A critical reanalysis of Vahey et al. (2015) “A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain”

Ian Hussey

Ghent University

Author Note

[Include any grant/funding information and a complete correspondence address.]

Abstract

Vahey et al.,

Keywords: [Click here to add keywords.]

A critical reanalysis of Vahey et al. (2015) “A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain”

Indirect measures of implicit attitudes have seen wide use in psychopathology research over the last twenty years (e.g., Roefs et al., 2011). Unlike direct measures, such as self-reports, these infer individuals’ attitudes through reaction time biases, misattributions, and other forms of automatic behaviour (De Houwer & Moors, 2010; although see Corneille & Hütter, 2020). These measures have led to important insights and useful predictions, such the role of self-esteem within depression (Gemar et al., 2001; Hussey & Barnes-Holmes, 2012; Remue et al., 2013) and the prospective prediction of repeat suicide attempts (Nock et al., 2010; Tello et al., 2020).

A meta-analysis of one implicit measure, the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2010), concluded that it possesses good criterion validity and has potential within clinical assessment (Vahey et al., 2015). In Vahey et al. (2015), the authors (a) provided an estimate of the association between IRAP effects and clinically-relevant criterion variables, (b) reported that the IRAP compares favorably to other a more popular implicit measure, the Implicit Association Test (Greenwald et al., 1998), and (c) used the estimate of effect size to conduct power analyses and make sample size recommendations for future IRAP research.

Vahey et al. (2015) has been well-cited and used to guide subsequent work: at time of writing, it has been cited 93 times, with roughly 20% of articles citing it to justify sample size decisions. IRAP studies have typically involved small sample sizes, many around 40 participants, which roughly correspond to Vahey et al.’s sample size recommendations of at least 37 participants to [have 80% power to detect a [medium?] correlation with external criterion variables]. McEnteggart (2015) provides a particularly clear characterization of the importance of Vahey et al.’s (2015) results for the broader IRAP literature: “the *N*s involved in [IRAP] studies … are often relatively small. ... Indeed, it could be argued that this impacts upon on the credibility of IRAP research. However, in a recent meta-analysis of IRAP studies, it was reported that even small *N* IRAP studies have sufficient statistical power.” (p.XX). Given that past and future work continues to rely on the conclusions of Vahey et al.’s (2015) meta-analysis, it is therefore important that its results are accurate and reliable.

Elsewhere, three meta-analyses have all concluded that the IRAP’s reliability that is poor at best (e.g., internal consistency alpha = .53, test-retest reliability ICC = .18: Hussey & Drake, 2020 REF; see also (Golijani-Moghaddam et al., 2013; Greenwald & Lai, 2020). Given that a degree of reliability is necessary for validity and utility, the conclusions of these meta-analyses of reliability versus criterion validity would seem to be at odds with one another. This serves as a second source of motivation to assess whether the results of Vahey et al. (2015) are indeed reliable.

In the wider psychology literature, the concepts of reproducibility and replicability have come to recent prominence as part of what has been called the Replicability Crisis (REF). This began in the field of social psychology but discourse around it has now also spread to clinical psychology (REF). Large-scale efforts to reproduce and replicate psychological research have demonstrated that the published literature has both a high rate of misreported results (REF) and that results frequently do not replicate when experiments are repeated with high fidelity (REF).

Concerns about reproducibility have been raised about not only original research articles but also the results of meta-analyses. Lakens et al. (2017) recently demonstrated that the results of the majority of a random sample of meta-analyses published in psychology cannot be reproduced [in what way?]. Maassen et al. (2020) found that almost half of effect-sizes reported in meta-analyses of psychology research could not be reproduced. This was attributed to due to a variety of issues such as errors in the extraction of effect sizes from original studies, insufficient details regarding data processing and transformation of effect sizes, insufficient details of the specific meta-analytic approach employed, or failures to adhere to meta-analysis reporting guidelines.

