Dear Veli-Matti Karhulahti and reviewers,

Thank you for the feedback regarding our submission of the manuscript "*Learning from comics versus non-comics material in education: Systematic review and meta-analysis*" to the PCI RR. We believe that your thoughtful comments and constructive suggestions have strongly benefited the quality of the manuscript.

We are excited to inform you that we have diligently incorporated your feedback, making the necessary changes to enhance the overall quality of our manuscript. Your recommendations have played a pivotal role in refining this updated version, and we truly appreciate your commitment towards this end. We have listed a point-by-point overview of the changes below.

Thank you for your invaluable input and the positive impact it has had on our manuscript.

Yours sincerely,
The authors.

Dear Marianna Pagkratidou and co-authors,

Thank you for submitting your highly interesting Stage 1 manuscript to PCI RR. We have been very lucky to receive no less than four helpful reviews, from experts of education, meta-analyses, and comics. All reviewers are generally positive about the research plan and support inviting a revision. Because the reviews are extensive, I will minimize my own comments and merely recap the most significant points that should be focused on in the revision.

1. The reviewers consistently point at the problematic comparison between "comics and non-comics". I agree with them and encourage you to follow any effective solution of your preference, perhaps one of those kindly suggested by the reviewers.
Response: We have considered that, and we thank you for the useful suggestion. As you will see in the updated manuscript, we have narrowed down the comparison between comics and texts. Please see on page 4 the following:

" However, there are inconsistent findings regarding the effectiveness of learning when using comics compared to non-comcis material, namely texts. In this study, we will conduct a systematic review on using comics in education, as well as a meta-analysis to quantify the overall effect of empirical studies that used comics versus texts."

2. The reviewers also voice an issue of testing hypotheses without an underlying theory or other explanation that would justify the hypotheses (a hypothesis without reasoning is just a guess!) I tend to agree and, as the reviews suggest, encourage you to either formulate a theoretical, empirical, or other basis for testing the chosen hypotheses or transforming the

plan toward a more exploratory direction. If you choose to keep the hypotheses for testing, please move the supplement table (from OSF) to the end of the revised manuscript text file.
Response: As you will see below and in Table 1 at page 9, we have taken into consideration the reviewers suggestions and now we have made it clear that our study is exploratory.

3. As the reviewers note, PCI RR generally discourages using rules of thumb effect sizes (like those by Cohen) and instead justifying the range of meaningful and meaningless effects. A good additional source for this topic suggested by the PCI RR guidelines can be found here (Dienes 2021): https://doi.org/10.1525/collabra.28202
Response: We agree with the recommendation of the reviewer that the interpretation of the effect size should not be arbitrary. We have now defined a minimum effect size of interest, based on the results from a previous, similar meta-analysis, as shown below:

"For our meta-analysis, we consider an effect size d = 0.4 as the minimum effect of interest. This is derived by a previous meta-analysis that investigated the effectiveness of comics in education (Topkaya et al, 2023). In the study by Topkaya et al. (2023), a meta-analysis with subject area as a moderator variable (similar to our STEM and non-STEM categorisation) resulted in an overall effect size of g = 0.50, 95% CI [.33, .68]. Using the lower bound of the confidence intervals as a heuristic, and considering the Hedge's g correction, we define d = 0.4 as the minimum effect of interest to guide our analyses and interpretation."

4. Finally, I echo the reviewers' concerns about including any languages. It feels challenging to build a robust systematic plan that can ensure access to any language (both in terms of locating and reviewing relevant studies). A good solution could be to either limit the languages to those for which the authors and their collaborators have direct fluency, and/or review other languages in a separate exploratory section.
Response: We acknowledge that, and thank you for bringing that to our attention. We do sympathize with the comment that jargon might be an issue, but we still think that this is enough to exclude studies in the field; especially now that we can use GenAI tools. Of course, we will acknowledge and address that in the limitations. For now, we have updated our manuscript on page 13 with this:

"Of note, publication language will not be taken into consideration as we will not only include reports written in the languages spoken by the research team, but also we will include papers in any language that can be translated through Google Translate Documents (https://translate.google.com/?sl=en&tl=el&op=docs) and ChatGPT4 (OpenAI, 2023; https://openai.com/gpt-4) large language model, using the prompt 'Translate the following text to English:' followed by the section text in quotation marks."

I hope the reviews, as summarized by these notes, help you to make the plan even stronger. Please kindly include point by point responses to all the review comments in the revision.

I want to be clear that the idea of RRs is not to block any preferred research goals or force researchers to directions that they are not interested. Accordingly, if you consider some of the feedback not justified or that revising based on the feedback changes the plan too much from your goals, you are free to rebut any comment with a counter argument. If needed, you can also contact me directly during the revision process and we can together negotiate solutions for any part of the feedback in more detail. Our goal is simply to collectively make this the best possible study on your chosen topic.

All the best wishes and much looking forward to the next version,
Veli-Matti Karhulahti====

Reviews

1. ***Reviewed by Adrien Fillon, 31 Oct 2023 08:32***

This is a stage 1 report for a meta-analysis about how comics can increase learning compared to the same material in a non-comics format.
I think this manuscript is in good shape and I only have few minor points. Since there are no number of pages, I will refer to the chapter or the exact sentence.
Response: We thank Dr Fillon for his kind words. Let us note that we have added page numbers to the manuscript.

In study search, I think that Web of Knowledge does not exist, only web of science by clarivate. A quick google search led me to think that web of knowledge is the old name for web of science. Anyway, please update it.
Response: We thank the reviewer for bringing this to our attention. Web of Science was changed to Web of Knowledge, but has now reverted back to Web of Science. We have now adjusted the manuscript accordingly replacing 'Web of Knowledge' with 'Web of Science' on p.11.

10. Achievement level:
Can you explain a bit what this variable means? I don't understand what are these levels.
Response: We have made the following changes in the manuscript on page 15 to clarify:"Achievement level: We will extract information about whether the level of knowledge-achievement in the topic of interest has been taken into consideration in the studies by using a yes (=1) / no (=2) coding. In addition, we will extract the level of knowledge-achievement as specified by the studies, by coding the level of the participants as low (=1), medium (=2), and high (=3)."

For the moderators, I think that you can also code the "duration" of the experience of learning. Did the teacher presented a course with the comics for l h or the whole semester etc... I am always doubtful of all the effects found through 1 experience, as it can be confounded with a "surprise" effect. However, if we find an effect for a 1 semester course, the participants had time to accommodate to the material, reinforcing the effect.
Response: That is a great idea, we have updated the manuscript on page 15 with the following:

"*Duration of the intervention*: We will document the duration of the intervention in days, to examine the duration of the learning effect during the intervention period."

