# Stage 1 Verification Report Submission Template

# Title

Verification Report: 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”

# Abstract

Vahey et al.’s (2015) concluded that the Implicit Relational Assessment Procedure (IRAP) has potential “as a tool for clinical assessment”. They reported power analyses which have been used frequently to determine sample sizes. This article assesses the computational reproducibility of Vahey et al.’s results. On the whole, conclusions could not be reproduced and many apparent errors were detected, generally in favour of over-estimating the IRAP’s validity. A new meta-analysis and power analysis suggested that the IRAP has weak criterion validity for clinically-relevant variables and requires very large sample sizes.

# Keywords

implicit relational assessment procedure; implicit attitudes; meta-analysis; criterion validity; verification report

# Introduction

At minimum, the introduction should include a brief introduction to the topic, and a clear justification of the importance of the verification attempt.

Indirect measures of implicit attitudes have seen wide use in many areas of psychology research over the last twenty five years, including psychopathology research (e.g., Greenwald & Lai, 2020; Roefs et al., 2011). Unlike self-reports, implicit measures aim to infer individuals’ attitudes through reaction time biases, misattributions, and other forms of automatic behaviour (De Houwer & Moors, 2010; although see Corneille & Hütter, 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 ”demonstrates the potential of the IRAP as a tool for 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 favourably to other a more popular implicit measure, the Implicit Association Test (Greenwald et al., 1998), and (c) used their meta-analysed estimate of effect size to conduct power analyses and make sample size recommendations for future research using the IRAP.

There are two strong rationales to perform a verification of Vahey et al. (2015). First, there is good a priori reason to believe that meta-analyses in general often contain non-replicable results. Lakens et al. (2017) recently demonstrated that the results of the majority of a random sample of meta-analyses published in psychology cannot be computationally reproduced, often because of differences in individual effect sizes between those reported in meta analyses and those reproduced from the original studies. Similarly, Maassen et al. (2020) found that almost half of individual effect-sizes reported in meta-analyses of psychology research could not be reproduced from the original articles. This was attributed to due to a variety of issues including errors in the extraction of effect sizes from original studies, insufficient details regarding data processing and transformation of effect sizes, and insufficient details of the specific meta-analytic approach employed.

Second, Vahey et al.’s (2015) article has been well-cited and used to guide subsequent work. At time of writing, it has been cited 119 times with roughly 20% of articles citing it to justify sample size decisions (i.e., in lieu of a power analysis for that study). Studies employing the IRAP have typically involved small sample sizes of around 40 participants. This is frequently argued to be acceptable because it is in line with Vahey et al.’s (2015) sample size recommendation: “a sample size of at least N = 37 would be required in order to achieve a statistical power of .80 when testing a continuous first-order correlation between a clinically-focused IRAP effect and a given criterion variable” (p. 63). Kavanagh et al. (2022, p. XX) provides a particularly clear characterization of the ongoing importance of Vahey et al.’s (2015) results for practices in the broader IRAP literature: “The general strategy for recruiting numbers of participants was guided by the results of a recent meta-analysis of IRAP effects in the clinical domain, indicating that a minimum of 29 is required to achieve a power of 0.8 for first-order correlations (Vahey et al., 2015).” Given that research continues to rely on the conclusions of Vahey et al.’s (2015) meta-analysis, and that meta-analyses in general have been shown to have poor computational reproducibility, it is therefore useful to verify Vahey et al.’s results.

Vahey et al. (2015) reported the steps in their analyses in the conventional order: they identified effect sizes in original article, applied inclusion and exclusion criteria, extracted them, converted them to Pearson’s *r*, averaged them when multiple effect sizes came from a given study, fit a meta-analysis model, and performed a power analysis on the meta-effect size to guide sample size determination in future studies. Attempts to verify these steps for this article were conducted, and are reported here, in reverse order. Subsequently, I report a new meta-analysis and power analysis using the reextracted individual effect sizes.

# Method

A detailed protocol describing the (re)analyses. This should be comprehensive in detail and include links to all materials and code required.

## Transparency statement

All data, code, and formulae (e.g., to convert effect sizes) to reproduce the verification and extension analyses can be found in the supplementary materials (see osf.io/XXXX). We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study (Simmons et al. 2012).