In this article, I therefore sought to assess the reproducibility of Vahey et al.’s (2015) results. In the first section of this article I provide details of the original meta-analysis. In the second section I assess its reproducibility in multiple ways: its power analyses, meta-analysis results, the calculation of weighted-mean effect sizes, and the extraction and conversion of individual effect sizes from original articles. Based on issues and errors that were detected in the second section, in the third section I conduct a new meta-analysis and power analyses to make sample size recommendations. All data and code to reproduce my analyses can be found on the Open Science Framework (osf.io/XXXX).

# Overview of original meta-analysis

Vahey et al. (2015) stated that the purpose of their meta-analysis was to “quantify how much IRAP effects from clinically-relevant responding co-vary with corresponding clinically-relevant criterion variables” (p.60). To this end, the authors conducted a non-systematic review of the available literature at the time, finding 46 empirical articles that employed the IRAP. Their inclusion criteria were that “the IRAP and criterion variables must have been deemed to target some aspect of a condition included in a major psychiatric diagnostic scheme such as the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, 2013) … The authors decided whether the responses measured by a given IRAP trial-type should co-vary with a specific criterion variable by consulting the relevant empirical literature.” (p.60). The authors extracted 56 effect sizes that met these inclusion criteria from 15 of the articles, as determined by two independent reviewers. One or more authors of the original meta-analysis was also an author of 12 of the 15 articles (80.0%) from which effect sizes were extracted, indicating that the authors of the original meta-analysis were familiar with the research they were meta-analysing. These extracted effect sizes were provided in Vahey et al.’s Supplementary Materials.

The results of the original meta-analysis could not be reproduced based on the details included in the original manuscript and its supplementary materials. The first author of the original meta-analysis was also contacted but declined to provide their analysis scripts or other research materials. Efforts to reproduce the original results are outlined below, along with details of any necessary approximations, in order to highlight the features that hindered reproducibility.

# Reproducibility of Vahey et al. (2015)

In this section, I work backwards through Vahey et al.’s (2015) results and data and attempt to reproduce each step. While Vahey et al. (2015) did share their meta-analysed effect sizes in supplementary online materials, they did not share their analysis code or scripts. Prior to attempting to reproduce Vahey et al.’s (2015) data and results, I contracted the first author and requested copies of their data, code and/or analysis scripts. However, he declined to share these materials. At times, the details reported in Vahey et al. (2015) and their supplementary materials were not sufficient to reproduce their data and results. This difficulty has been noted in previous publications which attempted to reproduce the results of published meta-analyses (e.g., REF). Where details were not sufficient to reproduce the original data and results, reasonable guesses were made. At times, this involved exploring more than one method, and settling on the closest approximation found.

The first author of the original meta-analysis was contacted again before this article was either submitted for publication or to a preprint server. The draft manuscript, along with all data and code was supplied to them, and they were encouraged to find possible mistakes in my reanalysis. [comment on whether Vahey et al. (2015) responded or found any issues]

## Power analyses

Vahey et al.’s reported meta-analysis effect size estimate was *r* = .45, 95% CI [.40, .54], 95% CR [.23, .67]. Using this effect size, they conducted power analyses for sample size planning. They reported that, to detect a zero order correlation with 80% power, 29 participants would be required when using the meta-analytic effect size, or 37 if using the lower bound of the CI (alpha = .05, one-tailed; following recommendations by Perugini, Gallucci, & Costantini, 2014). I used the R package pwr (REF) to reproduce these sample size recommendations, which were found to be computationally reproducible.

However, Vahey et al.’s (2015) choice of parameters for these power analyses could be questioned: one-tailed correlation with alpha = .05 are very uncommon in the literature, and regression analyses require two-sided testing. A two-tailed test with alpha = .05 would therefore correspond more closely to modal research practices. I therefore recomputed sample size estimates using these parameters: using the meta-analytic effect size (*r* = .45), 36 participants would be required (80% power, alpha = .05, two-tailed). When using the lower bound of the confidence interval (*r* = .40), 46 participants would be required. These suggested sample sizes were therefore 24% higher than those reported in Vahey et al. (2015).

