Dear Prof. Dar,

Thank you very much for your and the reviewers’ thoughtful comments and suggestions on this manuscript. I think the manuscript is greatly strengthened by these changes.

I have addressed each of them below, point by point. I have included relevant quotes from the manuscript that illustrate the changes. In the one case where I have elected not to make the suggested change (Reviewer 1 point 3), I provide a clear argument for why I do not feel this would be in the best interests of the scientific record and the function of this critique.

Kind regards

[anonymised for peer review]

**Editor’s comments**

Dear Dr Hussey,    
  
Thank you for submitting your manuscript to the Journal of Behavior Therapy and Experimental Psychiatry.   
  
I have obtained two expert reviews of your work, both of which like your manuscript and are recommending that you revise and resubmit it to the Journal. Having read your submission, and the reviews, I'm inclined to agree with this recommendation. Please resubmit your revised manuscript by **Oct 18, 2024**.   
  
When revising your manuscript, attend carefully to the reviewers' and Associate Editor's comments below: please outline every change made in response to their comments and provide suitable rebuttals for any comments not addressed. Please note that your revised submission may need to be re-reviewed.

**Reviewer 1’s comments**

**Reviewer 1 point 1**

In this manuscript, the author criticizes a well-cited meta-analytic paper published in this journal by offering detailed checks of the reproducibility of the meta-analytic analysis, effect size computations, and selection of results. The many methodological issues raised by the author are severe and cast serious doubts on the main conclusions of the Vahey et al. meta-analysis, particularly its problematic recommendation to use small samples to study the validity of the Implicit Relational Assessment Procedure (IRAP). The journal should publish this criticism in a shortened form.

*Author response: Thank you for this very clear endorsement of the value of the manuscript.*

**Reviewer 1 point 2**

The author provides a lot of details and adds some interesting topics for meta-researchers who seek to assess the validity of meta-analysis, but both the details and meta-research topics distract from the main relevant finding that the main results of Vahey et al. meta-analysis do not stand up to scrutiny and that the proper selection, effect size computations, and meta-analuysis point at a mean effect size that is much smaller than as reported in the original paper.

*Author response: Thank you, this was exactly my intention for this manuscript: not merely to highlight the issues in Vahey et al. (2015) but to demonstrate how such issues can be found and evidenced in other meta-analyses. I must disagree that these details distract from the critique of Vahey et al. (2015) itself. This is not merely because of my desire for this manuscript to have this dual audience of IRAP researchers but also meta-researches: it is because many of the problems associated with the lack of reproducibility or the errors in Vahey et al. (2015) are the direct result of inadequate documentation and transparency and therefore the potential for easy verification. I do not wish to repeat those problems in my critique, nor do I wish to expose the manuscript to any critique that it does not, as a standalone document, provide evidence of its claims of errors and lack of reproducibility in Vahey et al. (2015).*

*After corresponding with the editor, who as I understood it stated that he does not have strong views on whether complying with this reviewer point is a deal-breaker for the manuscript, I therefore have elected not to move these details to supplementary materials. Although this is my strong preference, it is not a hard line for me either, and I would do so if it meant the difference between publishing this manuscript or not.*

**Reviewer 1 point 3**

Apart from placing some details to an appendix, the comment can also be restructured and shortened to reflect the standard steps in a systematic review, namely problem formulation (including clear construct and operational definitions), literature search, study selection, computation of effect sizes, analysis, and reporting. Much of the trouble with the original M-A is already caused by an inadequate selection of studies (also raising questions about the literature search that was by no means reproducibly reported in my opinion) and awkward computation of effects sizes, sometimes even including within-subject effects that are by their nature very hard if not impossible to compare to between-subject effects of relevance to the clinical domain. It is crucial that the problem formulation, including clear construct and operational definitions, is well articulated and linked in a meaningful (psychometrically informed) way to the selected effect sizes. Next, the critique should focus on the selection of effects from the primary studies. Substantively, it makes little sense to reproduce the overall analysis of a meta-analysis of effects that are wrongly selected and/or computed in the first place. Perhaps for meta-researchers it is interesting to study the computational reproducibility of the analyses as a case study, but for subsantive researchers it is little more than a post-mortem inquiry. I would suggest focusing on the main results based on the right data and analysis rather than on the erroneously selected and computed effect sizes. In other words, it would be better for the readers of this journal to redo the systematic review and meta-analysis using the criteria proposed by the authors and reproducible effect size computations from the current author, and present these results as the main outcome of the critique, together with short conclusions, implications, and future directions. Additional points with relevance for meta-researchers were interesting for me as a meta-researcher but not very clearly presented and perhaps less relevant for readers of this journal with an interest in the validity and usefulness of the IRAP.

