Peer Community In
Health & Movement Sciences

Bayesian cumulative evidence synthesis and identification of questionable research practices in health & movement science

Wanja Wolff and *Jérémie Gaveau* *based on peer reviews by* *Maik Bieleke* *and 1 anonymous reviewer*

Hannah Corcoran, James Steele (2024) Cumulative evidence synthesis and consideration of ”research waste” using Bayesian methods: An example updating a previous meta-analysis of self-talk interventions for sport/motor performance. Sportrxiv, ver. 2, peer-reviewed and recommended by Peer Community in Health and Movement Sciences.
https://doi.org/10.51224/SRXIV.348

Submitted: 27 November 2023, Recommended: 20 November 2024

Cite this recommendation as:
Wolff, W. and Gaveau, J. (2024) Bayesian cumulative evidence synthesis and identification of questionable research practices in health & movement science. *Peer Community in Health and Movement Sciences*, 100044.
10.24072/pci.healthmovsci.100044

Published: 20 November 2024

Research is a resource-demanding endeavor that tries to answer questions such as, “Is there an effect?” and “How large or small is this effect?” To answer these questions as precisely as possible, meta-analysis is considered the gold standard. However, the value of meta-analytic conclusions greatly depends on the quality, comprehensiveness, and timeliness of the meta-analyzed studies, while not neglecting older research. Using the established sport psychological intervention strategy of self-talk as an example, Corcoran & Steele demonstrate how Bayesian methods and statistical indicators of questionable research practices can be used to assess these questions [1].

Bayesian methods enable cumulative evidence synthesis by updating prior beliefs (i.e., knowledge from an earlier meta-analysis) with new information (i.e., the studies that have been published on the topic since the earlier meta-analysis had been published) to arrive at a posterior belief - an updated meta-analytic effect size. This approach essentially tells us whether and how much our understanding of an effect has improved as additional evidence has accumulated; as well as the precision with which we are estimating it. Or to put it more bluntly, how much smarter are we now with respect to the effect we are interested in?

Importantly, the credibility of this updated effect depends not only on the newly included studies but also on the reliability of the prior beliefs – that is, the credibility of the effects summarized in the earlier meta-analysis. A set of frequentist and Bayesian statistical approaches have been introduced to assess this (for a tutorial with worked examples, see [2]). For example, methods such as the multilevel precision-effect test (PET) and precision-effect estimate with standard errors (PEESE) [2] can be used to adjust for publication bias in the meta-analyzed studies, providing a more realistic estimation of the effect size for the topic at hand. This would then help to assess the magnitude of the true effect in the absence of any bias favoring the publication of significant results.

Why does it matter for health and movement science?

The replication crisis and evidence of questionable research practices has cast doubts on various findings across disciplines [3–8]. Compared to other disciplines (e.g., psychology [9]), health & movement science has been relatively slow to recognize issues with the potential replicability of findings in the field [10]. Fortunately, this has started to change [10–14]. Research on factors that might negatively affect replicability in health & movement science has revealed evidence for various questionable research practices, such as publication bias [12,13], lack of statistical power [11,13], and indicators of p-hacking [12]. The presence of such practices in original research does not only undermine trustworthiness of individual studies, but also the conclusions drawn from meta-analyses that rely on these studies.

Open Science practices, such as open materials, open data, pre-registration of analyses plans, as well as registered reports are all good steps for improving science in the future [15–17] and might even lead to a 'credibility revolution' [18]. However, it is also crucial to evaluate the extent to which an existing body of literature might be affected by questionable research practices and how this might affect conclusions drawn from the research. Using self-talk as an example, Corcoran and Steele demonstrate this approach and provide a primer on how it can be effectively implemented [1]. By adhering to Open Science practices, their materials, data, and analyses are openly accessible. We believe this will facilitate the adoption of Bayesian methods to cumulatively update available evidence, as well as making it easier for fellow researchers to comprehensively and critically assess the literature they want to meta-analyze.