## Meta-analytic effect size

Vahey et al.’s (2015) aforementioned power analyses relied on the accuracy of their meta-analytic effect size. I attempted to computationally reproduce the meta-analytic effect size from the weighted-mean effect sizes, 95% Confidence Intervals and sample sizes reported in Vahey et al.’s (2105) forest plot (p.XX). 95% Confidence Intervals were extracted from the forest plot and converted to variances using the following formula:

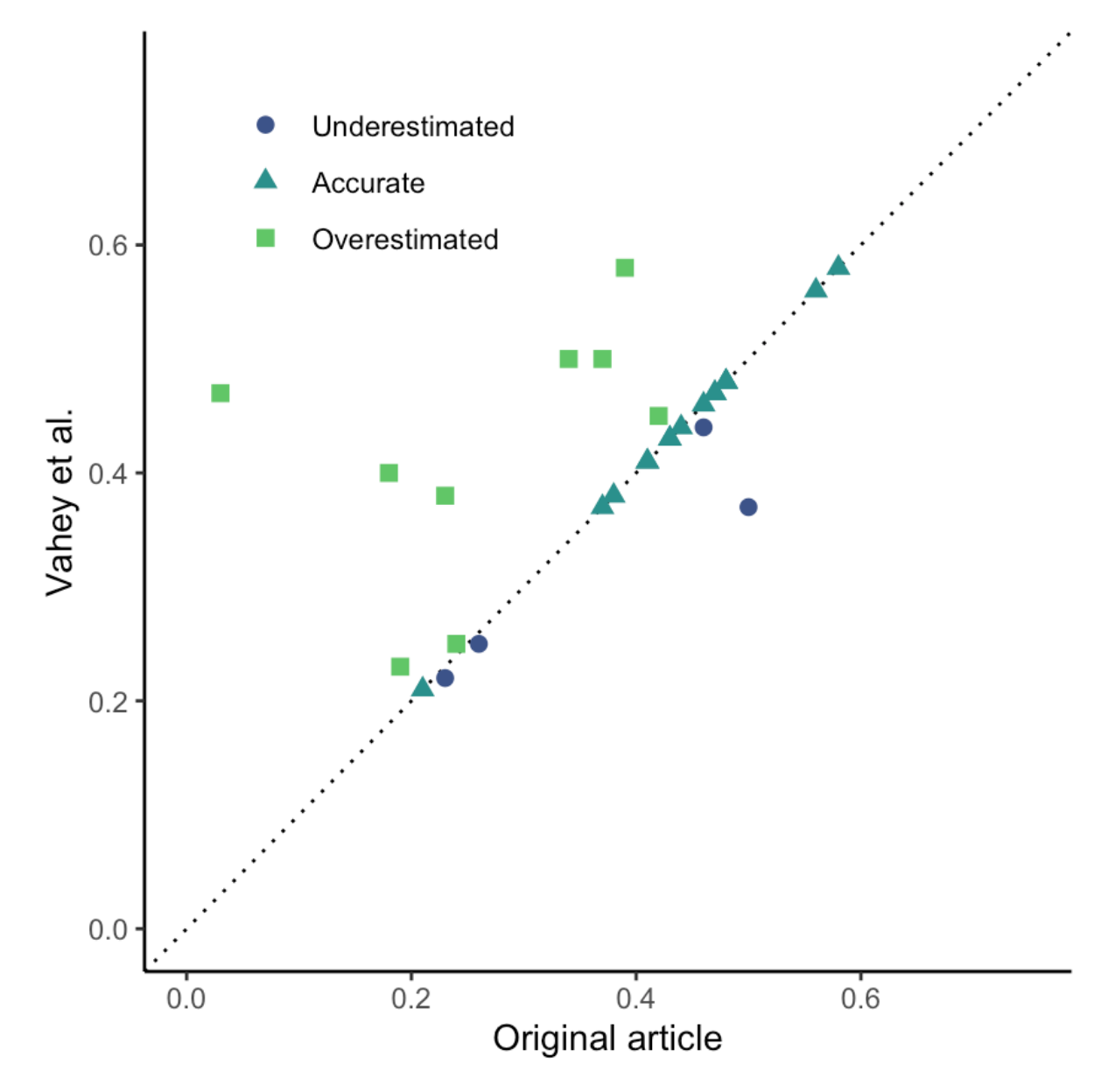
Vahey et al. (2105) reported employing a Hunter and Schmidt style meta-analysis. The meta-analysis model was implemented using the R package metafor (REF). The Hunter & Schmidt estimator was used, and the effect sizes were weighted by sample size, following the Hunter & Schmidt method.

