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. 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)? 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.

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). 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.

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.

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.

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 meta-analysis. 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.

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.

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