, for instance, Moreira et al. (2019)

However, I disagree with the conclusion implied that for deep learning, spacing is required. Although repeated RP always include spacing (hence repeated) the spacing of RP increase the mental effort and thereby enhance the testing effect- not the spacing per se, rather the effort associated with using spacing

Moreover, and as I think critically is the statement that “practicing twice will lead to a large learning gain, with diminishing returns for practicing three times and four times” is incorrect. With support from Rawson and Dunlosky (2011) statement “With both short-term and long-term goals in mind, our pick is the 3 \_ 3 schedule (i.e., practicing to three correct recalls during initial learning followed by three subsequent relearning sessions)” I argue that the critical aspect is successful retrieval, hence how many times an item is successfully retrieved. Which seem to be around three to four times, which in turn require perhaps ( on average) 6 repetitions before you reach 3-4 successful retrievals. Or, as Roediger Iii and Butler (2011) argued, there is a need for 5-7 retrievals, though depending on the spacing. Also in (Wiklund-Hornqvist et al., 2021), it was shown that brain activity (Hippocampus) in an RP paradigm was dependent on items having been retrieved multiple times during the initial learning session,

The statement on page 3 “One of the strongest learning techniques known to promote the construction of well-developed knowledge structures, is having students repeatedly retrieve to-be-learned material from memory, known as repeated retrieval practice” made me wonder about the statement “the construction of well developed knowledge structures” What does that really mean? I would argue that RP enhances or strengthen the memory consolidation of the practiced information, which makes subsequent retrieval easier/more effective. Hence repeated retrieval seems to reduce the cognitive load on the frontal lobe and boost long-term memory retrieval – potentially by inducing representation variability(Karlsson Wirebring et al., 2015). With that said, you should elaborate and substantiate your statement regarding “the construction of well-developed knowledge structures.”

2.2.3 CT practice.  
I would like the RP to be more clearly described. Did they get the MC question at the same time as they read the practice task? Hence they could look at the MC question at the same time as the premises were displayed; the which mean that they did not have to “reinstate” . So my question is, where were the retrieval components in the practice tasks. You have to be explicit in that respect

Hypotheses  
I don’t understand why you excluded the most apparent hypothesis, the Moment × Condition interaction, and instead went directly on the simple effects hypotheses ( e.g., whether RP twice is better than RP once). Hence the hypotheses should have addressed the:

1. pretest to posttest performance gains on learning items in all conditions (Hypothesis 1).
2. Moment × Condition interaction  
   This also corresponds to the analyses conducted in the paper

With respect to the design and the argument that you investigate “ repeated retrieval effects,” To get a RP effect ( hence a testing effect), you need something to compare with ( classic comparison is RP against study). If you would have gained an effect (gain from pre- to post-test) this could be caused by a test-retest- effect from pre to post-test ( which is common to get). Moreover, the option to address moment × condition interaction is also problematic since the number of repetitions is confounded with spacing effects

In sum, I think the study has many problems as outlined above. and I am skeptical that a major revision will help, and therefore recommended rejection

Karlsson Wirebring, L., Wiklund-Hornqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015, Jul 1). Lesser Neural Pattern Similarity across Repeated Tests Is Associated with Better Long-Term Memory Retention. J Neurosci, 35(26), 9595-9602. <https://doi.org/10.1523/JNEUROSCI.3550-14.2015>

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019, 2019-February-08). Retrieval Practice in Classroom Settings: A Review of Applied Research [Review]. Frontiers in Education, 4(5). <https://doi.org/10.3389/feduc.2019.00005>

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? Journal of Experimental Psychology: General, 140(3), 283-302. <https://doi.org/10.1037/a0023956>

Wiklund-Hornqvist, C., Stillesjo, S., Andersson, M., Jonsson, B., & Nyberg, L. (2021, Jan). Retrieval practice facilitates learning by strengthening processing in both the anterior and posterior hippocampus. Brain Behav, 11(1), e01909. <https://doi.org/10.1002/brb3.1909>

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  | ✔ |  |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) | ✔ |  |  |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  | ✔ |  |  |  |

**Author Response**  
Aug 30, 2021

**Response to Reviewer Comments and Suggestions**

“Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills”

***Dear Editor, dear Dr. Christopher Madan,***

***Thank you for providing us with the opportunity to revise and resubmit our manuscript. We would also like to thank the reviewers for the thoughtful comments and suggestions, which we feel helped us improve our manuscript. Please find below our reply (in italics and blue) clarifying the rationale of our study and detailing the changes we made in response to each of the comments (plain text) made in the reviews.*Editor’s comments**  
I have now received all reviews of your manuscript, “Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills” from qualified researchers. I also independently read the manuscript before consulting these reviews.  
  