Results demonstrated a meta-analytic effect size of *r* = .47, 95% CI [.40, .54], 95% CR [.40, .54], *p* < .0001. No heterogeneity was observed, *Q*(*df* = 14) = 7.05, *p* = .933, 𝜏2 = 0.0, *I*2 = 0.0, *H*2 = 1.0. Vahey et al.’s (2015) meta-analysis results could therefore not be precisely computationally reproduced using the data they reported in their forest plot and their descriptions of their analytic approach. Estimate of the meta-analytic effect size differed by only a small amount (Δ*r* = .02). Confidence intervals were identical widths (i.e., were reproduced). However, credibility intervals and therefore estimated heterogeneity differed by a large amount (large heterogeneity in Vahey et al. 2015, no heterogeneity in my reanalysis).

## [old text]

## Data extractions

In order to assess the reproducibility of the data extraction, effect sizes were re-extracted from the original articles and converted to Pearson’s *r* values and compared to those reported in the original meta-analysis’s supplementary materials. Wherever possible, the same effect size conversion method was employed as in the original meta-analysis, following those listed in their supplementary materials. However, while the names of the methods were listed, specific formulae or software implementations were not provided. In order to maximise the reproducibility of the re-extractions, all formulae, tools and workings used here are provided in the Supplementary Materials, and all extractions and conversions were performed independently by two researchers, cross checked, and then double checked. When rounding all effect sizes to two decimal places, inconsistencies were found between Vahey et al.’s extractions and the re-extractions here in nearly half of all cases (44.8%). A bias was found in the directionality of inconsistencies: 13.7% of effect sizes were larger in the re-extractions whereas 31.0% were larger in Vahey et al.’s extractions. When pairwise differences between our were calculated some were found to differ by large amounts (range -.44 to .13, see Figure 1). The inconsistencies between the effect sizes reported in Vahey et al. and the re-extractions here were deemed sufficiently important by themselves to raise important questions about the reproducibility of the results of the meta-analyses, separate from the other aspects of reproducibility discussed below. The re-extracted values are fitted to a new meta-analysis



*Figure 1.* Inconsistencies in the effect sizes reported in Vahey et al.’s (2015) meta-analysis vs. those reported in the original research articles.

## Data processing

Vahey et al. noted (pXX) that their meta-analytic model employed converting the effect sizes reported in the original articles to Pearson’s *r* correlations. In the majority of articles, multiple effect sizes were reported between a given IRAP and multiple criterion variables. Vahey et al. noted that this non-independence of data was dealt with by calculating mean correlations weighted by the degrees of freedom of the original test. In order to assess their computational reproducibility, the correlations and degrees of freedom reported in the supplementary materials were used to calculated weighted means. These weighted means were then compared to those reported in the Vahey et al.’s forest plot (p XX). Inconsistencies were found for 2 of 15 (13%) of the weighted mean effect sizes (discrepancies ranged from -0.02 to 0.05).

