This is a very interesting stage 1 Meta-analysis on the protective effect of retrieval practice on memorization when stressed. Since it is the first round of review, I have plenty of commentaries. All the following are about the length of the manuscript, the writing, and the easiness to read. At the end of this review, I have several more important questions that I would like to be answered because they are more crucial in the realization of this meta-analysis.

Commentaries about the write-up of the MA

Could you please add the osf link in the abstract?

Page 2 you can put the names of the contributors, or at least the first letters (MM)

Page 6 I don’t see the need for the last paragraph and it does not coordinate with the flow of the above and below sections. (The one beginning with “one explanation” to “contextual retrieval”.

Indeed, before you talk about the operationalization of stress and below the protective effect for retrieval practices. So yeah, I don’t see how this section belongs here (I would prefer you to keep it for the discussion actually).

p. 7 the term impervious seems to be too strong. I am not sure that in any part of psychology, something is impervious. Maybe use less sensitive.

-> Now that I read the whole MA, I get that what you mean is that retrieval practice protects from the detrimental effect of stress on memorization. But then you need to have a strong hypothesis here: does RP completely destroy the effect of stress on memorization (then use impervious) or does it lower of loss of memory compared to other practices? This is a more global commentary that you need to think about for hypothesis 1.

p.9 same observation. Since it is clear that the evidence is mixed, it is difficult to imagine that RP could lead to resistance to stress. I would rephrase the following sentences:

*This question is critical to address because it would suggest that memory traces could be made resistant to what would otherwise be detrimental effects on memory via non-invasive learning strategies. This investigation would challenge some of the major theories of stress, as it would suggest that there is a way to make memory resistant against what would be a stress-induced memory impairment*

As something more nuanced in terms of protecting factor or making the memory less sensitive.

*And, stressors experienced at retrieval decrease memory and learning. RP has been consistently shown to boost memory and learning*

I think the coma after the and can be deleted (the and too).

Page 10 Stress main effect

Why isn’t it « Retrieval practice main effect”? I am not sure about the clarity of the hypothesis. It is said that the effect should be seen with stress induction… even in a group that did not undergo stress? I went back to the introduction to see if I missed something but I am sorry, I don’t understand.

I think this hypothesis lacks the point. In my opinion, you tried to get two hypotheses inside one.

The first hypothesis is that stress has an effect on memory performance.

The second is that when we compared strategies to lower the effect of stress, the best one is retrieval practice.

Both are main hypotheses, but the second is the key of this meta-analysis.

Now let’s go deeper in the operationalization because it is tricky (you are doing two meta-analyses with the same dataset). You will mostly have studies 2(stress or not) x 2 (retrieval practice versus other practice).

Your datasheet will look like this

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Reference | Experiment | Es | H1 | Practice\_to\_Compare | M\_retrival | M\_PtC | M\_stress | M\_nostress |
| Study 1 | 1 | 1 | No | Reread | 2.01 | 2.03 |  |  |
| Study 1 | 1 | 2 | No | Nothing | 2.40 | 2.50 |  |  |
| Study 1 | 1 | 3 | Yes |  |  |  | 4.10 | 2.78 |

Reference is the name of the study (Mihaylova, 2023). The experiment is the experiment inside the study (1, 2, or 3) for the two-level MA. ES is the Group inside the experiment (the number of ES per experiment, for the 3-level MA). Practice to compare is the other practice that will be useful for the moderator hypothesis. The last columns are the means (and sd and n for sample size etc, etc.) The key point is the H1 column which would be No for Hypothesis 2 and Yes for Hypothesis 1. “No” lines are comparing RP to control only for the stressed groups. “yes” lines are comparing the stress group and the no stress group no matter the practice groups. For the analysis I would only filter this column to do the first or the second meta-analysis.

Another possibility is to have two datasets (two Excel sheets). I hope it helped.

Page 13: *We saved all final studies included in the total search into the Included Articles folder accessible on OSF*

I recognize it coming from our template. We thought of it as a good idea, but we don’t have the right to share the studies from other researchers on OSF. I recommend deleting this part.

P12 Psycnet is not a database, Psycinfo is (without an h, <https://www.apa.org/pubs/databases/psycinfo/>) I highly recommend adding a database with thesis such as proquest, because I suppose many papers are graduated or Ph.D. thesis.

While looking at Table 1 (p 15), I found that you planned to use “thesis common”. Can you explain better before that? (Especially given my previous commentary on proquest).

On another side, I recommend putting this table in supplementary material, since you already talked about the info inside in the text.

Given that this table will be supplementary, I would like to have the table of all articles included in the main manuscript (see the end of page 14). It is good to have a nice overview of the characteristics of the studies in the main manuscript. I view it as the “descriptives” of the meta-analysis.