The reviewers were particularly conflicted here, with two reviewers recommending acceptance as-is and the third reviewer recommending rejection. Here I am making the decision that you consider the critical feedback of the third reviewer and submit a revised version and I will likely make a decision on that version without sending it out for further review–though it may still be either acceptance or rejection. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.  
  
In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.  
  
*Thank you for the opportunity to revise and resubmit our manuscript.* *Please see our responses to the reviewers’ comments for detailed explanations of the changes we made in our revised manuscript.*

**Reviewer 1**I think this is well written, with a clear introduction/rationale, description of methods, and presentation of results. The authors have also pre-empted many questions/thoughts readers would have in their discussion section, particularly around potential weaknesses in their experimental design, number of participants, and reasons why certain results may have been found.

*Thank you for these positive comments.*

Table 1 is a little confusing, could the italicised text that indicates procedures were not part of the study just be deleted?

*This comment seems to refer to figure 1 instead of Table 1. We have now deleted the text indicating procedures that were not part of the study from figure 1. This has already been described in the procedure subsection.*

**Reviewer 2**  
Thank you for the opportunity to review the manuscript “Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills”.  
  
First, I would like to appreciate that author(s) have pre-registered the study on OSF. The problem of repeated retrieval practice as a way to improve critical thinking skills is interesting. I want to congratulate the authors on well-prepared research and a well-written study. I do not have any questions (the authors addressed all my questions and comments before I asked them).  
  
*Thank you for these positive comments.***Reviewer 3**  
The study aim at investigating whether using retrieval practice (RP) in combination with distributed practice can enhance critical thinking, operationalized by the ability to avoid biases in reasoning and decision making.

The authors argue that RP, as a learning strategy in combination distributed practice is a robust phenomenon, proven to be effective in educational contexts, across material, and test formats.

The authors hypothesized that explicit CT-instructions combined with retrieval practice would be effective for learning: thus, they expected;

1. pretest to posttest performance gains on learning items in all conditions (Hypothesis 1).
2. that practicing twice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2a) and a higher posttest performance on transfer items (Hypothesis 3a) than practicing retrieval once.
3. We expected that retrieval thrice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2b) and a higher posttest performance on transfer items (Hypothesis 3b) than practicing retrieval twice.
4. However, as cognitive psychology research has revealed that practicing twice leads to a large learning gain with diminishing returns for more repetitions, we expected these differences to be smaller than the differences between practicing retrieval once and twice. The authors posted their hypothesis in advance on the open since framework(OSF) A mixed design with distributed practice ( once, twice triche) as between-subject factor and pre-post and within-subject

I really like attempts to broaden the vast literature on RP and trying to evaluate potential benefits when we should use retrieval practice and, as important, when we should consider NOT use retrieval practice. The conclusion from this study seems on the face of it to be: Don’t use retrieval practice in this context – the effects are almost nonexistent (although progress was seen within the practice sessions). With that said, I need to question how the study was designed and how RP was used.

*Thank you for these positive comments. Furthermore, it is important to point out that our study is not a classical retrieval practice study in which a retrieval practice condition is pitted against a restudy control condition. Instead, the central question in the present study was whether repeated retrieval practice would be beneficial for the acquisition of critical thinking (in particular: unbiased reasoning) skills; that is, whether repeated retrieval practice would be positively related to measures of learning and transfer of unbiased reasoning. To address this question, we compared three conditions. Retrieval practice occurred in all conditions, but the frequency of retrieval practice differed. The conclusion of our study is that additional retrieval practice (compared to a control condition in which people practiced retrieval on six tasks) appears to have no added value with regard to the outcome measures. However, we cannot conclude from the results of the present study that retrieval practice should not be used for learning the critical thinking tasks we employed because we - intentionally - did not use a no-retrieval-practice control condition.*

Introduction

I think it was an adequate introduction with respect to both CT and RP. Concerning RP (which I know better) there are many studies to choose from. Since this is a classroom-based study, perhaps RP review on classroom studies would be good to include, for instance, Moreira et al. (2019). However, I disagree with the conclusion implied that for deep learning, spacing is required. Although repeated RP always include spacing (hence repeated) the spacing of RP increase the mental effort and thereby enhance the testing effect- not the spacing per se, rather the effort associated with using spacing.