In order to perform a meta-analysis, estimates of associated error are required in addition to the effect sizes themselves. However, the Supplementary Materials did not contain data on confidence intervals or standard errors associated with each effect size. As such, the analyses could not be reconstructed on the basis of the supplementary data alone. In lieu of this, confidence intervals for the mean effect sizes weighted by degrees of freedom were extracted from the forest plot reported in the article, and then converted to Standard Errors using the formula SEM = (ci\_upper – ci\_lower)/1.96\*2. This method was an imperfect approximation given that the confidence intervals reported in the forest plot were asymmetric, and no description was provided in the manuscript of how confidence intervals or standard errors were extracted from the original articles. The numeric results reported in the forest plot were also compared against estimations of the values displayed in the plot. No discrepancies were found in either the estimates or the confidence intervals.

## Meta-analytic strategy

XX

## Results of meta-analysis

The specific analytic strategy was then reconstructed from its descriptions in the text of the original meta-analysis. Vahey et al. describe applying “Hunter and Schmidt's (2000) random effects model … because it readily allows the calculation of credibility intervals to qualify its resultant meta-effects, and because it is a well established and versatile method (see Field & Gillett, 2010).” (p.XX). The authors discuss applying a specific feature of a Hunter and Schmidt style meta-analysis: the weighting the effect sizes by sample size. That is, mean effect sizes weighted by degrees of freedom were calculated to combined multiple effect sizes within a given study, and then these weighted means were then weighted by sample size within the meta-analysis. While the degrees of freedom used were reported in supplementary materials, it was less clear how the samples sizes used for weightings and reported in the forest plot were obtained, given that the individual effect sizes that were converted to mean effect sizes were in many cases calculated from different sample sizes, yet the reported sample sizes were even numbers. The sample sizes reported in the forest plot were employed here as an approximation. No discussion of other features of Hunter and Schmidt’s approach to meta-analysis were made in the text of the original meta-analysis, and so were not employed here (e.g., the deattentuation of correlations by the tasks’ internal consistency estimates; see REF).

All analyses in the current paper were conducted in R (REF) using the packages metafor (REF), tidyverse (REF), pwr (REF), and irr (REF). The mean effect sizes weighted by degrees of freedom and their associated standard errors that were extracted from the original meta-analysis’s forest plot were then subjected to a Random Effects meta-analysis model that employed the Hunter and Schmidt method (i.e., using the Hunter and Schmidt estimator function and weighted by sample sizes taken from the forest plot, see Supplementary Materials for all data and code). Results suggested a meta effect of *r* = 0.47, 95% CI [0.37, 0.57], 95% CR [0.37, 0.57]. Results were therefore broadly similar but not identical to those reported by Vahey et al.: *r* = .45, 95% CI [.40, .54], 95% CR [.23, .67]. In particular, the current results provide poorer estimation compared to the original study (CI width = .14 vs. .20), and diverge regarding the degree of heterogeneity between the analyses (i.e., confidence and credibility intervals are identical in the current analyses but were different in Vahey et al.’s results).

## Results of power analysis

As well as reporting estimates of effect size, Vahey et al. used their meta-analysis results to conduct *a priori* power analyses to inform sample size decisions in future research. Vahey conducted two sets of power calculations, one using their effect estimate and one using the lower bound of the estimate’s confidence interval, on the basis that this was recently argued to provide a more conservative method of sample size planning (REF). The results reported in their original paper were computationally reproducible using the R package pwr: to detect a one-sided zero order correlation with 80% power when alpha = .05 the minimum sample size is 29 participants (if using the estimate: *r* = .45) or 37 (if using the lower bound of the estimate’s CI: *r* = .40). However, it should be noted that one-tailed hypotheses for correlations are uncommon in the broader psychological literature (i.e., they equate a default two-tailed test with alpha = .10), and indeed in the subsequent research that cites the Vahey et al. meta-analysis for the purpose of sample size determination. While it is of course not the responsibility of the authors of the original meta-analysis if subsequent research misinterprets their findings, it would arguably have been more informative or intuitive to report power analyses for the two-tailed tests that are more likely to be used in subsequent research. Applying these suggested a minimum sample size of 36 (using the estimate: *r* = .45) or 46 (using the estimate’s lower bound CI: *r* = .40); 24% higher than when using one-sided tests.