"the symmetry of the forest plot to investigate the presence of small study bias" seems to be written in grey?
It seems that "use Cochrane [1]s risk of bias tool to" is also in grey
Response: Dear Dr. Fillon, thank you for the comment.

While the use of ROB2 is an important idea to assess biases, it would be good to also add other tools such as the 3PSM which performed well in a set of simulation (carter et al. 2019). Adding PET-PEESE and a z-curve (or p-uniform) analysis could also benefit this MA by providing more details on the possibility of publication bias. One can contact me if you need the R code to conduct these analyses.
Response: We thank the reviewer for recommending the addition of multiple tools to assess bias in our MA. We also appreciate the availability for providing the R code. We have now added (p. 16-17) that we will also employ *p*-value drapery plots, as well as *p*-curve binomial testing, as additional tools for measuring bias, as follows:

"Further, we will employ multiple approaches to investigate potential heterogeneity and small study bias considering that it is recommended to employ various methods to investigate bias for meta-analyses in psychology (Carter et al., 2019). Specifically, we will use *p*-value drapery plots (Rücker & Schwarzer, 2021) and funnel plots to visually investigate small study bias. In addition, *p*-curves will be generated (Simonsohn et al., 2020) and tested for skewness and flatness using a χ^2 Binomial test.

The symmetry of the effect sizes will be examined by generating funnel plots to visually investigate the symmetry of the forest plot to investigate the presence of small study bias and we will also statistically test it using the Egger's regression test (Egger et al., 1997)."

As regards to 3PSM and PET-PEESE, we choose not to register such methods for our Stage 1 report. This is because the successful utilization and interpretation of both approaches is heavily reliant on multiple parameters such as τ, *SMD*, and *k.* The dependency of the success of these tools on multiple parameters leads to a high-risk for no model convergence, overfitting, and/or overcorrection (see Stanley, Soc Psych and Pers Sci, 2020 in addition to Carter et al., 2019). However, we would be happy to conduct these analyses as exploratory in Stage 2, if assumptions are met and the results from these analyses would be meaningful.

Finally, I have not found anything about open script, data and supplementary. In the PCI-RR guidelines, it is stated that "In general, authors are required to make all study data, digital materials, and computer code publicly available (at Stage 2 submission) to the maximum extent permissible by relevant legal or ethical restrictions." Therefore, it would be important to at least state that all codes, data and materials will be shared on an (empty for the moment) OSF repository, in the abstract and/or in the method section.
Response: Thank you so much, we have made the changes in the method section on page 10 with the following:

". All codes, data and materials will be shared on this OSF repository: https://osf.io/cmqb6/?view_only=98677e4fcab84e47968556c7958817f3"

I look forward to the revision and conduct of this meta-analysis as I find it very interesting to understand better the relationship between entertainment and learning.

Response: Dear Dr. Fillon, we appreciate your constructive feedback a lot.Adrien Fillon

2. ***Reviewed by Benjamin Brummernhenrich, 23 Nov 2023 16:36***

The authors describe a planned review and meta-analysis that is concerned with the effect of using comics as educational materials on knowledge gains. I think the topic is relevant for the field and would find a systematic analysis of the effectiveness of comics worthwhile. However, although the Stage 1 Registered Report makes the procedure reasonably clear (with some exceptions which I will detail below), I am unsure whether the question, as the report currently poses it, is a well-formulated and reasonable one. I will first state the main problems that I see with the authors' reasoning and approach.

There are three main problems regarding the derivation of the research question (or the effects to be meta-analysed), that make me question its reasonableness:

1. What media comparisons are sensible?
Response: We have taken into consideration all the comments, and we have decided to focus on the comics vs texts comparison, and document all the other non-comics comparisons exploratory. Please see the updated *introduction* section, on page 15 where we added the following:

"Text type (control condition): We will document information regarding the type of the text; and for any non-comic educational material used, we will document the information by using the following coding system: text (=1), photo (=2), animation video (=3), etc."

2. What is a comic and in what way exactly does it differ from other (visual) media?
Response: We have updated our definition in the *introduction* section as follows:

"Comics can be defined as a particular type of social object, used by people of a particular cultural orientation, which use visual language (sequential images) and writing, typically associated with contexts and styles (Chute, 2008; Cohn, 2012; Cohn & Magliano, 2019). However, this definition is complicated by the term "comics" in education being conflated by how people use the term in the first place (Cohn, 2013a; Gavaler, 2022). In one sense, "comics in education" is meant as the promotion of the structural properties of the "medium" (i.e., sequential text/image units). Such an advantage would manifest in educational materials

being created in this manner. For example, this sense may include cases where textbooks or educational materials are created using multimodal text-image units put into a sequence. Another sense is that comics as existing social objects are beneficial in education. In these cases, the published works that carry the designation of "comics" are used in educational contexts. This would include using published comics to teach about literature or to teach second languages, or using Art Speigelman's memoir Maus about the holocaust as a way to teach history. This does not necessarily use the structural properties of visual and multimodal expression as a means to educate (such as in comparison to standard textbook formats), but rather uses published literature that may otherwise be viewed as entertainment within educational contexts. While advocacy for "comics in education" often conflates these senses both in theory and practice, these distinctions are important for disentangling their purported advantages."

3. What makes STEM subjects different and how does this impact learning with media?
Response: What is different in STEM subjects is the large association between spatial ability and the ability to form accurate mental representations of many STEM topics/concepts/processes. The ability to visualize and mentally manipulate images and representations is very important in STEM learning, and more important in some cases than verbal or mathematical abilities. Many students struggle with this type of thinking and can be greatly assisted by having external visualizations provided to them. Comics can provide such visualizations in the form of static images that can be brought to life or greater meaning through a comic character. They can, therefore, scaffold the learning of STEM topics, particularly for those with low levels of spatial ability. We have addressed that issue in the manuscript too, see *introduction* section, with the following:

"s. We specifically aim to examine potential differences between STEM and non-STEM contexts, as panels in comics, by providing external visualizations to students, might scaffold learning differently for STEM and non-STEM topics, especially for individuals with low spatial abilities, considering the relationship between visualization, spatial ability efficacy, and STEM learning (Newcombe, 2013; 2017; Zhu et al., 2023)."