Moreover, and as I think critically is the statement that “practicing twice will lead to a large learning gain, with diminishing returns for practicing three times and four times” is incorrect. With support from Rawson and Dunlosky (2011) statement “With both short-term and long-term goals in mind, our pick is the 3 x 3 schedule (i.e., practicing to three correct recalls during initial learning followed by three subsequent relearning sessions)” I argue that the critical aspect is successful retrieval, hence how many times an item is successfully retrieved. Which seem to be around three to four times, which in turn require perhaps ( on average) 6 repetitions before you reach 3-4 successful retrievals. Or, as Roediger Iii and Butler (2011) argued, there is a need for 5-7 retrievals, though depending on the spacing. Also in (Wiklund-Hornqvist et al., 2021), it was shown that brain activity (Hippocampus) in an RP paradigm was dependent on items having been retrieved multiple times during the initial learning session,

*The feedback of the Reviewer on our Introduction made us realize that the reference to the spacing literature might not have been very conducive to the clarity of the line of reasoning, so we rewrote the introduction and dropped the part on and the references to spaced repetition as this is not the focus on the present study. Moreover, we disagree with the reviewer that our statement regarding diminishing returns would be incorrect. However, this did make us realize that we should have made a stronger case to support our claim that additional retrieval practice has diminishing returns on the final test performance. That is, we fully agree with the Reviewer, that the number of successful retrieval practice attempts is positively related to the final test performance. However, we also argue that the added benefit of extra retrieval practice testing decreases because the retrieval effort decreases for successful retrieval attempts. We now explain this in more detail in the revised introduction. Furthermore, we included the useful literature suggestion of the Reviewer. The revised version of the text is presented below (see pages 3 and 4 of the revised manuscript):*

*“One of the strongest learning techniques known to promote the construction of meaningful knowledge structures, is having students retrieve to-be-learned material from memory, known as practice testing or retrieval practice (e.g., Dunlosky et al., 2013; Fiorella & Mayer, 2015, 2016; Roediger & Butler, 2011). The effect of retrieval practice seems to be extremely robust (for reviews, see Carpenter, 2012; Delaney et al., 2010; Moreira et al., 2019; Rickard & Pan, 2017; Roediger & Butler, 2011; Roediger & Karpicke, 2006a; Pan & Rickard, 2018; Rowland, 2014) emerging on measures of both learning and transfer, and with different kinds of materials and test formats (e.g., Butler, 2010; Carpenter & Kelly, 2012; McDaniel et al., 2012, 2013; Rohrer, Taylor, & Sholar, 2010).*

*1.1 Repeated retrieval practice*

*The effect of retrieval practice seems to be positively related to the number of successful retrieval attempts during practice (e.g., Greene, 1989; Rawson & Dunlosky, 2011; Roediger & Karpicke, 2006b), albeit with diminishing returns. For example, in Experiment 2 from the study by Roediger and Karpicke (2006b), participants either studied a prose passage multiple times (SSSS condition), studied a prose passage multiple times and took one free recall retrieval practice test (SSST condition) or studied a prose passage once and took a free recall retrieval practice test thrice (STTT condition). Subsequently, a delayed final free recall test on the prose passage was administered in all conditions. The results on this final free recall test showed that taking a single retrieval practice test increased the free recall performance relative to the control condition from a mean score of 40% correct to a mean score of 56% correct. Furthermore, repeated retrieval practice (i.e., the STTT condition) increased the free recall performance to a mean of 61% correct, hence showing diminishing returns for extra retrieval practice. That is, where a single retrieval practice test in the SSST condition lifted final test performance with 16% points, the two additional retrieval practice tests increased the final test performance with only 5% points. These diminishing returns of repeated retrieval practice might be due to the fact that the practice testing effect depends not only on the number of successful retrieval attempts but also on the effort that is required to successfully retrieval information from memory. According to the retrieval effort hypothesis (e.g., Rawson & Pyc, 2009) the effect of retrieval practice becomes larger when successful retrieval attempts require more effort. When information is repeatedly retrieved from memory, the effort associated with successful retrieval is likely to decrease, which will lead to diminishing returns of repeated retrieval practice.”*