These power analyses were then updated in light of the results of the reproduced meta-analysis: to detect a zero order correlation with 80% power when alpha = .05 (two-sided) the minimum sample size was 33 participants (if using the estimate: *r* = .47) or 55 (if using the lower bound of the estimate’s CI: *r* = .37). This represents sample size recommendations that are 50% larger than recommended by Vahey et al.

# Reanalyses

## Updated meta-analysis

Given the non-trivial differences between the effect sizes extracted by Vahey et al. and those re-extracted here, a new meta-analysis model was fit using the re-extracted effect sizes. Additionally, recent results from simulation studies suggests that the method employed by Vahey et al to deal with non-independence of effect sizes estimates (i.e., averaging across them with weightings) provides poorer statistical power than the alternative approach of employing a multi-level meta-analysis model. I therefore elected to employ a multi-level random effect meta-analysis, with random intercepts for study, without weightings (i.e., the default recommended), and using the Restricted Maximum Liklihood estimator function (n.b. Hunter & Schmidt style estimators are not possible for multi-level meta-analyses within the metafor package, therefore the recommended default estimator was employed). A meta-analyzed effect size of *r* = 0.38, 95% CI [0.32, 0.44], 95% CR [0.32, 0.44] was found. No evidence of heterogeneity was detected, *Q*(df = 28) = 12.09, *p* > 0.99, 𝜏2 = 0.00. Results were therefore significantly smaller than those reported by Vahey et al. (*r* = .45, 95% CI [.40, .54]). Results were also estimated more precisely than in the original meta-analysis (CI width = .14 vs. .12).

Updated power analyses using these estimates suggested that, to detect a zero order correlation with 80% power when alpha = .05 (two sided), the minimum sample size was 52 participants (using the estimate) or 74 (if using the lower bound of the estimate’s CI). This represents a required sample size that is two times larger than recommended by Vahey et al.

## Alternative meta-analysis

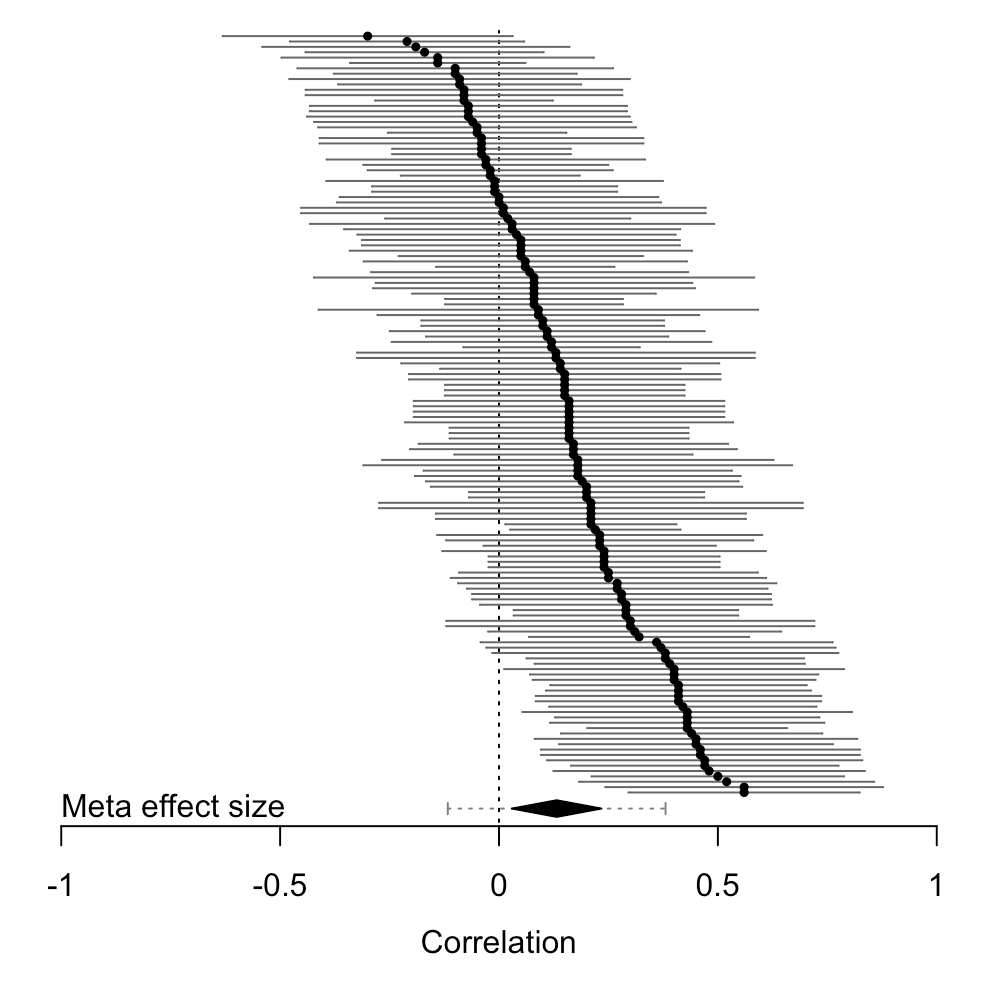
Vahey et al.’s extractions were incorrect, but also his choices for what to include or not were also highly questionable.