## Scope and purpose of Vahey et al. (2015)

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. They reported that they found 46 empirical articles that employed the IRAP. The authors extracted 56 effect sizes from 15 articles. A list of source articles and individual effect sizes were provided in their Supplementary Materials, in both their original form as well as the Pearson’s *r* value they were converted to. However, no details were provided on the specific methods or formulas that were used to convert these effect sizes. Vahey et al.’s (2015) article included a forest plot that contains both plotted and numerical values for effect sizes, confidence intervals, and sample sizes. No details were provided on how confidence intervals were estimated.

## Original inclusion and exclusion criteria

Vahey et al.’s (2015) stated that clinical relevance was an inclusion criterion: “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).

## Correspondence with the authors

One common barrier to reproducing results is the unavailability of code, whether openly or upon reasonable request. While attempting to reproduce Vahey et al.’s (2015) results, I contracted the first author and requested they send me their code. He declined to share these materials “until there are specific criticisms for [them] to address”, and suggested that I instead use the code provided by Field & Gillett (2010) to recreate their analyses. I provided specific criticisms in a conference presentation in June 2019, which the first author attended. I did not subsequently receive any materials from the authors.

## Power analyses

Details of the power analyses conducted by Vahey et al. (2015) were extracted. This included the meta-effect size used (i.e., using point estimate or lower bound confidence interval, following Perugini et al.’s recommendation, as adopted in Vahey et al. 2015), test (Pearson’s r correlation, independent t-test, dependent t-test), direction of hypothesis (one-sided vs. two sided), and the recommended sample size (i.e., the result of the test). Verification tests were performed using the pwr R library (REF). Table XX contains all details of the original power analyses reported by Vahey et al. (2015) and the results of the verification analyses. As can be seen in the table, Vahey et al.’s sample size recommendations were found to be computationally reproducible when their meta-analytic effect size was used. [add additional analyses to code; remove references to two sided tests not being reported]

## Implementation of the meta-analysis

Vahey et al. (2015) stated that they employed a Hunter & Schmidt style meta-analysis and cited Field & Gillett (2010). The latter authors’ described the Hunter & Schmidt style meta-analysis as involving an average Pearson’s r effect size that was weighted by sample sizes, and reporting credibility intervals rather than confidence intervals. I return to the definition of credibility intervals later. Vahey et al. (2015) did not specify in their article how they implemented their analyses nor did they make their code available. When contacted, the first author declined to share their code and suggested that the SPSS scripts associated with Field & Gillett (2010) should be used to recreate their analyses. The accompanying materials for Field & Gillett (2010) were therefore downloaded from Field’s website. This contained two different SPSS scripts, “Meta\_Basic\_r.sps” and “h\_s syntax.sps”.

The code in these scripts was examined in order to determine whether they (individually or jointly) were capable of calculating the various values that were reported in Vahey et al. (2015). That is, agnostic to the numerical results actually produced by running the scripts, I first assessed whether these scripts were capable of producing (a) all of the variables that Vahey et al. (2015) reported and (b) producing them in an unambiguous way (i.e., only calculating a given variable via one method). Inspection of these two scripts showed that (a) neither script was capable of computationally reproducing all four of the forest plot, the meta-analytic point estimate, the confidence intervals, and the credibility intervals. However, (b) the scripts were also not capable of jointly reproducing the variables reported in Vahey et al. (2015) in an unambiguous way, because they implemented slightly different methods.

Specifically, on the one hand, the “Meta\_Basic\_R.sps” script applies an Overton transformation to the Pearson’s *r* correlations prior to analysis, whereas “h\_s syntax.sps” did not. Field & Gillett, 2010, stated that the Overton transformation should only be employed in a Hedges’ style meta-analysis and not the Hunter & Schmidt method (i.e., this is Field & Gillet’s incongruity, not Vahey et al.’s). In addition to this, this script does not calculate confidence intervals for Hunter & Schmidt style meta-analyses. On the other hand, the “h\_s syntax.sps” script requires reliability estimates for both variables in each correlation (i.e., the reliabilities and for the correlation ) in order to correct the meta-estimates for attenuation. Vahey et al. (2015) did not report extracting or using reliability estimates in this way in their manuscript or supplementary materials. This script does calculate confidence intervals on the meta-estimate but not the individual effect sizes, therefore the forest plot could not be reproduced from this script.