*Author response: Thank you for this input. I very much sympathize with a desire for this evidentiary hole to be filled. Indeed, if the “apples and oranges” problem with integrating effect sizes with different estimands could be overcome (which I discuss in the manuscript), I would agree that a new and much improved systematic review and meta-analysis should be conducted. However, I must strongly disagree that it is the responsibility of critics to then also fully reconduct and improve the original flawed work.*

*Delays to scientific correction are very important to consider here. No parties involved seem to dispute the validity of my claims that Vahey et al. (2015) is extremely flawed. It is therefore important to communicate to the scientific community that Vahey et al. (2015) is flawed so that more work is not erroneously influenced by it or based on its recommendations. Vahey et al. (2015) had been cited 66 times when I first raised these concerns in 2019. It has now been cited 148 times – once more since I submitted this manuscript to JBTEP. The desire to have a better answer to the substantive question does not justify delaying the dissemination of the known problems with the original work, or delaying necessary action on the basis of this knowledge.*

*This request also risks misdiagnosing what this manuscript is and should be: it is not an attempt to provide the reader with a more accurate estimate of the IRAP’s criterion validity, it is a critique of an existing estimate of that estimand. Its core point is that Vahey et al. (2015) is so flawed that it should be substantially corrected or retracted, not merely that the literature should be expanded with a new answer to the question. The reviewer notes that the “readers of this journal with an interest in the validity and usefulness of the IRAP”, and I argue these readers would be best served by being informed of the flaws in Vahey et al. (2015) in a timely fashion. This is in line with COPE guidelines and Cochrane recommendations.*

*With all that said, again, I completely agree that if the “apples and oranges” problem with integrating effect sizes with different estimands could be overcome, that a new and much improved systematic review and meta-analysis should be conducted – but by the original authors of Vahey et al. (2015) or others with substantive interests in this domain. This is also in line with the principle of the burden of proof is on those asserting a claim (i.e., substantive claims about the IRAP are made by the original authors; whereas claims about errors in the original work are made here and I provide evidence for them).*

*In order to address these concerns or potential for misreadings of the results presented in the manuscript, I have added numerous reminders to the reader that the results are presented only in order to illustrate the compound impact of the errors found, and not to present the results as a better answer to the substantive question of the IRAP’s criterion validity. For example:*

“A new meta-analysis was calculated to understand the compound impact of these errors (i.e., without endorsing its results as a valid estimate of the IRAP’s criterion validity).” (abstract)

“Corrected meta-analysis and power analyses to illustrate the compound impact of the issues

In order to understand the compound impact of the various errors on the conclusions of the meta-analysis, I fitted a new meta-analysis to the 156 effect sizes re-extracted from the original articles. I then used the meta-analysis effect results to calculate new power analyses. Importantly, the purpose of this new meta-analysis was not to present its results as a more accurate estimate of the IRAP’s criterion validity, but rather to illustrate the compound impact of the various errors that were outlined above on the meta-estimate. I return to this point in the discussion.” (p. 35, methods)

“With that said, it is important to reiterate that purpose of the new meta-analysis and power analyses is to illustrate the compound impact of the observed errors on the results, and to illustrate that, by the original article’s logic, the IRAP literature is in general under powered. This should not be mistaken for an endorsement of the results of as a more accurate or valid estimate of the IRAP’s criterion validity.” (p. 39, discussion)

“Recalculated results suggested that the compound impact of the errors reduced the meta-effect size to less than half the original result ( = .45 vs. .22), and increased the sample size recommendations by more than 15 times the original results (minimum *N* = 37 vs. 346).” (p. 43, conclusion)

**Reviewer 2’s comments**

**Reviewer 2 point 1**

This verification report provides a rigorous attempt to replicate findings that have proven impactful for the field, for example by providing a guide for the required sample sizes in this research area. The verification attempt points out several important problems that peers in this field should be aware of. Overall, my conclusion is that this paper makes an important contribution, both to the content people who use this IRAP procedure, but also more generally about the awareness that meta-analyses can have large impact, but are not always performed to high standards. If the articles updates the sample sizes people need in this field, and raises the quality of meta-analysis, this is good for the field. One point not highlighted in the abstract, but also an important insight, was the 'apples and oranges' nature of pooling all these estimates (I think this deserves to be mentioned in the abstract as well).

*Author response: Thank you for this comment – I agree this point is important, but its surprisingly difficult to find a citation to support it (e.g., Cochrane doesn’t actually caution against it, oddly), and I didn’t want to have to argue the point from first principles. On your suggestion, I have emphasized this point slightly more. The abstract now includes the line:*

“The original results were also undermined by combining effect sizes with different estimands.”