References

[1] Corcoran H. & Steele, J. Cumulative evidence synthesis and consideration of ”research waste” using Bayesian methods: An example updating a previous meta-analysis of self-talk interventions for sport/motor performance. SportRxiv, ver.2, peer-reviewed and recommended by PCI Health & Movement Sciences (2024). https://doi.org/10.51224/SRXIV.348

[2] Bartoš, F., Maier, M., Quintana, D. S. & Wagenmakers, E.-J. Adjusting for publication bias in JASP and R: Selection Models, PET-PEESE, and robust bayesian meta-analysis. Adv. Methods Pract. Psychol. Sci. 5, 25152459221109259 (2022). https://doi.org/10.1177/2515245922110925

[3] Yong, E. Replication studies: Bad copy. Nature 485, 298–300 (2012). https://doi.org/10.1038/485298a.

[4] Hagger, M. S. et al. A multilab preregistered replication of the ego-depletion effect. Perspect. Psychol. Sci. 11, 546–573 (2016). https://doi.org/10.1177/1745691616652873

[5] Scheel, A. M., Schijen, M. R. M. J. & Lakens, D. An excess of positive results: Comparing the standard psychology literature with registered reports. Adv. Methods Pract. Psychol. Sci. 4, 25152459211007467 (2021). https://doi.org/10.1177/2515245921100746

[6] Perneger, T. V. & Combescure, C. The distribution of P-values in medical research articles suggested selective reporting associated with statistical significance. J. Clin. Epidemiol. 87, 70–77 (2017). https://doi.org/10.1016/j.jclinepi.2017.04.003

[7] Errington, T. M. et al. An open investigation of the reproducibility of cancer biology research. eLife 3, e04333 (2014). https://doi.org/10.7554/eLife.04333

[8] Hoffmann, S. et al. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. R. Soc. Open Sci. 8, 201925 (2021). https://doi.org/10.1098/rsos.201925

[9] Open Science Collaboration. Estimating the reproducibility of psychological science. Science 349, aac4716 (2015). https://doi.org/10.1126/science.aac4716

[10] Mesquida, C., Murphy, J., Lakens, D. & Warne, J. Replication concerns in sports and exercise science: A narrative review of selected methodological issues in the field. R. Soc. Open Sci. 9, 220946 (2022). https://doi.org/10.1098/rsos.220946

[11] Abt, G. et al. Power, precision, and sample size estimation in sport and exercise science research. J. Sports Sci. 38, 1933–1935 (2020). https://doi.org/10.1080/02640414.2020.1776002

[12] Borg, D. N., Barnett, A. G., Caldwell, A. R., White, N. M. & Stewart, I. B. The bias for statistical significance in sport and exercise medicine. J. Sci. Med. Sport 26, 164–168 (2023). https://doi.org/10.1016/j.jsams.2023.03.002

[13] Mesquida, C., Murphy, J., Lakens, D. & Warne, J. Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: potential barriers to replicability. J. Sports Sci. 41, 1507–1517 (2023). https://doi.org/10.1080/02640414.2023.2269357

[14] Büttner, F., Toomey, E., McClean, S., Roe, M. & Delahunt, E. Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. Br. J. Sports Med. 54, 1365–1371 (2020). https://doi.org/10.1136/bjsports-2019-101863

[15] Chambers, C. D. & Tzavella, L. The past, present and future of Registered Reports. Nat. Hum. Behav. 6, 29–42 (2022). https://doi.org/10.1038/s41562-021-01193-7

[16] Soderberg, C. K. et al. Initial evidence of research quality of registered reports compared with the standard publishing model. Nat. Hum. Behav. 5, 990–997 (2021). https://doi.org/10.1038/s41562-021-01142-4

[17] Wunsch, K., Pixa, N. H. & Utesch, K. Open science in German sport psychology. Z. Für Sportpsychol. 30, 156–166 (2023). https://doi.org/10.1026/1612-5010/a000406

