Response to Reviewers

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

Dear Professor Wolff,

Many thanks for considering our manuscript and placing it under review with *Peer Community in Health and Movement Sciences*. We are very grateful for your efforts in obtaining reviews. We are very impressed with the quality and constructiveness of the feedback provided by the reviewers which we feel has significantly improved the quality of our manuscript.

Note, because of the extensive rewrite of the manuscript due to the shift in focus we have not used track changes or highlighting throughout, though changes can be viewed in the GitHub repository commits and versions. However, please see below for our comments in response to some of the key points raised by the reviewers.

Again, we would like to thank all involved in the process of reviewing this manuscript and look forward to the next round of reviews in considering our revisions.

Many thanks

The authors

Reviewer Comments and Author Responses

Reviewer 1 - Maik Bieleke

Reviewer Comment

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.

Author Response

Thank you for your suggestion. It is actually a point which in retrospect after submitting the manuscript we realised more ourselves and so it is gratifying to have this reinforced by both reviewers. For clarity, the absence of certain methodological details such as those noted did stem partly from us wanting to replicate the methods employed by Hatzigeorgiadis et al. as closely as possible. As this was an undergraduate project being completed in a short period of time we opted to try to replicate and update to make the project more efficient under such constraints. We did initially intend though to include a PRISMA flow diagram to describe our updated searches, however we encountered an issue and realised that some of the searches we had conducted were not correctly recorded by the tool being used to manage the process. Unfortunately we were unable to exactly replicate the searches despite attempting to reproduce the searches again. We have added this detail to footnote 2 for transparency.

We do agree though that positioning this paper instead as an example of using Bayesian meta-analytic methods for cumulative science, whilst using the self-talk literature as an exemplar case-study given how long it has been since a quantitative evidence synthesis was conducted, is perhaps more suitable. As such we have attempted to rewrite the manuscript with this in mind, and have also changed the title to: *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*

Reviewer Comment

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.

Author Response

[Given the reframing of the paper, we have now subsectioned the introduction and include a section more explicitly introducing self-talk.

Reviewer Comment

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.

Author Response

This is an interesting, and somewhat counterintuitive result of our analysis. From the additional work we have conducted whereby we have utilised simulation to add a single new study and explore it's effects upon the prior distribution we were surprised that even a study simulated with a large sample of participants and either a null or large effect had next to no impact in updating the prior estimate. These simulations have been added into the manuscript as an example of what can be done in determining whether a prior estimate is already precise enough.

However, we conducted a little additional exploration regarding how the prior distribution was set to better understand this. We utilised a t-distribution ($t(k, \mu, \sigma)$) with $k - 2$ degrees of freedom as suggested by Higgins et al. (2009) cited in our manuscript. But, given the large degrees of freedom due to the large number of studies in the previous meta analysis from Hatzigeorgiadis et al., the t-distribution approximates a normal distribution more closely and thus has tight tails which influence the impact that an additional evidence might have on the prior when updating to a posterior (see this example from Richard McElreath - https://x.com/rlmcelreath/status/1701165075493470644?s=20). Out of curiosity we simulated a single study of one million participants and a null effect with a more skeptical t-distribution using only 3 degrees of freedom thus allowing the tails to be fatter. Even with this though ~85% of the entire posterior distribution was still within the 95% interval for the prior. To our interpretation this goes to show how certain the estimate from Hatzigeorgiadis et al. already was and highlights that further intervention research could be considered wasteful (assuming it has not contributed anything additional to our understanding). But, as we explain in response to reviewer 2, we have now examined using various methods the possibility of small sample and publication bias in the studies included since 2011. We think this paints a concerning picture, though one that is perhaps not that surprising given that there is widespread evidence of such questionable research practices across many fields including the sport and exercise sciences. Hatzigeorgiadis et al. only examined this possibility using the now known to be flawed fail-safe N and from it concluded that there was likely no publication bias present prior to 2011. We find this hard to believe given that it is highly likely that prior to 2011, when many consider the current replication crisis in psychology (of course which sport psychology sits close to)

kicked off, there were not many of the methodological reforms such as pre-registration and even registered reports, open data, materials, and code etc. in place that might help to mitigate somewhat the influence of QRPs such as publication bias. We didn't really want to end up down this route with the paper, and as noted due to it being an undergraduate project had deliberately conducted the updated systematic search to reduce the burden given the time available to complete the work. So we have not replicated the systematic search of Hatzigeorgiadis et al. for pre-2011. However, we have contacted the authors and asked for whether they still have their extracted data such that we might be able to examine this in the pre-2011 studies. We will be happy to update the manuscript with this if the authors make it available to us.

Reviewer Comment

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.

Author Response

We have now noted this shift towards the overall pooled effect and also included additional commentary on it in the discussion. We also relate this to the comments regarding whether or not the goals of studies were to examine self-talk interventions or more specific questions regarding mediators to highlight that even in this regard the newer evidence has not added a great deal to update our beliefs.

Reviewer Comment

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.

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.