In order to better emphasize this point in the manuscript itself, while also being clearer that the reported new meta-analysis should not be interpreted substantively, I have reworked and expanded the following paragraph that now appears between page 38 and 40, immediately before the limitations section.

*“With that said, it is important to reiterate that purpose of the new meta-analysis and power analyses is to illustrate the compound impact of the observed errors on the results, and to illustrate that, by the original article’s logic, the IRAP literature is in general under powered. This should not be mistaken for an endorsement of the results of as a more accurate or valid estimate of the IRAP’s criterion validity. The most important reason for this that the the analytic strategy employed in Vahey et al. (2015) and reproduced here pools effect sizes with distinct estimands. Although it is mathematically possible to convert some of these effect sizes to a common scale such as Pearson’s r (with the exceptions of the partialized effect sizes that were erroneously converted, as discussed previously), this does not mean that these effect sizes have a common estimand. That is, they estimate fundamentally different properties. As Borenstein et al. (2009) put it: “even if there is no technical barrier to converting the effects to a common metric, it may be a bad idea from a substantive perspective” (p. 46). Additionally, the effect sizes themselves are often derived from different IRAP data (e.g., single trial types, multiple different forms of averaged trial types), again changing the estimand (e.g., from the IRAP’s criterion validity to the criterion validity of one or more trial types on a given IRAP while ignoring the remaining trial types). Additionally, all the following effects were meta-analyzed together: effect sizes representing the magnitude of the compatibility effect on the IRAP itself, interaction effects between IRAP trial types and group allocations, and correlations between IRAP trial types and criterion tasks. In doing so, different types of IRAP data were combined as one: data from single trial-types, overall effects for the whole task, and effects averaging the trial types in different ways. Lastly, effects treating the IRAP as the dependent variable, the independent variable, and purely associative effects were combined as one. It is exceptionally difficult to know what the resulting meta-analyzed effect size is an estimate of, i.e. what the estimand is, and whether it applies to the type of effect that a researcher may wish to observe in their own future work. For example, to what degree is the interaction effect between a depression IRAP’s trial types and high vs. low experiential-avoidance group informative to a separate study on the correlation between self-reported self-esteem and a self-esteem IRAP in a prisoner population? Even if all implementational issues with the original analysis were fixed, I would argue that this ‘apples and oranges’ approach to pooling effect sizes fundamentally undermines the interpretability and validity of the results and leads to misleading conclusions. Nonetheless, the point of the verifications presented here is to highlight that erroneous analytic approach reported in Vahey et al. (2015) was also erroneously implemented.”*

**Reviewer 2 point 2**

I do have a number of comments that I think will improve the article.

1) Bias  
I think it is clear the author has been frustrated about this meta-analysis, the lack of cooperation from the original authors, and the large amount of work needed to correct the scientific record. At some places, this leads to statements that are unnecessarily subjective, speculative, or personal. I will highlight these below, and I think the article will improve if these sentences are changed. Currently, the are largely separate of the main contributions of the paper anyway, and they distract. Although it might be understandable that affect influences scientists, we should work as hard as we can to not let it influence our writing. I have often been on the receiving end of this feedback myself, as I also have a tendency to let affect influence my writing when I am passionate about something - but this is exactly why we have peer review, to provide an extra check.  
The authors also accuses the original authors of 'cherry-picking'. This might well have happened, and it looks like it, but a more neutral description is possible 'it seems the authors selectively included larger estimates' for example. This is a description on the level of the data, not on the level of the intentions of the authors.

*Author response: Thank you for checking me on this – I agree it is important to remove any language that distracts from the scientific critique here. To this end:*

*- I have removed all references to cherry picking and replaced them with discussion of the potential for “selection bias”.*

*- I have rephrased the manuscript throughout to talk about the results presented in Vahey et al. (2015) (the article) rather than those presented by Vahey and colleagues (the authors). For example, all instances of “Vahey et al.’s (2015) results” have been changed to “the results presented in Vahey et al. (2015)”.*

*- I have revised the language throughout, keeping an eye on language that could be considered personal, speculative, or subjective, lest it detract from the evidence.*

**Reviewer 2 point 3**

The sentence "a measure which was also created by the last author of Vahey et al. (2015)." can be removed.