The statement on page 3 “One of the strongest learning techniques known to promote the construction of well-developed knowledge structures, is having students repeatedly retrieve to-be-learned material from memory, known as repeated retrieval practice” made me wonder about the statement “the construction of well developed knowledge structures” What does that really mean? I would argue that RP enhances or strengthen the memory consolidation of the practiced information, which makes subsequent retrieval easier/more effective. Hence repeated retrieval seems to reduce the cognitive load on the frontal lobe and boost long-term memory retrieval – potentially by inducing representation variability(Karlsson Wirebring et al., 2015). With that said, you should elaborate and substantiate your statement regarding “the construction of well-developed knowledge structures.”

*Repeated retrieval practice contributes to more than just memory consolidation. We now elaborated more on what well-developed knowledge structures are and how these can aid future problem solving (see page 3 of the revised manuscript).*

*“Previous research has demonstrated that to establish learning and transfer, learners have to actively construct meaningful knowledge from to-be-learned information, by mentally organizing it in coherent knowledge structures and integrating these with one’s prior knowledge. (Bassok & Holyoak 1989; Fiorella & Mayer, 2016; Gick & Holyoak, 1983; Holland et al, 1986; Wittrock, 2010). This, in turn, can aid future problem solving (Kalyuga, 2011; Renkl, 2014; Van Gog et al., 2019): if a situation presents similar requirements and the learner recognizes them, they may select and apply the same or a somewhat adapted learned procedure to solve the problem.”*

2.2.3 CT practice

I would like the RP to be more clearly described. Did they get the MC question at the same time as they read the practice task? Hence they could look at the MC question at the same time as the premises were displayed; the which mean that they did not have to “reinstate” . So my question is, where were the retrieval components in the practice tasks. You have to be explicit in that respect

*In our study, participants had to deliberately recall information on multiple tasks during one practice session, two practice sessions, or three practice sessions. We now provided this information in our methods section (p.8):*

*“Participants were instructed to read the problems thoroughly and to choose the correct MC-answer option, provided directly below the problems. They had to deliberately recall the relevant information from their memory to solve the problems.”*

*Offering the MC-questions directly below the problems does not prevent students from engaging in retrieval processes. On the contrary, it might encourage them to actively process the information and retrieve the relevant information from their memory to successfully solve the problems. See for instance the conditional syllogism explained in the Appendix (p.30). Students were provided with the problem (consisting of two premises and a conclusion), the MC-answer options, and an open text field. To correctly answer the MC-question, that is, whether the conclusion follows logically from the premises, they had to retrieve the rules of logic that were taught in the instruction phase in week 1, from their long-term memory and to apply that knowledge onto the task.*

Hypotheses

I don’t understand why you excluded the most apparent hypothesis, the Moment × Condition interaction, and instead went directly on the simple effects hypotheses ( e.g., whether RP twice is better than RP once). Hence the hypotheses should have addressed the:

1. pretest to posttest performance gains on learning items in all conditions (Hypothesis 1).
2. Moment × Condition interaction  
   This also corresponds to the analyses conducted in the paper

*Our hypotheses do include the interaction between Test Moment and Condition (as do the analyses). Hypothesis 1 refers to the main effect of Moment. For the learning items, we expected the mean post-test score to be higher than the mean pre-test. Hypothesis 2a and Hypothesis 2b concern the expected differences between conditions in performance gains on learning items; these hypotheses reflect the Moment x Condition interaction. Given that the transfer items were not included in the pretest, we were not able to test transfer gains and, therefore, Hypothesis 3a and Hypothesis 3b concern the expected differences between conditions on posttest performance on the transfer items. In response to the Reviewer’s point of feedback, we adapted the text on the hypotheses to make it clearer that we expected this interaction. The revised text (see page 5 of the revised manuscript) is presented below:*