Other than the above divergences, the code in each script was confirmed to accuracy implement the equations described in Field & Gillett (2010; documentation of this validation is available in the supplementary materials). As such, in the absence of additional information, the two scripts provided by Field & Gillett (2010) were therefore not capable of fully and unambiguously reproducing the same variables as those reported in Vahey et al. (2015) without further alternation. In light of this, I therefore altered the implementations in multiple ways in order to attempt to reproduce Vahey et al.’s (2015) results.

### Definition of credibility intervals

Vahey et al. (2015) report what they refer to as Credibility Intervals (CR), which attempt to estimate the generalizability of the meta-effect size (Field & Gillett, 2010). However, there is some ambiguity around the term credibility interval. In order to ensure that the verification analyses were correctly implemented, it is important to first define them.

Broadly speaking, and using the language of linear mixed effects models, whereas confidence intervals are based on the standard error of the intercept (), credibility intervals are based on the standard deviation of the random effect (often denoted ). Credibility Intervals posed two challenges to verification: (1) they are referred to by different names by different authors (e.g., Vichtbauer REF refers to them as prediction intervals), and (2) even when using the same name, they are defined differently by different authors (this point will be returned to later). It is therefore important to be precise when defining the interval defined and implemented by Field & Gillett (2010), which the first author of Vahey et al. (2015) stated in a personal communication that they used for their analyses. Field & Gillett (2010) defined the Hunter & Schmidt style credibility interval as the meta-analytic effect size ± the critical *t* value (1.96 for 95% intervals) multiplied by the square root of the variance in population correlations (Field & Gillett, 2010, equation 5):

They define the variance in population correlations as the variance of sample effect sizes (which Vahey et al. 2015 denote as ) minus the sampling error variance (Field & Gillett, 2010, equations 2, 3, and 4 combined):

Field & Gillett’s definition of the Hunter & Schmidt style credibility interval is therefore based solely on the standard deviation of the population variance. Vahey et al. (2015) state that such “[Hunter & Schmidt style] credibility intervals are generally wider and thus more conservative than corresponding confidence intervals” (p.61). However, as Vahey et al. (2015) acknowledge in this quote, this may be generally true but it is not always the case. For example, if the sampling error variance is found to be larger than the variance in the sample effect sizes, then will be negative. If is negative, then credibility intervals cannot be calculated, as the square root of a negative number is non-real. Although Field and Gillett (2010) do not discuss this possibility in their article, they cover this case in their implementation by setting negative values of to zero (see “h\_s syntax.sps” script). In such cases, both the lower and upper bound of the credibility interval will equal the point estimate (i.e., ). This would represent an important case in which confidence intervals are much more conservative than credibility intervals. The point to be appreciated here is that the definition and implementation of credibility intervals used in all verification analyses here precisely follows the definition of these intervals in Field & Gillett (2010) and their implementations in by Field’s SPSS syntax, which Vahey et al. (2015) reported using.

## Meta-analysis

### The meta-analytic results to be reproduced

Vahey et al.’s reported a meta-analytic effect size, 95% confidence intervals, and 95% credibility intervals. These were extracted from Vahey et al.’s (2015) forest plot (point estimate and CR) and the text on page XX (CI): r = .45, 95% CI [.40, .54], 95% CR [.23, .67] (see also Table XX). Prior to any attempt to reproduce these results, it is useful to notice the asymmetric confidence intervals around the point estimate: the upper bound is +9 from the point estimate, whereas the lower bound is -5 from the point estimate. While asymmetric intervals are indeed possible (e.g., when using a transformation such as Fisher’s r-to-z), this would result in the opposite pattern where the lower bound is further from the point estimate than the upper bound. There is a priori reason to expect that the analytic method that Vahey et al. (2015) state that they employed, and indeed other variations of meta-analysis, should not produce confidence intervals that are asymmetric in this way. I will return to this point later.

### Verification attempt 1

All verification analyses were conducted using the individual averaged effect sizes and sample sizes reported numerically in Vahey et al.’s (2015) forest plot. The forest plot verification analysis also used the confidence intervals reported numerically in their forest plot.