*Author response: Removed.*

**Reviewer 2 point 4**

2) Clarify the interactions with the original authors  
There is some information in the paper about not getting code from the authors, but also that there has been some interaction about the code used, and links provided to the code. I would suggest to have a dedicated header for all this information, and be much clearer about what has been provided by the authors. If the author responsible for the data analysis replied 'sorry, I no longer work in academia, I did not store the code, but you should be able to take numbers in the papers and run these 2 scripts we used' that is much clearer for me (and fairer to the original authors) that ' the authors did not share their code' . Of course you need to maintain confidentiality, but I feel there is space for a more objective summary of which information was received, and why. Here, it is useful to be clear about the details. For example, if a researcher has the code, but actually declines to share it, you should probably report them to their scientific integrity officer - that is clearly a violation of research integrity. But if 'decline' means ' I do not have the code, but you should be able to use these files and I do not want to recreate the code myself' that is more acceptable.

*Author response: Thank you for this prompt to make these details clearer. I have collated information that was present in the previous version of the manuscript under a new heading, “Correspondence and source of original code”, that comes immediately after the Transparency Statement on page 8. The content of this section was moved from the Transparency Statement and the “Implementation of the meta-analysis” heading.*

*In practice, it is not clear cut whether Vahey declined to share scripts that he had. He stated that he probably had them but I shouldn’t need them as I could go to Field’s website; he said that they were probably on a different laptop or harddrive and it would take too long to dig them out; he later promised twice to share them but said he was moving between jobs, and eventually stopped responding. Functionally, I would say this amounts to refusing to share, but it wasn’t a simple “no”. Dr. Vahey still works in academia. I have now contacted his Research Integrity office about this.*

**Reviewer 2 point 5**

Somewhat related, if the authors ran multiple scripts, I do not really think that is a problem. The reproduction stresses if any single script can repeat all numbers - but if they ran multiple and combined results, what is the problem exactly?

*Author response: Thank you for the opportunity to clarify this important point. Simply put, the point here is that Vahey et al. (2015) stated that they used a given set of scripts, but the evidence shows that they did not. The methods reported in Vahey et al. (2015) are therefore inaccuate and its results are not reproducible, at minimum.*

*To expand on this point: the issue here is not whether a single script can reproduce all numbers, it is that the Dr. Vahey states that he employed the scripts associated with Field and Gillett (2010), but (a) those two scripts implement different meta-analytic methods, one of which corresponds on to the method Vahey et al. (2015) state was used and one which does not; and (b) nether script nor some combination of their code allows third parties to reproduce the results of Vahey et al. (2015). The point here is therefore whether the methods of Vahey et al. (2015) are accurately described, and the results demonstrate that the answer is no. Scrutinising both scripts, even the one that should be wrong, and combinations of code between them, was necessary to demonstrate this fact while trying to be generous to the original authors (e.g., in case they merely mislabelled the method and script employed, rather than there being much more serious issues of reproducibility).*

*In order to clarify the point of this section further, I have modified the “Implementation of the meta-analyses” section on page 14, including adding an explicit statement that:*

“The point here is to highlight my best efforts to try to reproduce the results in Vahey et al. (2015) using the tutorial scripts Dr. Vahey reported using, and the complications that not having direct access to the original authors’ code presented.**”**

**Reviewer 2 point 6**

3. Verification by including errors  
  
I have a big problem with one approach in this verification report, and I believe it should be completely removed from this manuscript before the manuscript can be accepted. The author at one point introduces intentional errors, and then concludes these bring the results from the reproduction closer to the reported results. The results are still very different from the reported results. So, it is not like the author has identified errors that were actually made. The fact that introducing errors gets the results closer to reported results is in no way any evidence that errors were made, but it is extremely suggestive of the idea that the original authors made errors. Of course there is something weird that happened with the credible intervals. But it is just as possible that they ran a script they should not have, or entered a parameter, changed a setting, or something like this, and got the results. In other words, there is nothing interesting in the fact that the reproduction with intentional errors is closer to the reported results. If it had replicated the original results, sure. But now if is insinuating erroneous analyses that are not supported. The last paragraph on page 26 should be removed, the section on page 27, and verification attempt 7 should be removed.  
  
*Author response: I reported these results for transparency on the basis that I had run them, but as you say the results are unclear. I am happy to remove them for the reasons you suggest. Verification attempt 7 has been removed, as have the paragraphs you mention. The abstract and discussion have both been appropriately adjusted to remove references to this speculation.*

*e.g., the following was cut from the abstract:*

*“*The reproduction attempt with the closest compressive set of results required making two serious errors: using the wrong data set and mislabelling Confidence Intervals as Credibility Intervals and vice versa.”

**Reviewer 2 point 7**

Minor points  
  
The bottom of verification attempt 1 abruptly stops after (See - there is no attempt 2. But then there is an attempt 3. Almost as if part is missing. It looks very sloppy.