* No mention of how many effect sizes were considered or rejected.
* Questionable omissions and blinding. Examples.
* Significance from zero effects
* IRAP as the DV
* Retrospective *a priori* predictions
* Inclusion of effects that do not meet the inclusion criterion of clinical relevance.

A new meta-analysis was therefore conducted. Vahey et al. extracted 56 effect sizes from 15 articles, but provided no information about the number of effects that were not included or details of these excluded effects. I re-extracted all effect sizes reported in these 15 articles, resulting in 334 effect sizes. Some additional effect sizes were found that were non-independent with the extracted ones (e.g., follow-up *t* tests after ANOVA, correlations with the overall IRAP score when its component trial types were also correlated, or correlations with a scale’s sum score when its subscale sum scores were also available). Two independent raters then rated each effect (both the IRAP domain and the criterion) for clinical relevance using Vahey’s definition. No exclusions were made on the basis of ‘retrospective a priori predictions’ on the basis that I strongly disagree that this is a meaningful classification effects in terms of its experimental replicability or its measurement reliability or validity. If either rater rated the effect as clinically relevant it was included in the meta-analysis. Agreement was found in 90% of cases (Cohen’s Kappa = 0.88, *p* < .0001).

In addition, results from types of analyses that were defined a priori as producing problematic or misleading results or inferences were also excluded. This included ……

After excluding effects that were rated as not being clinically relevant or which were based on analyses that were determined a priori to be problematic, 144 effect sizes remained for inclusion in the meta-analysis. The same choice of multi-level meta-analysis model was again employed. Given the large number of effect sizes being meta-analyzed, results are illustrated using a Caterpillar plot rather than a Forest plot (i.e., no article labels are included and effects are sorted by size; see Figure 2).



*Figure 2.* Caterpillar plot of the effect sizes and meta effect size estimate for the new meta-analysis.

Results demonstrated a meta effect size *r* = .20, 95% CI [.12, .29], 95% CR [-.04, .44], *p* = .000005. Evidence of heterogeneity was found, *Q*(df = 141) = 195.21, *p* = .0017, 𝜏2 < 0.00. Based on the non-overlap of their confidence intervals, this estimate is significantly smaller than the effect size reported in the original meta-analysis (i.e., *r* = .45, 95% CI [.40, .54]).