The first verification attempt employed Field’s “h\_s syntax.sps” script. Copies of all original and altered scripts are available in the supplementary materials. The credibility interval widths were changed from 80% to 95%, as Vahey et al. (2015) reported using the latter. One other key assumption was made in order to allow the script to run. To take a step back, a Hunter & Schmidt style meta-analysis is sometimes referred to as a form of psychometric meta-analysis because it typically involves de-attenuating the effect sizes based on the reliability of the measures that produced them (REF). Although Hunter & Schmidt did describe what they referred to as a “bare-bones” meta-analysis that did not perform this deattentuation based on reliability, Field’s “h\_s syntax.sps” script requires the researcher to provide reliability values for each effect size for the script to run. Vahey et al. (2015) do not report any extracting or estimating reliabilities or deattenuating the effects based on them, and no reliability data is available in their manuscript or supplementary materials. In the absence of other information, I set reliability for all variables to 1 in order to allow the script to run. Table XX presents the meta-analysis effect sizes estimates reported by Vahey et al. (2015) as well as the results of all verification analyses conducted here. As shown in Table XX, this verification attempt did not reproduce the original results for any estimate (point estimate, confidence interval, or credibility interval). The point estimate was off by a small amount (r = 0.02), but more than could be accounted for by rounding. Confidence intervals were nearly four times wider in the verification analysis than the original results. Credibility intervals were infinitesimally narrow (i.e., because was negative and interval width was therefore set to zero), whereas they were wider than the confidence intervals in the original analysis.

### Verification attempt 2

The second verification attempt employed Field’s “Meta\_Basic\_r.sps” script. Note that, unlike the “h\_s syntax.sps” script, this script does not contain code to calculate confidence intervals for the Hunter & Schmidt approach. One alteration was made to the implementation: if was negative it was set to zero, as in the “h\_s syntax.sps” script. Without this alternation, if was negative, the script would fail to run. Table XX again presents the results. As shown in the table, this verification attempt also did not reproduce the original results. The point estimate was off by a small amount (r = 0.01), but again more than could be accounted for by rounding. Credibility intervals were again infinitesimally narrow (i.e., because was negative and interval width was therefore set to zero).

### Verification attempt 3

On close inspection, the Field’s “Meta\_Basic\_r.sps” script applies an Overton correction (REF) to the correlations prior to meta-analysing them. This is likely an error, as according to Field & Gillett (2010) themselves, this correction is intended to be applied prior to a Fisher’s r-to-z transformation, which the Hunter & Schmidt style meta-analysis does not employ. Note that this implies that verification attempt 2 should not be considered as a valid estimate of the true effect size, as it involves an inappropriate correction. To cover the possibility that Vahey et al. (2015) also noticed this error in Field’s code and corrected it, the third verification attempt was identical to the second other than employing the weighted-average effect sizes without an Overton transformation. As shown in Table XX, this verification attempt also did not reproduce the original results. The point estimate was again off by a small amount (r = 0.02), and was now identical to the result found in verification attempt 1. Credibility intervals were again infinitesimally narrow because was negative.

### Verification attempt 4

In a fourth verification attempt, I switched from using Field’s implementation in SPSS to Vichtbauer’s Hunter & Schmidt implementation in R (REF) using the metafor package (REF). This provided another form of robustness (i.e., to implementation), as well as a new avenue to attempt to reproduce the original results in a programming language I was more familiar with. Hunter and Schmidt style meta-analyses following Field & Gillett’s (2010) equations were implemented. As shown in Table XX, this time the confidence intervals reported in Vahey et al. (2015) were reproduced. However, the point estimate and credibility intervals again did not reproduce the original results, and matched the results found in verification analyses 1 and 3.

This verification attempt also attempted to reproduce the original forest plot (Vahey et al.’s, 2015, Figure 1), which was more feasible in R. It is useful to note that the original forest plot reported asymmetric confidence intervals around individual effect sizes. That is, the lower bounds are typically further from the point estimate than the upper bounds. This implies that some form of non-linear transformation was employed, such as a Fisher’s *r*-to-*z* transformation. However, Vahey et al. (2015) do not report employing any transformations in their meta-analysis or forest plot. The forest plot associated with this verification attempt can be seen in Figure XX. Confidence intervals around individual effect sizes were symmetric and therefore did not reproduce the original plot.

