ERROR REPORT #3

Report by Ian Hussey

Study Title: A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain

Journal: Journal of Behavior Therapy and Experimental Psychiatry

DOI: 10.1016/j.jbtep.2015.01.004

Citations: 152

Summary: I have detected a number of apparent errors in this article (Vahey, Nicholson, and Barnes-Holmes 2015). I provide a full report, including data and code, in a published article at this link: sciencedirect.com/science/article/pii/S0005791624000740. I originally sent my report to the editor with request they initiate an integrity investigation following COPE guidelines, which they declined to do. The editor elected to peer review and publish the report, and has not contested any of my findings, but has so far declined to initiate any correction or retraction of the original article.

This report serves as a brief overview of some of the more salient issues found in the paper. Vahey, Nicholson, and Barnes-Holmes (2015) appears to need substantial correction at minimum. In particular, researchers should not rely on its results for sample size justification. A list of suggestions for error detection in meta-analyses is provided. A reply by Vahey and colleagues (10.1016/j.jbtep.2024.102016) does not substantively solve the issues raised by my report.

Table of contents

Mathematically implausible meta-analytic effect size

The meta-effect size is reported as $\bar{r}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67]. These confidence intervals are asymmetric around the mean: [-.05, +.09]. This cannot be explained by data transformations such as Fisher's r-to-z transformation, which would produce intervals with a skew in the opposite direction. I cannot think of a method that can generate a point estimate with intervals like this.

Data is internally incompatible

The weighted average effect size estimates in the forest plot (figure 1, Vahey, Nicholson, and Barnes-Holmes 2015) don't match those in the funnel plot (figure 2, Vahey, Nicholson, and Barnes-Holmes 2015). At least one data point seems to differ between the two plots.

Two of the 15 weighted average effect sizes reported in the forest plot cannot be reproduced from the individual effect sizes reported in the supplementary materials (i.e., by weighting by df, as stated in the manuscript).

Inappropriate effect size conversion

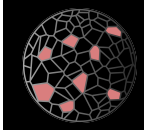
At least some of the effect sizes reported in the original articles were partial eta-squared (η_p^2) estimates. The authors report converting these to Pearson's r values using the formula for the conversion of eta-squared (η^2). This erroneous, as η_p^2 and η^2 are different effect sizes.

Not computationally reproducible

I attempted to reproduce the results using the effect sizes reported in the article and supplementary materials, and the scripts provided by Field and Gillett (2010) that are cited in Vahey, Nicholson, and Barnes-Holmes (2015). I tried many different approaches and variations but cannot reproduce the reported results. In particular, all attempts produced Credibility Intervals with a width of zero, [.47, .47], and most attempts produced a different mean effect size of r = .47.

Incorrect inclusions

The manuscript states that this is a meta-analysis of criterion validity, i.e., between the Implicit Relational Assessment Procedure (IRAP) and other clinically relevant criterion variables. However, 23 of the 56 individual effect sizes reported in the supplementary materials do not involve a variable other than the IRAP, ie they are taken from one sample t-tests that quantify the size of the IRAP effect, which is related to the different in reaction times between consistent and inconsistent blocks. These cannot provide evidence of the criterion validity of the IRAP, just as the magnitude of a Stroop effect or the mean score on a questionnaire cannot tell you about the criterion validity of those measures.



Incorrect omissions

I examined the same articles included in the meta-analysis for other effect sizes meeting the stated inclusion criteria. I found 308 additional effect sizes meeting criteria. These effect sizes were generally smaller than the included ones. An updated meta-analysis provides an effect size less than half of the size of the original: $\bar{r}_{updated} = .22$, 95% CI [.15, .29], 95% CR [.22, .22], 95% PI [-.01, .42]; $\bar{r}_{original}$ = .45, 95% CI [.40, .54], 95% CR [.23, .67].

Conclusion

The results of Vahey, Nicholson, and Barnes-Holmes (2015) were found to have poor reproducibility at almost every stage of the analytic strategy. In aggregate, these seriously undermine the credibility and utility of the conclusions and recommendations of the origin. Recalculated results suggested that the compound impact of the errors reduced the meta-effect size to less than half the original result ($\bar{r} = .22$ vs. $\bar{r} = .45$) and increased the sample size recommendations by more than 15 times the original results (minimum N = 37 vs. 346). Vahey, Nicholson, and Barnes-Holmes (2015) therefore requires substantial correction at minimum, and researchers should not use it for sample size planning. The editor of the journal has declined to initiate such a correction or retraction of the original article, apparently in violation of COPE guidelines.

References

Field, Andy P, and Raphael Gillett. 2010. "How to Do a Meta-Analysis." *British Journal of Mathematical and Statistical Psychology* 63 (3): 665–94.

Vahey, Nigel A, Emma Nicholson, and Dermot Barnes-Holmes. 2015. "A Meta-Analysis of Criterion Effects for the Implicit Relational Assessment Procedure (IRAP) in the Clinical Domain." *Journal of Behavior Therapy and Experimental Psychiatry* 48: 59–65.