There is a long history of media comparison research in educational psychology, but it is also a very rocky one. A prominent example is the Clark-Kozma debate on learning with digital media, that revolved around the question whether the use of (digital) media per se had specific impacts on learning or whether any effects were exclusively a consequence of teachers structuring the instruction in a specific manner around the medium. I think the same question has to be asked here: Is there a specific effect of comics as a medium or does this depend too strongly on how they are used, in which context and to what end? Even if the authors expect a specific effect, I found that the introduction did not make it very clear how this effect comes about.
Response: We have taken into consideration all the comments and we have changed our Research Questions and hypotheses. Please see *Table 1* the following:

"1. What are the claimed benefits of comics vs text for education?

We refrain from forming concrete hypotheses, as our analysis will be exploratory in nature.

2. Is there a difference in the putative effectiveness of comics in STEM vs non-STEM subjects?
We refrain from forming concrete hypotheses, as our analysis will be exploratory in nature.

3. Is there a moderating effect in the putative relationship between comics and learning of factors such as age, target population, experimental design, intervention type and alternative non-educational material such as videos etc?
We refrain from forming concrete hypotheses, as our analysis will be exploratory in nature."

Does it have to do with the sequentiality of comics? With the combination of text and images? With the fact that comics will be perceived as entertaining by students? This leads to the second problem: What exactly is the comparison point?
Response: We have updated our definition in the *introduction* section with the following:

"Comics can be defined as a particular type of social object, used by people of a particular cultural orientation, which use visual language (sequential images) and writing, typically associated with contexts and styles (Chute, 2008; Cohn, 2012; Cohn & Magliano, 2019). However, this definition is complicated by the term "comics" in education being conflated by how people use the term in the first place (Cohn, 2013a; Gavaler, 2022). In one sense, "comics in education" is meant as the promotion of the structural properties of the "medium" (i.e., sequential text/image units). Such an advantage would manifest in educational materials being created in this manner. For example, this sense may include cases where textbooks or educational materials are created using multimodal text-image units put into a sequence. Another sense is that comics as existing social objects are beneficial in education. In these cases, the published works that carry the designation of "comics" are used in educational contexts. This would include using published comics to teach about literature or to teach second languages, or using Art Speigelman's memoir Maus about the holocaust as a way to teach history. This does not necessarily use the structural properties of visual and multimodal expression as a means to educate (such as in comparison to standard textbook formats), but rather uses published literature that may otherwise be viewed as entertainment within educational contexts. While advocacy for "comics in education" often conflates these senses both in theory and practice, these distinctions are important for disentangling their purported advantages.
Indeed, although often associated with entertainment, the comics "medium" has been shown to be valuable and beneficial for conducting and communicating science and they have been further used for educational purposes for over 80 years by a range of other scholars (Farinella, 2018; McCloud, 1993; Topkaya et al., 2023; Yang, 2008)."

However, we would like to answer that comment too as all of these things have been claimed by people as providing an educational advantage, either within comics specifically or with comics as an aggregation of those various traits (i.e., sequence, multimodality, entertainment). All of these have been claimed as advantageous, and thus that comics are

advantageous. Our study is exploratory and it is not *our* claim that comics are beneficial for comics, but rather to interrogate the claims by a range of other scholars that comics are beneficial for education. Our study aims to see whether there is evidence of such a claimed advantage.

The authors define comics as "as a particular type of social object, used by people of a particular cultural orientation, which use visual language (sequential images) and writing, typically associated with contexts and styles" (p.5). Some of these concepts are not explored further (what cultural orientation? what kind of contexts and styles?), so I have to assume they are not important here. If it is the sequentiality and the combination of images and writing, I wonder what it is about these that should make comics superior to other media.
Response: We indeed did not differentiate between particular styles or cultural orientations as we average across such distinctions (if they are even made in studies at all). The reviewer is correct that this often makes sequencing and multimodality the primary structural traits that are compared, although really the definition of "comics" is "whatever people call comics." So, there is potential for structural overlap between what people call comics and what people call other things (like diagrams). We differentiated within our comparisons their properties in the *inclusion and exclusion criteria* section. Also, we have addressed this concern on page 4-5 with the following:

"Comics can be defined as a particular type of social object, used by people of a particular cultural orientation, which use visual language (sequential images) and writing, typically associated with contexts and styles (Chute, 2008; Cohn, 2012; Cohn & Magliano, 2019). However, this definition is complicated by the term "comics" in education being conflated by how people use the term in the first place (Cohn, 2013a; Gavaler, 2022). In one sense, "comics in education" is meant as the promotion of the structural properties of the "medium" (i.e., sequential text/image units). Such an advantage would manifest in educational materials being created in this manner. For example, this sense may include cases where textbooks or educational materials are created using multimodal text-image units put into a sequence. Another sense is that comics as existing social objects are beneficial in education. In these cases, the published works that carry the designation of "comics" are used in educational contexts. This would include using published comics to teach about literature or to teach second languages, or using Art Speigelman's memoir Maus about the holocaust as a way to teach history. This does not necessarily use the structural properties of visual and multimodal expression as a means to educate (such as in comparison to standard textbook formats), but rather uses published literature that may otherwise be viewed as entertainment within educational contexts. While advocacy for "comics in education" often conflates these senses both in theory and practice, these distinctions are important for disentangling their purported advantages.
"

The problem here is that the authors plan to compare comics to "non-comics", i.e. basically all other kinds of media - or at least visual media - that are not comics! That in itself seems like a very asymmetric comparison. The authors state that "media comparisons also differ

between text [...], animation [...], or video." (p. 7) This not only ignores the heterogeneity of these media types in themselves but also the host of other visual media. Images (moving or non-moving) can be (such as a diagram) or analog/depicting (such as a photograph). Images and text can be combined, which is a whole area of research (e.g. Schnotz's "integrated model of text and picture comprehension", 2005), of which comics would arguably be a special case. The problem that I see is that whether comics are more or less effective depends not only on the context and content (see problem 1.) but also on which medium specifically they will be compared to, because in what way comics differ from another medium determines whether an effect obtains or not. Comics and videos are both sequential, but comics are combined with text and videos (typically!) are not. If you compare comics with infographics the opposite is true!
Response: We have acknowledged the concerns about this broad comparison, and in the updated version we have narrowed it down to comics vs texts; while documenting all the other non-comics material for exploratory purposes. Specifically, on page 8 we have added the following:

"Given the conflicting results, the main question driving the current research is whether comics, compared to non-comics education material, namely text material, are an effective learning tool. To our knowledge, no past research has integrated the range of empirical studies that conducted interventions to improve learning using comics versus non-comics materials. The aim of the present study is to systematically review and to quantify using meta-analysis the overall effect of comics vs text material that have been used in empirical studies that targeted learning for STEM and non-STEM fields."