*Author response: Thank you for catching this mistake, which was introduced when converting the file from preprint format to APA format. The end of attempt 1 + attempt 2 now reads (p.22):*

“This verification attempt did not reproduce the original results for the point estimate, Confidence Interval, or Credibility Interval (see Table 3).

**Verification attempt 2**

The second verification attempt employed Field’s “Meta\_Basic\_r.sps” script, which implements a Hedges’ style “basic” meta-analysis. I was unable to get this script to run in SPSS. It makes use of commands such as nrow(), csum(), sd(), and t(), which are apparently not SPSS commands. R does have similarly named functions, but the script employs these commands within lines of SPSS syntax. It does not appear that these commands were simply deprecated between versions of SPSS. It is unclear how these apparent errors in Field’s script have apparently not been publicly detected or corrected given they are still distributed on Field’s website and Field & Gillett (2010) continues to be cited (>1,200 citations at time of writing).”

**Reviewer 2 point 8**

There are some spelling errors throughout the paper, and some sloppy sentences copy pasted with hard returns, and differences in font.  
very good ( = 0.90). This would imply that > sentence should continue  
Hunter and Schimdt style  
Welches

*Author response: Thank you for catching these mistakes, some of which were introduced when converting the file from preprint format to APA format. I have corrected them all and given the manuscript careful proof readings.*

**Reviewer 2 point 9**

The 'assessment of inclusions' header should make clear a whole new meta-analysis is performed! This was a lot of work - highlight it.

*Author response: Thank you - I have explicated this. Page 30 now reads:*

“I therefore re-examined the same 15 articles included in Vahey et al. (2015) and searched for other effect sizes that met the original inclusion criteria. Note that this does not represent an endorsement of those criteria, it was merely an assessment of the correct application of the original criteria. There data were later used to conduct a new meta-analysis. As I discuss later, this too does not represent an endorsement of its results (e.g., as a valid estimate of the IRAP’s criterion validity), but rather it was conducted to illustrate the compound impact of the errors on the final results reported in Vahey et al. (2015).”

**Reviewer 2 point 10**

Please do not have paragraphs of a single sentence.

*Author response: Corrected throughout.*

**Reviewer 2 point 11**

It remains totally unclear > just unclear without totally is fine and sounds less biased.

*Author response: Corrected.*

**Reviewer 2 point 12**

Another minor point is that the N = 34 vs 36 difference could be due to rounding. Sample sizes should always be rounded up (we can not collect .5 of a participant) but not all software does this - maybe the 34 is incorrectly rounded down.

*Author response: Thank you for this thoughtful point. The reconstructed values were calculated not by rounding at .5 but by calculating the ceiling values. The N per group was calculated and then doubled: power analysis returns 16.94, ceilinged to 17, X2 to find total sample = 34. As such, the reproduced result does not appear to be erroneous due to the reasons you suggest. I’m glad to be checked on a fine grain detail such as this though.*

**Reviewer 2 point 13**

I found the following sentence to add nothing - please try to convey it, or remove it : Trying to unravel this was extremely challenging, to a degree that is difficult to convey.

*Author response: Removed.*

**Reviewer 2 point 14**

Why were the bias detection tests not reproduced? This is just a few lines of code in R, right?

*Author response: Unfortunately it’s not as simple as this. The original SPSS scripts by Field call from within them R scripts that calculate some of these bias detection methods but not others. Those sections of the SPSS scripts and/or R scripts are no longer functional, so reproducing the results would first require fixing Field’s code, once again expanding the scope of the project. When paired with my critique that, as the authors of many of the included articles the authors have direct knowledge of publication bias (which you mention in your next point), this reduced the effort to reward ratio for attempting to also reproduce the risk of bias tests.*

**Reviewer 2 point 15**

I very much like the statement that the original authors should have had insights about their own file drawer, and included this in their meta analysis!

*Author response: thank you for this positive feedback – I wish I saw it more often in meta-analyses conducted on original research by the same authors.*

**Reviewer 2 point 16**

The author writes "This also brings the average criterion association observed for the IRAP closely in line with the average correlation observed across social and personality psychology (i.e., around r = .2: Hemphill, 2003; Hussey, 2023; Richard et al.,2003). > But these estimates are not really comparable, right. This is also an 'apples vs oranges' comparison.

*Author response: Hemphill’s estimate is itself a highly heterogeneous set of results from across an entire field – very much apples to oranges, but nonetheless informative to the reader regarding normative distributions at a field level. I thought that this reference point might be useful, following Cohen’s (1988) recommendation that we only use rule of thumb cutoffs when normative data isn’t available. But I hear that you and readers might feel otherwise, and so I have removed this.*