### Verification attempt 5

In the final verification attempt, I apply Fisher’s *r*-to-*z* transformations to the individual effect sizes prior to meta-analysis, and back transformations prior to reporting and plotting. The analysis was otherwise identical to the previous attempt. All estimated values were identical to the previous attempt, therefore the original results were not reproduced. The forest plot associated with this attempt reproduced the confidence intervals around the individual effect sizes in the original plot (see Figure XX), suggesting that Vahey et al. (2015) employed these transformations but did not report them.

### Summary

Confidence intervals around individual effect sizes in the forest plot were only reproduced when Fisher’s *r*-to-*z* transformations were applied (verification attempt 5) and not when they weren’t (verification attempt 4). However, (a) Vahey et al. (2015) did not report applying this transformation in their meta-analysis or plot, and (b) Field & Gillett (2010) stated that this transformation is not part of the Hunter & Schmidt method of meta-analysis. This would therefore be an ad-hoc mix of the Hunter and Schmidt method and the Hedges and colleagues’ method, as Field & Gillett (2010) describes them. A point estimate of r = .47 was produced by in all four correctly implemented analyses (i.e., excluding attempt 2). Interestingly, as noted previously, if we assume that Vahey et al.’s (2015) confidence intervals were accurate and symmetric – and it should be noted that they are congruent with the results of verifications 4 and 5 – they would imply a point estimate of r = .47. One possible explanation is that the point estimate of .45 is a typo that was propagated throughout Vahey et al.’s manuscript, plots and tables. The confidence intervals were not reproduced using the scripts that Vahey et al. (2015) reported using, but were reproduced using different implementation of the analysis in R that Vahey et al. (2015). This is difficult to account for. The credibility intervals could not be reproduced in any analysis.

## Meta-analytic effect size

Vahey et al.’s (2015) aforementioned power analyses relied on the accuracy of the meta-analytic effect size. I attempted to computationally reproduce the meta-analytic effect size from the weighted-mean effect sizes and sample sizes reported in Vahey et al.’s (2105) forest plot (p.XX).

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).

## Weighted-average effect sizes

Vahey et al.’s (2015) meta-analysis results relied on the accuracy of the weighted-mean effect sizes used in it. I attempted to computationally reproduce the weighted-mean effect sizes presented in their forest plot from the individual effect sizes and degrees of freedom presented in their supplementary online materials. Weighted-mean effect sizes are one strategy that can be employed to deal with the non-independence of multiple effect sizes taken from a given study or sample. Vahey et al. (2015) reported that they followed the method suggested by [REF] and weighted by degrees of freedom. Results were not computationally reproducible in 2 of 15 (13%) of cases. The magnitudes of the differences were small (Δ*r* = -.02 and .05).