* Actually, you already have this table in Table 5. Why then have you said you put it in supplementary?

The screening section is long and redundant in my view. It also lacks a mandatory input, the PRISMA flow diagram, while it could also be in the literature search section. In text, one need to explain how they perform the procedure of the Prisma figure, succinctly. I suggest a complete rework of this section with the addition of the figure.

In the same vein, the included studies coding is too long. You can just say that it was made by two authors who resolved conflict by talking together and reporting reasons for exclusion in the excluded studies tab, while check the interrater agreement with cohen’s kappa. The last paragraph seems not useful since the datasheet is available and results will be reported from it. I suggest deleting it.

Regarding the variables and design in the studies section (page 21).

First, it is said “When possible, we checked whether assumptions of normality and linearity were fulfilled in included studies and whether the data was transformed or not, as this can affect the effect size (Meteyard & Davies, 2020).” What does it imply, will you exclude the ES when the assumptions are not met? will you just report that they are not? It seems that this part lacks a conclusion about what you will do with this information.

On the overall, I also suggest deleting this section, because I have not learned anything new (beside what I just mentioned) that was not in the inclusion/exclusion section. Especially, you talk a lot about the design, and since you already mentioned that you will exclude non-experimental 2(RP)x2(stress) designs, this section is not that useful. Regarding the last part:

*In case articles conducted several experiments or if several effect sizes existed per article, these were entered as separate effect sizes and not combined in order to prevent unit-of-analysis error (Harrer et al., 2021)*

It would be better to say that you will perform a three-level meta-analysis to account for dependencies within the studies as mentioned in Harrer book and in our template. I would have written that in a section called “analysis”.

* Oh, actually you wrote that just after page 23. I then suggest deleting this part as well.

Again when you say « Whenever available, we collected statistical information (i.e., standardized effect sizes) from the original papers or directly from authors […] tab to facilitate reproducibility.”

This was already said in the inclusion/exclusion part and is redundant.

P.24 *« Moderate to high heterogeneity in our sample is indexed by an I2 of over 50% and a significant Q statistic (Harrer et al., 2021)”*

Heterogeneity is calculated by the Tau² statistics. I² represents the percentage of the total variation in a set of studies that is due to heterogeneity, The Q statistics indicated how surprising it is if accounted by sampling error. I would rephrase it a bit to say that you determined a threshold of 50% of I² as an indicator to perform subsequent moderator analysis. (Ok it’s a detail but it’s better not to say that i² is an indicator of heterogeneity).

Page 25 : *Publication bias was assessed with various measures in order to obtain a range of values.* I don’t understand this sentence. Then you used two times the word « illustrated »

Page 26 I would recommend deleting the “we did not employ p-curve…” because, well you don’t have to mention everything you don’t want to do. Maybe consider adding it in the supplementary materials instead (not the p-curve, the paragraph).

More Broadly, I have questions/concerns regarding some aspects of the MA that needs more reflexion.

One is related to the operationalization of stress. I do get that two tasks are used to stress the participant (while I don’t know if it is not possible to use other tasks), is there a possibility to control/measure/take into account the stress felt by one participant? In terms of magnitude? Right now, it seems that researchers in this area said “we used this task to stress the participants so we assume they are all stressed”. But we don’t know if it affected everyone (with the use of a check for example) and how one particular individual will cope with this stress. Also, in the studies explained in the introduction, only the “recall performance” was measured. It seems a bit weak to me. Also, in H2 you only focus on TA and TSST. Are you sure they are the only types of tasks? (I don’t know I am not an expert in the field). I think we need to have a basis on how stress is operationalized before thinking that something can protect people from it.

In the same view, I have difficulties to understand how it is possible to compare the strength of Retrieval Practice when, for some, the performance was measured directly after the task, and for others as long as 1 week after. It is a bit the problem of apples and oranges, where we aggregate measures in a MA that are not equivalent. This could be resolved by using the length of retention as a moderator.

Finally, I think that the “type of task” is important in term of generalizability. As you mentioned in the Intro with the “context dependency” part, I think it is relevant to test of strong is the effect depending on the context – especially because we can assume the experiment was mostly, if not only, conducted with students in psychology. Even in this situation, Learning materials for a course, with the stress of a grade, is different than learning a scenario or an imaginary passage of a book. It also think of it as something to be controlled via a moderator analysis.

As I mentioned in the beginning of this review, I think this MA has potential and I would be happy to know about the results, but it needs clarification and rework before having the possibility to conduct it. If you need any help regarding the template (especially how we created the datasheet and Rmarkdown, so that it can be better tailored to your need), don’t hesitate to send me an email at [fillon.adrien@ucy.ac.cy](mailto:fillon.adrien@ucy.ac.cy).

Best Regards,

Adrien Fillon