In my opinion, the authors need to offer a clear account what characteristic of comics it is that provides the benefit, and in what situation. That precludes, in my view, a broad comparison of comics vs. non-comics because the characteristics that differ will vary. In some cases e.g. sequentiality may provide a benefit, in some cases it may not, or may even be detrimental. I am not convinced that the moderator variables that the authors consider adequately capture these factors. In this way, I think the authors' first hypothesis is not well argued.
Response: Now the study is exploratory. Also, we accounted for the characteristics of comics based on the cluster of features typically associated with comics (sequence, multimodality, etc.). We would like to note here that it's not us who are defining comics, but it's the studies that are reviewed in our meta-analysis. Specifically, if the authors of the studies call the education material as comics (or the associated terms), we count them as comics. Also, we will like to note here that the comparison, after reflecting on reviewers' feedback, is on comics vs texts.

The third point is related to this one, and that is the STEM/non-STEM comparison. Here, again, it is unclear which characteristics of STEM subjects make comics especially apt for these contexts: The authors state that "the background information provided in the panels might operate as scaffolding for more effective learning about STEM than non-STEM concepts" (p. 9) - in what way are STEM concepts different such that comics are a superior

medium than others to learn them? Again, I think the report lacks a coherent theoretical reasoning from which to derive this hypothesis.
Response: We have addressed that comment by adding a new paragraph in the manuscript; and we have made it clear that our study is exploratory. The changes, in page 8, are the followings:

" We specifically aim to examine potential differences between STEM and non-STEM contexts, as panels in comics, by providing external visualizations to students, might scaffold learning differently for STEM and non-STEM topics, especially for individuals with low spatial abilities, considering the relationship between visualization, spatial ability efficacy, and STEM learning (Newcombe, 2013; 2017; Zhu et al., 2023)."

I realise that I'm basically asking "What is the process?", a kind of question that is receiving some push-back recently. I acknowledge that investigating effects can be enlightening without a strong theoretical reasoning, but this is, in my reading, not what the authors are setting out to do, especially since they are formulating hypotheses around the specific comparisons. Thinking about the title it occurred to me that the problem may be that the authors want to make this a systematic review as well as a meta-analysis. Although the review part is not elaborated upon very strongly, it suggests to me that the authors are looking at both theoretical as well as empirical aspects of the effects of comics.
Response: We have acknowledged this concern and we have made our research questions exploratory, please see Table 1. We are interested in finding out from empirical studies if the so-called comics (as defined by researchers) are indeed more effective in learning acquisition than texts (or any other non-comic material - exploratory) or not. We hope that this clarifies the purpose of our study.

In summary, I personally do not think it would be worthwhile to go through with the plan as the authors have formulated it. The reasoning why comics should be superior to all "non-comics" media does not seem sound to me, neither why comics should be more effective in STEM than in non-STEM comics. However, as I said above, I think the topic in itself is worthwhile of consideration. In the sense of constructive criticism, I personally see two ways for the authors to proceed in order to make this a worthwhile line of inquiry:
The first would be to elaborate more specifically on the characteristics of comics that make them special, that distinguish them from some specific other type(s) of media. Think about which of these characteristics should improve learning gains in specific situations (i.e. regarding a certain type of content, when used in a certain kind of way, etc.), what kinds of affordances do they bring to the learning situation. Then focus on this (set of) affordance(s) and review/meta-analyse studies that pertain to it. Let us assume (just for the sake of argument, I am not an expert in this specifically) that the sequential nature of comics should be especially beneficial for beneficial for learning procedural knowledge. Then studies could be sought out that compare comics with text-image combinations that are not sequential. The type of knowledge learned could be a moderator here, and effects should be stronger for procedural knowledge than, say, conceptual knowledge.
Response: "Comics" are defined as such both by their internal properties (sequencing,

multimodality) and by being assigned to that social category. There are certainly structural overlaps between diagrams and comics, but they have different social contexts and often use different visual languages. The key is that we are making sure that what we compare within the studies is adequately described as "comics" in one grouping (as defined by the authors of the studies we review) and texts in another (again, as defined by the authors of the papers we review). We think the reviewer is right to raise this issue, but it's not one that's well grappled with by most of the people who make these claims. As far as we would understand it, most comparisons are between "comics" and "textbook style materials". We would like to clarify at this point that there are a cluster of these traits that people say are beneficial, and this review provides the first attempt at integrating results across a growing body of literature; and we will work with what we have for now and make suggestions for future research of the nature the reviewer suggests

However this may be premature if the theoretical ideas around learning from comics are not yet specific and precise enough. In this case I would suggest - as my second proposal - to defer the meta-analysis to a later point in time and focus on the systematic review in order to tease out those characteristics. The strategies that the authors describe for literature search and categorisation would still be useful for this. The authors could more strongly focus on how comics are actually implemented in learning situations, with what kinds of media they are contrasted etc.
Response: The focus of our study is to see what they say are the purported benefits of comics for education. Mentioning these claims from the literature directly would be helpful.
For that reason, we are interested in this question: What are the claimed benefits of comics vs text for education? Also, now the study is exploratory, so getting quantitative results will help us drive the discussion. E.g., if no effects are found at all, then maybe we have no reason to tease out the characteristics. However, we expect to find an adequate number of studies, so not including a meta-analysis will be a waste of resources/time.

I hope the authors take these comments in the manner they were intended: As constructive criticism and in the hope that they are helpful for revising their approach, if they wish to do so.

In the following I will go through the text of the report in order, to point out additional smaller things as they occurred to me:

Introduction
I assume ASD refers to "autism spectrum disorder" but I am not entirely sure and I think the authors should write it out.
Response: Done, thank you.

The researchers seem to be active in the field - they should be explicit about how they can ascertain that their own studies will not be given preferential treatment or weight in their literature search and the following categorisation and analysis.

Response: Indeed, we are doing research in the field of comics cognition and memory, but we will follow the PRISMA protocol and we will not give preferential treatment to relevant publications. .

In Table 1, the authors first formulate directional hypotheses ("We expect comics to have a greater impact..."). However in the "Interpretation" column, they also allow directional interpretations in opposite directions. So I am unsure if there are competing, directional hypotheses (which would be fine if both were argued for!) or if the authors allow themselves to support any kind of hypothesis.
Response: We have updated *Table 1*.