[18] Korbmacher, M. et al. The replication crisis has led to positive structural, procedural, and community changes. Commun. Psychol. 1, 1–13 (2023). https://doi.org/10.1038/s44271-023-00003-2

Reviews

Evaluation round #2

Reviewed by Maik Bieleke, 21 September 2024

I have carefully reviewed the authors' point-by-point responses to my comments and thoroughly examined the revised manuscript. The revisions are substantial, with a much clearer focus on cumulative evidence synthesis, which I believe has significantly strengthened the manuscript. While the authors have retained the term ”research waste,” they now present a balanced rationale for its use. I appreciated the addition of the new section simulating the impact of adding a single study into the existing body of evidence, which emphasizes a particularly intriguing aspect of the authors' findings and provides valuable guidance for researchers making informed decisions.

Overall, I found the revised manuscript highly engaging and informative. All of my comments have been satisfactorily addressed, and I am pleased to fully recommend the mansucript for acceptance.

Reviewed by anonymous reviewer 1, 12 November 2024

I would like to thank the authors for their answers regarding my comments from the previous round. I am really impressed how thoroughly the authors engaged with my comments. I enjoyed reading both the revised paper and the authors' thoughts on the reviewers' comments. So, I do not see any further issues with the manuscript in its current form, and I do not have any further recommendations to make.

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.51224/SRXIV.348
Version of the preprint: 1

Authors' reply, 19 August 2024

Download author's reply

Decision by Wanja Wolff and Jérémie Gaveau, posted 11 March 2024, validated 11 March 2024

Decision on your manuscript: Revision invited

I have now received two reviews of your manuscript "Self-talk interventions and sport/motor performance: An updated systematic review and Bayesian meta-analysis", and I have read the paper carefully myself. As you can see from the notes below, the reviewers are very sympathetic to the submitted paper. I fully agree with this; and I particularly like the approach of conducting a Bayesian meta-analysis to update our prior beliefs about self-talk effectiveness with new knowledge (also: I really enjoyed the Figures, showing the prior and posterior distributions. This is very intuitive and transparent to the reader). The reviewers also raised some important points that could be addressed to further strengthen the paper. I agree with the reviewers' assessments. This pertains especially to:

- the rather slim section on self-talk
- the positioning of the paper as a true meta-analysis on self-talk interventions vs. a primer on the benefits of cumulative science
- the insinuation that the relative non-change in posterior estimates indicated that self-talk intervention research since 2011 has been "wasted research"

I will not risk confusing matters by fully reiterating the reviewers' comments (please consult the high-quality comments of the reviewers at the end of this letter).

One thing that I found to be deserving a bit more space were the rare occasions where posterior distributions actually changed. Given that everything else seemed to pretty much stay the same, wouldn't there be merit in diving a bit deeper into the moderators where the updated effects have changed?

Thank you very much for submitting your manuscript to PCI Health & Movement Science. I am looking forward to reading the revised version of the paper.

Best regards,
Wanja

Reviewed by Maik Bieleke, 16 January 2024

Review of the article: **"Self-talk interventions and sport/motor performance: An updated systematic review and Bayesian meta-analysis"**

The authors employ the existing literature on self-talk to demonstrate how Bayesian methods can enhance and update previous meta-analyses. The utilization of prior meta-analytical results in a Bayesian framework is an innovative and highly compelling approach to cumulative science. My comments, therefore, focus primarily on presentation aspects rather than the approach itself, as well as on some questions I had while reading the article.

1. Contrary to what the title suggests, I did not have the impression that the article is best framed as a meta-analysis of self-talk interventions. It does not align particularly well with established standards for meta-analyses

(PRISMA guidelines; e.g. flow charts, risk-of-bias analyses) and there are only very limited references to the self-talk literature. Rather, the article seems better described as a methodological contribution that demonstrates the benefits of Bayesian methods for cumulative science, using self-talk interventions as a convenient but largely interchangeable example. The keywords chosen by the authors reinforce this impression by omitting any reference to self-talk and meta-analysis. Therefore, a clearer and more consistent positioning of the article as valuable methodological contribution to cumulative science, rather than a standalone meta-analysis, might enhance clarity in this regard.