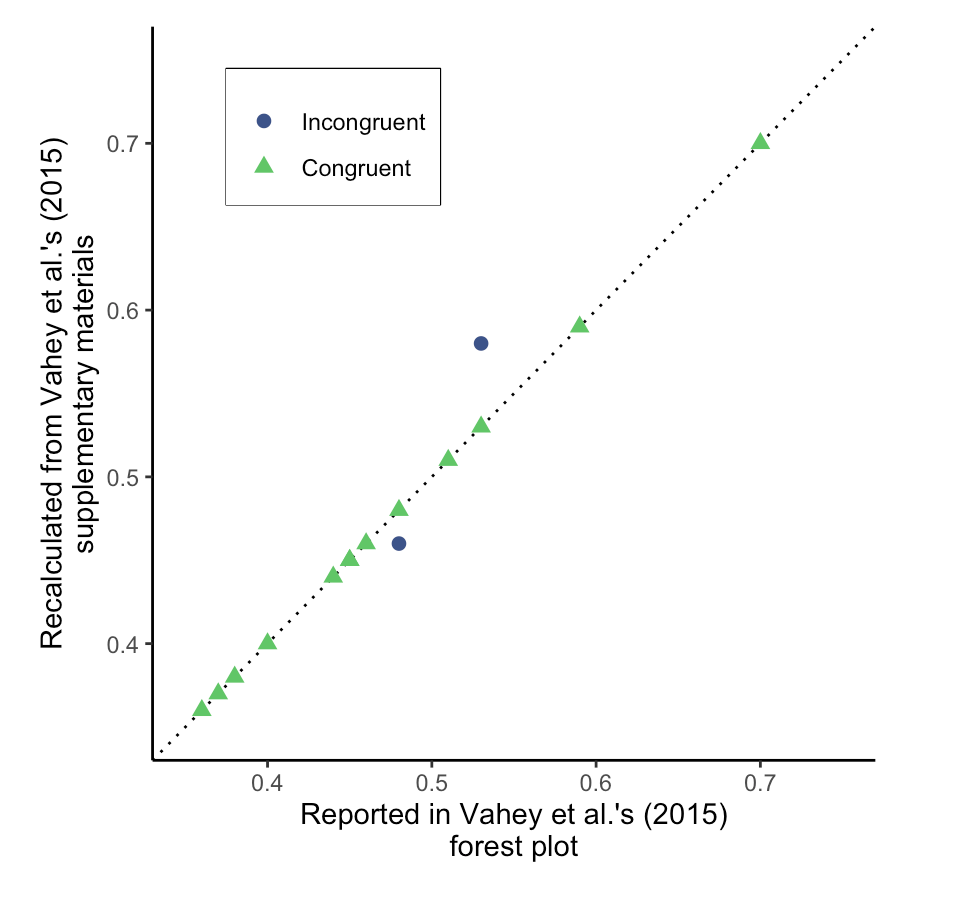


Figure XX. XXXX

## Individual effect sizes

Vahey et al.’s (2015) weighted-mean effect sizes in turn relied on the accuracy of the individual effect sizes that were extracted from original research articles (along with other statistics such as *N* and *df*) and, where applicable, the mathematical conversion between other effect sizes to Pearson’s *r*. I therefore attempted to computationally reproduce the individual effect sizes presented in Vahey et al.’s (2015) supplementary online materials. I make a distinction between two subsets of effect sizes and their reproducibility.

The first subset refers to effect sizes that could be reextracted and converted to Pearson’s *r*. In these cases, reproducibility refers to the numerical congruence between the effect sizes I obtain and those reported by Vahey et al. (2015). Wherever possible, the same effect size conversion method was employed as in the original meta-analysis, following the approaches listed in their supplementary materials. However, while these approaches were listed by name, specific formulae or software implementations were not provided. 29 (52%) effect sizes could be reextracted. When rounding all effect sizes to two decimal places, nearly half of the effect sizes reported by Vahey et al. (2015) could not be computationally reproduced (13 effect sizes, 45%). The magnitude of the differences between Vahey et al.’s effect sizes and mine were large in some cases (Δ*r*max = -.44). Where differences were observed, Vahey et al.’s (2015) effect sizes were generally skewed in favour of the IRAP’s validity (see Figure XX).