Method
- "We will sequentially screen titles, abstracts, and full-texts." (p. 11) I am unsure what this entails and what exactly the criteria for inclusion are. The goal of the registered report is to enable replication and guard against procedural flexibility and I think this is one point where subjectivity may come in.
Response: As detailed in the methods section we have presented the inclusion and exclusion criteria after taking into consideration all the comments from the four reviewers while trying to be as clear and transparent as possible to ensure easy replication of this study following the PRISMA protocol. In. the method section, you will find all the changes that we did, but if something is still unclear please let us know to improve further.

Will the Zotero database be published?
I am unfamiliar with Rayyan and would need more information about how it works. One relevant question regarding reproducibility would be how determinate its output is.
Response: Zotero and Covidence are online platforms providing tools that can help in the screening of big datasets, please check here https://www.zotero.org/ for Zotero and here https://www.covidence.org/ for Covidence. Note that we decided to use Covidence because one of the authors will have full access to this platform and we could exploit at 100% it's potential since there are studies indicating that Covidence is a very effective platform for such purposes. Please note that all materials will be made available, including zotero bib files, and csv files from Covidence where reviews stay archived in their website. On page 11 you will find the following change:

"We will use the open-source online reference management software-platforms that will help us in the screening of big datasets. Specifically, we will use.."

Regarding inclusion criteria:
I was unsure what the authors meant by "general population samples" (p. 11). Are there any samples that this excludes? If not, then I think that this is not really a criterion. It could be a moderator variable though.
Response: For the purposes of our study it is considered a criterion because we are only interested in the general population. We are going to exclude studies conducted, for example, in populations with clinical diagnoses, such as learning disorders.

I was unsure what exactly the authors consider to be "sufficient data" (p. 11). What does it mean for data to be "usable for the analysis"? What kind of statistical values would need to be reported?
Response: We have now added on page 12 the following information to clarify what we consider 'sufficient data' in terms of our inclusion criteria:

"We will use data that are either reported in the studies in a usable way for the analysis or provided by the authors after request. To be usable the data should enable us to calculate an effect size (Cohen's *d* or Cohen's drm) that can be used to pool an overall effect size in a random effects meta-analysis model (e.g., mean, sd/SE, n; see Statistical Analyses). If no data are provided by the studies or the authors and if possible, we will extract data from figures with the use of WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/)."

As I mentioned above, the authors would need to be explicit about which outcome measures would be eligible and which would be excluded. What kind of learning will be targeted? Will studies be excluded that report motivation measures? Small-group interaction? Procedural knowledge? etc.
Response: We will address this issue by using any outcome that is close to a knowledge measurement, as the focus of our study is not the type of knowledge. However, this is a clear limitation of our study that we will make sure to discuss in the discussion. If a study includes only motivation measures, and there is no relevant knowledge measurement, then yes that study will be excluded. We have clarified that now on page 13 in the manuscript:

"6. Outcome variable: We will include only studies that have an outcome variable for any kind of "knowledge" measurement."

Re: publication type/research design: What about studies with more control groups or different kinds of comics? What will suffice as a control group? What if the second group is not labelled as a control group?
Response: We are interested in studies that use comics as one group and text as another group. We will code a control group as a control group if it serves the purpose of a control group, whether it is being labeled or not. Also, in the case a study results in more than one effect size, we will explore how this might bias our results by employing a multi-model meta-analysis, by including a third-level representing the study. Please see that change on page 15 in a footnote at the manuscript.

I am very doubtful about including studies in languages other than those that the authors are proficient in. Although Google Translate has come a long way, I find that especially scientific jargon can be a gamble.
Response: We acknowledge that, and thank you for bringing that to our attention. We do sympathize with the comment that jargon might be an issue, but we still think that this is

enough to exclude studies in the field; especially now that we can use GenAI tools. Of course, we will acknowledge and address that in the limitations. For now, we have updated our manuscript on page 13 with this: Of note, publication language will not be taken into consideration as we will not only include reports written in the languages spoken by the research team, but also we will include papers in any language that can be translated through Google Translate Documents (https://translate.google.com/?sl=en&tl=el&op=docs) and ChatGPT4 (OpenAI, 2023; https://openai.com/gpt-4) large language model, using the prompt 'Translate the following text to English:' followed by the section text in quotation marks."

I was unsure what the authors meant by "We will extract the sex of the participants separately for each group (experimental vs. control)" How will this enter the analysis. The ratio of of male to female participants per experimental group? Whether the study tested male vs. female? Whether gender was a covariate?
Response: These data will be extracted this way for descriptive purposes, please see page 14 where we added the following:
"Sex of the participants: We will extract the sex of the participants separately for each group (experimental vs. control) as male, female, or other (if reported), for descriptive purposes."

The distinction of comics as complementary vs. main medium seems very vague to me. There should be more information what characterises each application.
Response: Please see pg 14 where we made the following changes:

"We will extract information regarding the type of intervention by using the coding system of comics as main teaching material (=1) - *referring to studies that used comics and non-comics as the main medium to educate the participants* - and comics as supplementary material (=2) - *referring to studies that used comics and non-comics as a complementary medium to the existing course or text material."*

In general, because some of these categorisations are at least partly matters of judgment, involving some uncertainty, I think only resolving inconsistencies by a third person is inadequate. My suggestions would be for the two coders to categorise a subset of the data, for which inter-rater reliabilities are then calculated, refining the system until agreeable consistency is achieved. This procedure and what is an acceptable level agreement (e.g. Krippendorff's Alpha or similar) should be made explicit in the pre-registration.
Response: We have addressed that, please see the *study selection* section on page 12 with the following:

"All materials will be made available, including zotero bib files, and csv files from Covidence where reviews stay archived in their website. Authors MP and PP will select the studies independently and author GD will resolve any inconsistencies (i) by taking the records to the full-text stage of the review in the titles and abstracts screening stages, even if only one author accepts them or is unsure and (ii) via discussion in the full-text stage. MPP author, specializing in meta-analyses, will resolve any cases of uncertainty."

I was unsure What exactly was meant by "level of knowledge-achievement". I thought that knowledge was the outcome. Maybe the authors mean prior knowledge and whether that was entered as a covariate.
Response: We have addressed that, please see the *data extraction* section on page 15 for the following change:

"We will extract information about whether the level of prior academic knowledge-achievement in the topic of interest has been taken into consideration in the studies by using a yes (=1) / no (=2) coding. In addition, we will extract the level of knowledge-achievement as specified by the studies, by coding the level of the participants as low (=1), medium (=2), and high (=3)."