2. Whether or not the authors follow my previous suggestion, I think that a brief introduction to research on self-talk and self-talk interventions would aid readers in comprehending the narrative of the article (e.g., understanding and interpreting the moderator analyses). For example, the term self-talk is never explicitly introduced or defined, and the ”key distinction” (p. 13) between strategic and organic self-talk is only mentioned towards the end of the discussion.

3. On a quantitative note, I found it surprising that the confidence interval around the overall effect size remained virtually unchanged, although the amount of evidence doubled in terms of studies and effect sizes. Interestingly, the posterior CIs of the various moderators were indeed narrower than the prior CIs, an increase in precision one would expect when sample sizes increase. I wondered whether this might reflect unobserved qualitative differences between studies of some kind (e.g., additional moderators; also see comment 6 below) but maybe there is a simpler explanation that I'm missing.

4. In the moderators section, the authors assert minimal differences between posterior estimates and priors. However, looking at Figure 2 it seems that the most extreme differences consistently decreased across moderator variables, pushing the posterior distributions closer to the overall effect size. Admittedly, this visual impression may be challenging to quantify or test, but it would be interesting to know the authors interpretation.

5. I wondered whether ”research waste” really is an appropriate label for all of the studies included in the present analysis. While these studies might have replicated the self-talk intervention effect, maybe this was not their main and/or only goal? Isn't it possible that these studies relied on the already established effect to answer a different question? For example, one might think of studies comparing self-talk as established intervention to a novel intervention. Such a study would likely be included in the present analysis due to its replication of the self-talk intervention effect, but its contribution mainly relates to examining the novel intervention. The term ”research waste” then might overlook important contributions from these studies that go beyond (possibly unnecessary) replication efforts.

6. Relatedly, it might be worth noting that the analysis focused solely on quantitative aspects of replication (effect sizes, confidence intervals). This is fine but the authors also mention theoretical and conceptual advancements in the literature on self-talk. For somebody unfamiliar with the self-talk literature, it seems implausible that such improvements did not affect in any way how post-Hatzigeorgiadis self-talk interventions were delivered and/or evaluated. However, if administration or evaluation methods have changed (and potentially improved), the fact that our beliefs about intervention effects still remain mostly unchanged actually is reassuring, no? A qualitative analysis of the included studies might go beyond the scope of the present study, but it is still worth discussing whether and how it limits the interpretation of the quantitative analysis.

7. In the introduction, the authors argue that ”the effect estimate from the meta-analysis of Hatzigeorgiadis et al. (2011) was already fairly precise”, implying that further research was unwarranted to begin with. I found the statement rather vague and the implication drawn with the benefit of hindsight. When is an estimate precise enough to discourage further replications, and aren't there other reasons for replicating an effect (e.g., robustness checks, changes in methodology) that warrant investigation?

8. There is currently no discussion of the potential strengths and limitations of the study. I think that adding this would make it easier for readers to gauge the contribution of the present research.

Minor points:

- In the abstract, there seems to be a wording error in this sentence? ”total of 34 studies providing 128 effects nested in 64 groups across experiments 42 were included in the final updated meta-analysis representing data from 18761 participants”

- In Figure 2, it would be helpful to also have the original estimates and CIs as numbers (maybe using the same color coding as for the distributions).

Reviewed by anonymous reviewer 1, 08 March 2024

The present paper presents a Bayesian meta-analysis on the effects of self-talk on motor performance. I enjoyed reading the manuscript and I congratulate the authors on their work. The writing of the paper was very accessible, the authors put a lot of effort into their study and I was impressed by the methods they employed. So, there were many things about the paper that I liked.