*“We hypothesized that explicit CT-instructions combined with retrieval practice would be effective for learning: thus, we expected an overall mean pretest to posttest performance gain on learning items in all conditions (Hypothesis 1). Furthermore, and more importantly, we expected that practicing retrieval twice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2a) and a higher posttest performance on transfer items (Hypothesis 3a) than practicing retrieval once. We expected that practicing retrieval thrice would lead to a higher pretest to posttest performance gain on learning items (Hypothesis 2b) and a higher posttest performance on transfer items (Hypothesis 3b) than practicing retrieval twice. However, as outlined before, prior research suggests that additional retrieval practice will have diminishing returns on the final test, so we expected these differences to be smaller than the differences between practicing retrieval once and twice.”*

With respect to the design and the argument that you investigate “ repeated retrieval effects,” To get a RP effect ( hence a testing effect), you need something to compare with ( classic comparison is RP against study). If you would have gained an effect (gain from pre- to post-test) this could be caused by a test-retest- effect from pre to post-test ( which is common to get). Moreover, the option to address moment × condition interaction is also problematic since the number of repetitions is confounded with spacing effects

*It is important to point out that we did not aim to investigate a standard RP effect, which would indeed require comparing RP against restudy. Instead, and we hope that this is made clearer in the revised Introduction (see our reaction to a previous point of feedback of the Reviewer), we interested in the impact of repeated retrieval on learning and transfer of CT. More specifically, we investigated whether learning and transfer of CT-skills, would increase with increasing retrieval practice opportunities (i.e., higher pretest to posttest gains on learning items and higher posttest performance on transfer items).*

*In reaction to the pre-test design issue brought up by the Reviewer, we can say the following. First, we included a pre-test in our design to assess whether learning occurred, which is important in educational psychological experiments that focus on instruction and practice. Also, a pre-test can be helpful to interpret the results in case the conditions do not differ on mean post-test performance. Based on the pre-test post-test comparison, one could then determine whether a lack of post-test difference was due to the fact that participants have not learned anything at all (i.e., did not improve from pre-test to post-test), or that the learning gains were similar across conditions. Second, the Reviewer is correct that pre-testing in itself might have an effect. This might have led to interpretation problems had we found that all conditions had revealed comparable, non-zero gain scores. However, given the items that we used on the pre-test and the procedure during the pre-test, it seems unlikely that pre-testing alone would have a strong effect on the post-test scores. Specifically, participants did not receive any feedback on their pre-test performance and, thus, they did not know whether their answers were correct or not, making it unlikely that they learned directly from taking the pretest. Moreover, we know from prior studies on teaching unbiased reasoning (e.g. Heijltjes et al., 2015), that it is the specific combination of instruction and practice that is effective for learning unbiased reasoning; practice alone without preceding instruction did not improve learning. This makes it unlikely that pretesting in and of itself provides a learning opportunity with these kinds of tasks.*

*Regarding the comment that the number of repetitions is confounded with spacing effects: if we had found a difference between conditions 2 and/or 3 and the control condition, this would indeed have been an issue. In this case, however, spacing cannot be confound because the gains are similar between the conditions.*

In sum, I think the study has many problems as outlined above. and I am skeptical that a major revision will help, and therefore recommended rejection

Karlsson Wirebring, L., Wiklund-Hornqvist, C., Eriksson, J., Andersson, M., Jonsson, B., & Nyberg, L. (2015, Jul 1). Lesser Neural Pattern Similarity across Repeated Tests Is Associated with Better Long-Term Memory Retention. J Neurosci, 35(26), 9595-9602. <https://doi.org/10.1523/JNEUROSCI.3550-14.2015>

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019, 2019-February-08). Retrieval Practice in Classroom Settings: A Review of Applied Research [Review]. Frontiers in Education, 4(5). <https://doi.org/10.3389/feduc.2019.00005>

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? Journal of Experimental Psychology: General, 140(3), 283-302. <https://doi.org/10.1037/a0023956>

Wiklund-Hornqvist, C., Stillesjo, S., Andersson, M., Jonsson, B., & Nyberg, L. (2021, Jan). Retrieval practice facilitates learning by strengthening processing in both the anterior and posterior hippocampus. Brain Behav, 11(1), e01909. <https://doi.org/10.1002/brb3.1909>

**Editor Final Decision: Accept**

Sep 21, 2021

Dear L.M. van Peppen,

I have now had a chance to read over your manuscript “Repeated Retrieval Practice to Foster Students’ Critical Thinking Skills”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,  
Christopher Madan