The categorisation of the control conditions, as mentioned above, falls short, in my opinion. Media use in education is very diverse, other forms of images and visual media, with and without text, sequential or non-sequential, symbolic (such as diagrams) as well as analogue (such as depictions) are ubiquitous. To choose only photos and animations as contrasts offers a very restricted view of visual media in the educational context.

Response: We have taken into consideration all the comments, and we have decided to focus on the comics vs texts comparison, and document all the other non-comics comparisons exploratory. Please see the updated *introduction* section at page 4.

Statistical analysis
The outcome is stated as "knowledge". Both the concept of knowledge itself but also its measurement are very complex. In educational studies, several different kinds of knowledge are common as outcome measure. However, these are not equivalent. Knowledge can be declarative, conceptual, procedural etc. It can be measured by multiple choice tests, written or oral exams, behavioural measures, teacher ratings, peer ratings, self-report etc. The studies will surely differ widely in this regard and needs to be represented in the categorisation of the studies.
Response: This is true, and we expect to find inconsistencies. We will address this issue by using any outcome that is close to a knowledge measurement, as the focus of our study is not the type of knowledge. However, this is a great point to consider, and a clear limitation of our study that we will make sure to discuss in the discussion.

Apart from this, I found the statistical analysis section the most convincing. (However, I am not very familiar with meta-analytical methods so an expert on the topic may have more to say on this topic.)
Response: Thank you.

How do the authors justify choosing r = 0.5 for studies that do not provide a coefficient for calculating d for repeated measures?

Response: We have now revisited our approach and decided to approximate the correlation between repeated measures. On page 18, we now report this method and the respective equation in the updated report as also shown below:

"If a correlation coefficient is not provided by the primary studies to calculate d_{rm}, a correlation coefficient (r) will be approximated using the formula $r = (t^2 * (sd^2_{pre} + sd^2_{post}) - N * mean^2_{change}) / (2 * t^2 * sd_{pre} * sd_{post})$, where t is the corresponding t-value, sd_{pre} and sd_{post} is the standard deviation of the sample before and after the intervention, respectively, N is the sample size, and $mean_{change}$ is the difference in means before and after the intervention."

3. ***Reviewed by Solip Park, 03 Nov 2023 18:48***

Summary of this RR

Aim: The authors seek to explore whether comics affect learning differently.

Problem statement: Despite the increase in the usage of comics in classrooms daily either as independent reading or as a supplement to the main lesson, there are inconsistencies in findings regarding the effectiveness of learning when using comics compared to non-comics material.

Goal: To systematically review and quantify using meta-analysis the overall effect of comics vs non-comics material that have been used in empirical studies that targeted learning for STEM and non-STEM fields.

Research setting:

In RQ1: How authors will measure the "better learning"?
Response: By calculating the effect sizes of the knowledge measurement and by comparing that among the studies that used comics vs non-comics. We do not define better learning, but we will measure which medium is more effective for learning. Please see page 8 that we added the following footnote:

" Note that we will measure which medium is more effective for learning by calculating the effect sizes of the knowledge measurement and by comparing that among the studies that used comics vs texts."

In RQ2: How authors will measure the "putative effectiveness"?
Response: By examining whether there is in comics a greater impact on learning than non-comics (please see Table 1).

This is perhaps the main point to measure to extract 'Achievement level' from the data.

Method:

Perhaps a few examples of what the authors expect when saying "…empirical studies that compare comics with any non-comics material". Because the abstract says "e.g., text or video" and the data extraction chapter mentions the three main non-comics materials in criteria: text (=1), photo (=2), animation video (=3). But what about else? What about those that are 'somewhere' in between? What first comes to my mind is whether 'visual novels' and 'comical storybooks' could be regarded as 'comic' in this systematic review, as long the authors of the original paper have identified the object as "comic". And what about games (e.g., board games) – would they be regarded as text, photo, or animation video? To avoid these potential terminological confusions, perhaps the authors could consider listing some examples and rationale behind these choices.
Response: We have taken into consideration all the comments, and we have decided to focus on the comics vs texts comparison, and document all the other non-comics comparisons exploratory. Please see the updated *introduction* section on page 8, as follows:

"Given the conflicting results, the main question driving the current research is whether comics, compared to non-comics education material, namely text material, are an effective learning tool. To our knowledge, no past research has integrated the range of empirical studies that conducted interventions to improve learning using comics versus non-comics materials. The aim of the present study is to systematically review and to quantify using meta-analysis the overall effect of comics vs texts material that have been used in empirical studies that targeted learning for STEM and non-STEM fields."

Google Translate's effectiveness in some languages can be somewhat questionable. For example, Korean and Japanese (the language that I can speak) have multiple vocabularies of "comic" (e.g., comic, manga/manhwa, toon, webtoon, etc) with subtle differences in nuance and tone, which I wonder how accurately the Google translation AI can able to articulate. I quickly checked that my native language Korean, comics ("Manhwa") either translated as "comic" or "cartoon" into English Google Translation.
Response: We acknowledge that, and thank you for bringing that to our attention. We do sympathize with the comment that maybe some language specific differences may affect the results (i.e., Manga vs. Comic), but we still think that this is something necessary that can be achieved with the new GenAI tools that we have. Of course, we will acknowledge and address that in the limitations. We have now updated our manuscript on page 13 with this:

"Of note, publication language will not be taken into consideration as we will not only include reports written in the languages spoken by the research team, but also we will include papers in any language that can be translated through Google Translate Documents (https://translate.google.com/?sl=en&tl=el&op=docs) and ChatGPT4 (OpenAI, 2023; https://openai.com/gpt-4) large language model, using the prompt 'Translate the following text to English:' followed by the section text in quotation marks."

I consider it would also be beneficial for the readers whether the comic or non-comic materials used in this research will be the ones developed by the researcher team themselves or outsourced (or using or modifying existing materials) from external sources. It would also

be interesting to see how it corresponds to the learning outcomes. (Or is the "Experimental design" criteria already covering this aspect? I wasn't sure from the current RR.)
Response: We will extract information as presented in the existing papers in the literature but we acknowledge the reviewer's comment and we will also add extract the data and as a moderating variable whether the researchers have created their own comics or have used existing ones.We have now updated our manuscript on page 13 with this:

"Comics type (treatment condition): We will document information about whether the researchers have created their own comics (=1) or have used existing ones (=2)."