Author Response

This is certainly a possibility and we initially in revising tried to soften the language regarding "research waste" somewhat throughout. But, given the minimal impact upon estimates of theoretically informed moderators and the concerns we have now raised regarding publication bias and QRPs we do still find it difficult to conclude anything other than that a lot of the research could indeed be considered wasteful. We had considered in light of this comment whether to conduct a qualitative appraisal of the included studies to determine the extent to which they had either merely estimated the effects of self-talk interventions or explored moderators which were already well understood (though this we now question due to the possibility of bias in prior estimates of main effects anyway) as opposed to contributing something new theoretically or methodologically. However, given the shift in the focus of the manuscript we opted not to do so but instead to highlight this as something those engaged in cumulative evidence synthesis should consider as well when determining whether research could be considered wasteful, or in the justification of a new trial by dint of the fact it might address theoretically or methodologically novel considerations.

Reviewer Comment

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?

Author Response

We have now attempted to address this concern through the addition of simulations of the effects that a new study might have had post the meta-analysis of Hatzigeorgiadis et al. in terms of changing our conclusions regarding the effect. This lends some quantitative precision the the claim that it was already fairly precise and that further research may have been

unwarranted. Of course there may be other reasons to continue further research attempting to replicate an effect, and indeed in this case it could be argued that further research might be needed following more modern open research practices including pre-registration and sufficient power/precision for effect estimation in order to actually establish in light of the presences of publication bias and QRPs whether self-talk interventions actually have an effect at all anyway.

Reviewer Comment

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.

Author Response

Given the reframing we instead have added discussion of the benefits of engaging with the kinds of methods demonstrated, but have also noted that supplementing this with qualitative appraisal is of value in determining whether additional research is warranted.

Reviewer 2 - Anonymous

Reviewer Comment

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.

Author Response

This is a good question and one which we encounter often. Specifically in a random effects meta-analysis the overall main estimate is the weighted average of the effect*s* that each study has estimated. It begins with the assumption that the specific effect being estimated does differ across the studies for a range of factors mostly relating to the fact that they are all conceptual replications as opposed to direct replications. The benefit though is that we can quantify the heterogeneity via the τ parameter and determine whether it is meaningful enough to be concerned about. In addition, the moderator analyses allow for us to explore whether or not study, experiment, or group level factors might be influences of the effects that each study estimates within the overarching distribution of effects. So, we would argue that in the case of asking the question about the effects of an intervention, such as self-talk upon an outcome, such as sport/motor performance, meta-analysis can be appropriate if conducted appropriately.

Reviewer Comment

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?

Author Response

Thank you for this suggestion. Somewhat in our haste and our focus on the updating aspect of the novel approach taken, we failed to consider this. However, we have now examined using various methods the possibility of small sample and publication bias in the studies included since 2011. The PET-PEESE, mixture model for *p*-hacking, and robust Bayesian meta-analysis model averaging methods reveal evidence of publication bias and *p*-hacking and that the adjusted effect may in fact be null considering the former. This paints a concerning picture to our interpretation, though one that is perhaps not that surprising given that there is widespread evidence of such questionable research practices across many fields including the sport and exercise sciences. Hatzigeorgiadis et al. only examined this possibility using the now known to be flawed fail-safe N and from it concluded that there was likely no publication bias present prior to 2011. We find this hard to believe given that it is highly likely that

prior to 2011, when many consider the current replication crisis in psychology (of course which sport psychology sits close to) kicked off, there were not many of the methodological reforms such as pre-registration and even registered reports, open data, materials, and code etc. in place that might help to mitigate somewhat the influence of QRPs such as publication bias. We didn't really want to end up down this route with the paper, and as noted due to it being an undergraduate project had deliberately conducted the updated systematic search to reduce the burden given the time available to complete the work. So we have not replicated the systematic search of Hatzigeorgiadis et al. for pre-2011. However, we have contacted the authors and asked for whether they still have their extracted data such that we might be able to examine this in the pre-2011 studies. We will be happy to update the manuscript with this if the authors make it available to us.

Reviewer Comment

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?

Author Response

Reviewer 1 also raised this concern and as mentioned we orginally in revising this work tried to soften the language regarding "research waste" somewhat throughout and also emphasise this limitation more in the discussion. We had considered in light of this comment whether to conduct a qualitative appraisal of the included studies to determine the extent to which they

had either merely estimated the effects of self-talk interventions or explored moderators which were already well understood (though this we now question due to the possibility of bias in prior estimates of main effects anyway) as opposed to contributing something new theoretically or methodologically. However, given the shift in the focus of the manuscript we opted not to do so but instead to highlight this as something those engaged in cumulative evidence synthesis should consider as well when determining whether research could be considered wasteful, or in the justification of a new trial by dint of the fact it might address theoretically or methodologically novel considerations.

Reviewer Comment

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?

Author Response

We have tried to better address this by leveraging simulating the addition a single new study, varying the underlying effect size and sample size, and explore it's effects upon the prior distribution. This is to demonstrate that such an approach can aid in determining exactly when the addition of new studies, whether small or indeed large, might be unnecessary in light of current evidence.

Reviewer Comment

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.

Author Response

As noted we have tried to make it clear that such qualitative considerations should also be taken into account when determining whether research has been wasteful, or indeed further research might be. Although, for the specific example of self-talk the process of working on this project has been enlightening and left us in considerable doubt of the value of much of the research in the area (though notably that would be the case for many areas within the sport and exercise sciences… unfortunately we've become somewhat cynical).