At the same time, I have a couple of points that I would like to see addressed. I do not consider these points to represent ‘flaws’ or something like that, but rather suggestions that the authors might want to consider. In my point of view, dealing with these points might further strengthen the present paper, but of course the authors are free to disagree here.

1)

Meta-analyses (as the authors of course know) estimate an average effect and deviations from this effect. Thus, meta-analyses only make sense when it is theoretically or conceptually meaningful to estimate an average effect over a set of studies. The Data Colada team makes this point better than I could, and they illustrate it with some nice examples (please see `http://datacolada.org/104` and the related posts).

So, I was wondering: Does it make sense (both from a theoretical and an applied perspective) to estimate the average effect of self-talk? I am not saying that it doesn’t, I simply do not know. In other words: The more the different self-talk interventions resemble each other (both in terms of theoretical foundations and in terms of their implementation), the more sense it makes to conduct a meta-analysis regarding their average effectiveness. Maybe the authors can address this point and add a couple of words on the similarity of the analyzed interventions.

2)

I was a bit surprised that the authors did not assess potential publication bias, but maybe I missed something. It is my understanding that publication bias (e.g., only positive results are published) poses a serious threat to meta-analyses, and that therefore tools to assess publication bias have been developed. These tools may range from funnel plots to techniques such as Trim-and-fill, PET, PEESE and so on (e.g., Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. Advances in Methods and Practices in Psychological Science, 2, 115-144. `https://doi.org/10.1177/2515245919847196`). Why did the authors not address potential publication bias? As far as I am concerned, one reason why results have not changed from 2011 till now might be that results that fit the picture get published easier?

3)

I do not understand the authors’ point that all research in the meantime between the first meta-analysis and theirs was “a waste of research”, simply because their meta-analysis comes to the same conclusion as the first one. To me, this statement is somewhat questionable for several reasons (at least in the way it is presented in the present form of the article).

First, the argument is only logically valid if we suppose that the _only_ goal of the studies that were analyzed was to assess the effectiveness of self-talk. However, if some of the studies had additional goals, then the “waste of research” conclusion is not valid anymore. Please let me illustrate this point with a couple of examples.

Suppose that a study aimed to find out whether self-talk works better for soccer players than for basketball players. The results show that it doesn't. In this case, we have learned something from the respective study, even if in a meta-analysis the study gets the same effect size like in a previous meta-analysis.

Or suppose that a study aimed to find out whether self-talk works better for handball players than for volleyball players. The results suggest that it works above average for handball players and below average for volleyball players. Again, the average effect would be the same as in the previous meta-analysis, but again I assume that we have learned something.

On a very basic level, I would like to argue that in order to decide whether a study was a waste of research and resources you have to take into account that very study's goals. Therefore, I was wondering: Was the only goal of the studies analyzed in the present meta-analysis really to assess the average effectiveness of self-talk?

Second, I was wondering: When exactly does additional research become "a waste"? I totally agree that research is not efficient anymore when it does not add anything to established knowledge anymore. However, when is knowledge really established in psychology? Considering the average sample sizes in psychology, even 10 psychological studies might not have the sample size that might be considered necessary to yield established knowledge in other fields. Again, I am not saying that there will not eventually come a point where further research becomes unnecessary and thus inefficient. I am simply wondering: When does it come?

Third, the authors themselves write (p. 2 / 3) that research on self-talk might have matured post 2011 regarding "... efforts to improve operationalisation/measurement, and efforts to improve methodology used in studying self-talk". Quite frankly, I cannot reconcile this observation with the statement that this research was a waste. I mean, finding the same effects with better measurement and methods should increase our confidence both in self-talk per se, but also in the results prior to 2011, or not? Suppose you have a study that finds a certain effect, but this study has weaknesses. Now a study with better methods finds the same effect. Would you really want to argue that the second study was a waste, because it finds the same effect? I surely would not.

Of course, I may be missing something here, but other readers might miss these things as well. So even if the authors disagree with me, they might want to add a couple of words.