4. ***Reviewed by Pavol Kačmár, 27 Nov 2023 20:39***

Thank you for the opportunity to review the protocol entitled: Learning from comics versus noncomics material in education: Systematic review and meta-analysis. The study aims to provide a systematic review and quantification of the overall effect of non-comics vs. comics materials on learning and examine whether learning is affected differently in STEAM and non-STEM fields and by selected moderators. Below, I will provide a review of all sections of the protocol separately, while the general evaluation and recommendation will be provided at the end of the review.

In the introduction, the authors argue that despite the inclination for comic book materials in education and students, inconsistent findings regarding the effectiveness can be found and systematic review and quantification of the effect size of comics vs. non-comic material in STEM and non-STEM fields is needed. There are three research questions and two hypotheses provided. The introduction is written in an engaging style and logically well structured. I like the introductory example and the logical flow of the text, pointing out that there is a lack of information concerning effectiveness and that effectiveness could depend on many factors.
Response: Thank you!

The authors argue that the difference between comics and non-comics is mainly in visualisation, leading to richer examples and more engaging ways of presenting materials. However, as a reader, I pondered whether there is no further theoretical basis. If there is, I would appreciate it if the authors could further elaborate on why it is expected that comics are more effective (i.e., are there any theories that could be mentioned as an example and will be used later in the discussion for interpretation of positive findings)?
Response: We have updated table 1. We will examine the potential benefit as implied by the existing claims by a range of researchers in the field to find out whether comics are more beneficial or not.

In addition, the authors would like to examine the moderating factors in the second and third research questions. It was mentioned that the lack of consensus in the findings of studies investigating the effectiveness of comics in learning could be attributed to differences in the experimental procedures. However, to bolster the mapping between theory, research questions, and hypotheses, I would recommend providing a further theoretical basis and

explanation as to why authors think that comics have a greater impact on learning than noncomics for STEM vs. non-STEM subjects (e.g., maybe technical materials could benefit more from visualisation and engaging style of presentation). Relatedly, although this is an exploratory part, I would recommend bolstering the argumentation of why authors think that selected categories should be examined and why these were selected and maybe also briefly elaborating on why there should be a difference in effectiveness in selected categories. These aspects are essential and are related to the research questions that will be addressed.
Response: We have updated *Table 1* towards an exploratory path and we have added one paragraph on page 8 regarding the STEM vs non-STEM, as follows:

"We specifically aim to examine potential differences between STEM and non-STEM contexts, as panels in comics, by providing external visualizations to students, might scaffold learning differently for STEM and non-STEM topics, especially for individuals with low spatial abilities, considering the relationship between visualization, spatial ability efficacy, and STEM learning (Newcombe, 2013; 2017; Zhu et al., 2023)."

The research questions and related hypotheses are clear. Formulated hypotheses are capable of answering the research question. Interpretation of possible results is provided (but as mentioned before, relation to some further theoretical basis could be beneficial).

The protocol is detailed and provides sufficient information. For the study search, authors aim to strive for completeness; search terms (e.g., comic*) and databases for search (i.e., Scopus, WOS, and PubMed) are provided. The authors will also conduct a search based on references from reviewed articles and contact authors, which is a good strategy. I am thinking about a way that can help cover grey literature (e.g., conference proceedings/theses) more thoroughly, but I am not sure here (maybe a search index with broader coverage, e.g., Google Scholar or databases such as OPENGREY.EU can be helpful).
Response: We have considered that option too, and thank you for the useful suggestion, but we have decided that we are only interested in empirical peer-reviewed published work, as peer-review can provide some evidence of study quality. However, we do appreciate the suggestions a lot and we decided to extend our research sources and add Google Scholar and opengrey.eu.

Study selection, inclusion, exclusion criteria, and data extraction template are provided in sufficient detail. The authors will follow the guidelines of the PRISMA statement (Page et al., 2020); and we will present the PRISMA 2020 Main Checklist and the PRISMA 2020 Abstract Checklist.

Although the planned statistical analysis is sound, I have some suggestions and minor tips based on my readings of literature dedicated to the topic of effect size and meta-analysis. Please note that these are intended as a way of improving the quality of proposal.

The authors plan to work with Cohen s d and interpret the effect size as low, moderate, or high, according to the Cohen benchmarks. This is common practice in research literature. However, these benchmarks are not optimal for interpreting the size of the effect, as they were suggested by Cohen for power analysis in situations where no other information is provided. Also, these benchmarks are arbitrary (see, e.g., Correll et al., 2020). Therefore, the

interpretation of effect size can be rather based on empirically derived benchmarks (e.g., Bosco et al., 2015; Gignac & Szodorai, 2016; Paterson et al., 2016; Schäfer & Schwarz, 2019), or alternative approaches such as the accumulation of the effect over time (Funder & Ozer, 2019) or probability of superiority/common-language effect size (PS/CLES; McGraw & Wong, 1992). These options seem like more meaningful solutions that can help the reader to understand the magnitude of the examined effect.

Response: We agree with the recommendation of the reviewer that the interpretation of the effect size should not be arbitrary. We have now defined a minimum effect size of interest, based on the results from a previous, similar meta-analysis, as shown below:

"For our meta-analysis, we consider an effect size $d = 0.4$ as the minimum effect of interest. This is derived by a previous meta-analysis that investigated the effectiveness of comics in education (Topkaya et al, 2023). In the study by Topkaya et al. (2023), a meta-analysis with subject area as a moderator variable (similar to our STEM and non-STEM categorisation) resulted in an overall effect size of $g = 0.50$, 95% CI [.33, .68]. Using the lower bound of the confidence intervals as a heuristic, and considering the Hedge's g correction, we define $d = 0.4$ as the minimum effect of interest to guide our analyses and interpretation."

I also have some suggestions based on my readings of the work of Borenstein and his colleagues (Borenstein, 2019; Borenstein et al., 2021) dedicated to common misconceptions when conducting and interpreting the results of meta-analysis. First, although I agree that the random effects model is preferable in the present context, justification of this decision should be provided – i.e., why the random effects model is preferred over fixed effect/effects should be explicitly justified as this is crucial analytical choice (e.g., studies in the analysis are representative of a large universe of studies and goal is to make an inference o that universe – beyond the included studies). Also possible violations of assumptions should be discussed (at least later in the limitation in the discussion section (e.g., studies in the analysis might not be representative of studies actually performed – comment related to grey literature). Relatedly, if random effects meta-analysis is used and a number of studies is currently unknown but it could be small and heterogeneity substantial (as indicated in the introduction), I would recommend using the Knapp-Hartung adjustment.