As in the original meta-analysis, this estimate of effect size was used to calculate a power analysis for future sample size planning. To detect a zero order correlation with 80% power when alpha = .05 (two-sided), the minimum sample size was 194 participants (using the estimate) or 542 (using the lower bound of the estimate’s confidence interval). This represents a required sample size that is nearly fifteen times larger than recommended by Vahey et al. According to the systematic review (see Supplementary Materials), both of these estimates are more than ten times larger than the mean sample sizes employed in IRAP research to date.

In order to explore the potential bias for bias within the original research studies based on whether the creator of the task was included as a co-author, a moderator meta-analysis was then conducted. Each study was first coded dichotomously based on whether it included “Barnes-Holmes” among its authors. Results demonstrated that effect sizes were significantly larger when the creator of the IRAP was listed among a manuscript’s authors, *r* = .25, 95% CI [.09, .41], than when this was not the case, *r* = .08, 95% CI [-.05, .22], *Q*(df = 1) = 4.06, *p* = .044.

# Discussion

## Summary of findings

The meta-analysis reported by Vahey et al. was found to have poor reproducibility on multiple fronts. Nearly half of the effect sizes included in the original meta-analysis did not match those reextracted from the original articles. In one third of cases, the effect sizes used in the original meta-analysis were biased upwards relative to the re-extractions done here. Data processing was found to not be reproducible, with 13% of cases demonstrating disagreement between the weighted average effect sizes reported in the forest plot and those recalculated from the effect sizes reported in the supplementary materials. The specifics of the meta-analytic strategy were not completely described in the text. Unfortunately, requests made to the first author of the original meta-analysis for the original data and code were refused. When the data reported in the original meta-analysis’s forest plot were refitted using a best estimation of the original meta-analytic strategy, results differed from those reported in the original (albeit, by a small amount). More worryingly, when all effect sizes were reextracted from the original articles a large number of questionable inclusions and inclusions were highlighted. When all effect sizes were included that a) met Vahey et al.’s inclusion criterion of being clinically relevant and b) were not derived from types of analyses that were defined a priori as producing invalid or misleading results or conclusions, the meta effect size estimate reduced greatly (original *r* = .45, 95% CI [.40, .54], new: *r* = 0.13, 95% CI [0.03, 0.23]). Power analyses calculations for future research using this updated effect size estimate suggest minimum sample sizes of more than 460 participants; an estimate that is 16 times larger than recommended by Vahey et al. and 10 times larger than the mean sample sizes employed in IRAP research to date.

At first glance, these sample sizes seem unfeasible, especially given many researchers experience of conducting IRAP research and obtaining significant results. However, results are not incompatible with this: IRAP papers frequently include a large number of statistical tests and comparisons and a very high ratio of tests to sample size. As such, the false positive rate is inevitably inflated. Future research should attempt to estimate the false positive rate in IRAP research, possibly via simulation studies (e.g., due to analytic degrees of freedom and multiple testing).

## Improving the reproducibility of future meta-analyses

Results have implications for both the IRAP specifically (e.g., the interpretation of previously published findings and use in future studies), and also meta-analysis more generally. potential pitfalls involved in producing reproducible meta-analyses and interpreting the reproducibility of existing meta-analyses more generally.

Provide all data, including a codebook, and data regarding the excluded effect sizes. Provide all code and scripts for data processing and analyses. No written description of the analytic strategy will provide the same precision as the code used to implement them (along with session info information that includes the versions of software used along with details of the operating system and hardware used). Supplementary materials should not only be hosted on the journal’s website but also on reliable archival services (e.g., OSF, Zenodo, etc.). Organizing and publicly archiving such data ahead of time removes avoids many issues likely to be encountered in the future. For example, loss or misplacement of data and materials over time or unwillingness to search for them (all of which were encountered here when attempting to obtain data and materials from the authors of the original meta-analysis).

Explicate more details in text. For example, the weighting strategy was unclear in Vahey’s meta-analysis.

## Conclusion

XXX