Response: We thank the reviewer for mentioning the choice of random over fixed effects. It is our understanding that meta-analytic research (especially within the social sciences) is now relying on random effects model, since the assumption of a fixed effect model (i.e., a homogenous "true" effect) can very rarely be assumed. As such, a very strong rationale would be required in the case of choosing a fixed effect model. Though, following the recommendation, we have now added the following, in support of our choice for the random effects model in p. 16-17:

"We will use the random effects model, which controls for the possibility that the true effect size may vary from study to study, providing a more flexible and robust analysis (Kanters, 2021)"

Further, we now report that we will employ the Knapp-Hartung adjustment, if the number of studies is less than 20, following recommendations from previous work, as below:

"If the total number of identified studies is less than 20 ($k < 20$), we will apply the Knapp-Hartung (Knapp & Hartung, 2003) adjustment to our random effects model. The Knapp-Hartung adjustment has been shown to reduce the chance of false positive findings (Langan et al., 2018), and has been recommended when the number of studies is less than 20 (IntHout et al., 2014)."

Also, I would like to appreciate that prediction intervals will be provided since this interval captures the extent of dispersion of effect, and this is done in the same metric as the effect size. This is important for the reader to assess heterogeneity in an intuitive way.
Response: Thank you!

The authors also plan to evaluate heterogeneity "using the I^2 index, which, according to Higgins et al. (2003), can be described as low, moderate, and high, when it falls close to 25%, 50%, and 75%, respectively"; however, I have some reservations about this strategy. Of course, I^2, Q, and related statistics, should be reported and interpreted. Nevertheless, although it is a common practice to interpret I^2 in this way, there are some problems with this interpretation, as further argued by Borenstein (2019). In particular, I^2 can be beneficial and help to understand the forest plot and to examine the sampling error, but I^2 speaks about the proportion (i.e., what proportion of the variance in observed effects reflects variation in the true effect, rather than sampling error), not the variation per se. Therefore, it does not tell the reader much about the amount of variation in an absolute sense. Relatedly, although a relatively common practice, categorising *I2* as low, moderate, or high is not optimal as what was considered high in the context of Higgins study could be low in other contexts and vice versa. Therefore, the idea that I^2 captures the dispersion outside the original context of Cochrane database used Higgins study is questionable. Third, the authors note that moderator analyses will be conducted if significant heterogeneity is found. I understand logic here. However, the nonsignificant p-value is a function of thing other than the estimated amount of heterogeneity, namely the precision of individual studies and the number of studies in metaanalysis. Therefore, the p-value may not be statistically significant even when the estimated heterogeneity is substantial or may be significant even if it is practically trivial. These issues are further discussed by Borenstein (2019) and Borenstein et al. (2021) - these resources could be beneficial for interpretations related to heterogeneity and authors can consult them if they wish.
Response: We thank the reviewer for the valuable material on the limitations of I^2. We have now updated the manuscript, at page 17, to note that additional approaches that we will use to explore potential heterogeneity, following also the recommendations of Reviewer 1 (Dr. ***Adrien Fillon***). The adjustments are also provided below:

"Heterogeneity will be quantified using the I^2 index (Higgins et al., (2003) and will be visually examined through *p*-value drapery plots (Rücker & Schwarzer, 2021). Mikolajewicz and Komarova (2019) provide a comprehensive summary for how Cohen's *d*, tau^2, *Q*, and I^2 are formulated. If significant heterogeneity is suspected, as reflected through significant

heterogeneity (I^2 , $p < .05$), high heterogeneity variance ($I^2 > 25\%$; Higgins et al., 2003), or prediction regions broader than the overall *p*-curve in the drapery plot (Rücker & Schwarzer, 2021), then moderator analyses will be conducted. ”

A forest plot will be used for visualisation and a funnel plot will investigate small study bias. Egger´s regression and the trim and fill method will be used. It is mentioned that if a small study bias is identified through visual inspection and Egger's regression test, authors will proceed with adjustments to the funnel plot using the Duval and Tweedie (2000) trim and fill method. However, the exact criteria would be beneficial. Authors also mention that "the adjusted funnel plot will then be visually inspected to identify the direction of bias" however, would they also provide adjusted effect size due to publication bias and other reasons? If yes, this should be stated. If not, it should be explained why not. I appreciate the plan to conduct a sensitivity analysis.
Response: Following also the recommendations of Reviewer 1 (Dr. Adrien Fillon) we now provide a more thorough plan for the investigation of bias. We also want to thank the reviewer for pointing out the reporting of the adjusted effect size. We plan to report the trim-and-fill adjusted effect size and we have now added this information in the manuscript. Of note, if possible, we will also attempt further sensitivity analyses exploring bias and heterogeneity, including 3PSM and PET-PEESE. However, as the validity of these analyses depend heavily on various parameters which are out of our control for the purpose of a registration (e.g., overall effect size, *k,* τ), we choose not to register them at this stage, but conduct them as exploratory if the necessary assumptions are met at Stage 2. The adjustments made in the report, at page 18, are provided below:

"Further, we will employ multiple approaches to investigate potential heterogeneity and small study bias considering that it is recommended to employ various methods to investigate bias for meta-analyses in psychology (Carter et al., 2019). Specifically, we will use *p*-value drapery plots (Rücker & Schwarzer, 2021) and funnel plots to visually investigate small study bias. In addition, *p*-curves will be generated (Simonsohn et al., 2020) and tested for skewness and flatness using a χ^2 Binomial test, while the symmetry of the effect sizes will be examined using the Egger's regression test (Egger et al., 1997). If small study bias is identified through the visual inspection and/or the Egger's regression test, we will proceed with adjustments of the funnel plot using the Duval and Tweedie (2000) trim and fill method. The adjusted funnel plot will then be visually inspected to identify the direction of bias, and an adjusted overall effect size based on the trim and fill correction will be estimated."

In sum, I would like to thank the authors for their work on study proposal. I evaluate the protocol positively (e.g., the research topic is interesting and practically important; research questions are scientifically justifiable and fall within established ethical norms; clarity and degree of methodological detail are sufficient to replicate the proposed study closely; hypotheses stem from a theory (to reasonable degree) and methodology and analytic pipeline are sound, considering the existing standards. However, as detailed in the text, there are some suggestions that authors should consider before principal acceptance.

P. Kačmár, PhD.

References:

Borenstein, M. (2019). *Common Mistakes in Meta-Analysis and How to Avoid Them*. Biostat, Inc.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (Second edition). Wiley.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'Small', 'Medium', and 'Large' for Power Analysis. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2019.12.009

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An Assessment of the Magnitude of Effect Sizes: Evidence From 30 Years of Meta-Analysis in Management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813