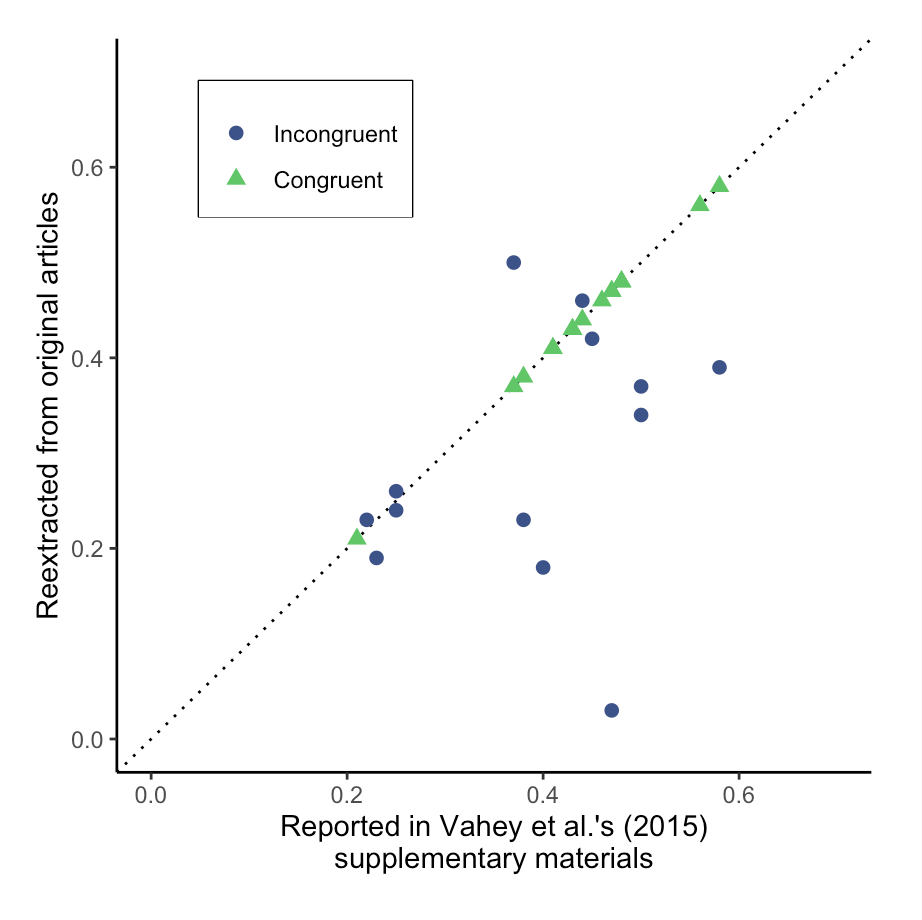


Figure XX. XX

The second subset of effect sizes refers to cases where I have a documented reason to believe that the effect size should not have been included in the meta-analysis for one or more of the following reasons. First, Vahey et al. (2015) appear to have treated as if it was equivalent to , which it is not: (a) has a relatively simple mathematical transformation to Pearson’s *r*, which Vahey et al. (2015) appear to have incorrectly applied to . However, cannot be converted to Pearson’s *r* as it is partial correlation. Additionally, has non-equivalent interpretation between different factorial designs (Lakens, 2013). As such, a number of effect sizes included by Vahey et al. (2015) were not reproduced.

Second, in some cases, effect sizes reported in Vahey et al.’s (2015) supplementary materials did not refer to effect sizes that were reported in the original article (e.g., Timko et al., 2010 Study 1: correlation between overall IRAP *D* score and DASS-total).

Third, in some cases, effect sizes referred to ANOVAs where mean IRAP *D* scores were used as the Dependent Variable (e.g., Kosnes et al., 2013, Parling et al., 2012; Hussey et al., 2012; Timko et al., 2010). Predicting mean IRAP effects from known groups tells us little about the IRAP’s validity, which would be appropriately assessed by through the IRAP’s ability to predict group membership. This analytic issue of swapping the IV and DV when attempting to provide evidence for a measure’s validity has been well documented elsewhere as a threat to research findings (Fried & Kievit, 2016).

Fourth, Vahey et al. (2015) included a large number of effect sizes that referred to tests of whether an IRAP effect had been demonstrated. That is, whether mean IRAP *D* scores were non-zero, or whether a reaction time differential was found between the consistent and inconsistent blocks. However, criterion validity can by definition only be established with reference to external variables. Quantifying the evidence for IRAP effects in isolation is at odds with Vahey et al.’s (2015) stated goal of assessing the IRAP’s clinically relevant criterion validity. As such, a number of effect sizes were not reproduced for this reason.

Finally, some effect sizes were not reported in sufficient detail in the original paper to allow for the calculation of an effect size. In such cases, I contacted the original authors, however in many cases I was not able to obtain additional data. These cases represent greater success by Vahey et al. (2015) in assembling results than I was able to achieve.

In total, only XX of XX effect sizes included in Vahey et al.’s (2015) supplementary materials were found to be computationally reproducible. Where reextracted values were found to differ, these differences were generally in the IRAP’s favour in Vahey et al. (2015, see Figure XX).

## Omitted effect sizes

XXX

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.

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).

## New meta-analysis

The majority of the step in Vahey et al.’s (2015) meta-analysis were not found to be computationally reproducible (i.e., meta-analysis results, calculation of weighted-mean effect sizes, or extraction and conversion of individual effect sizes). Where steps were found to be computationally reproducible, they were found to be poorly justified (e.g., power analyses). In some cases, one could argue that differences between the results reported by Vahey et al. (2015) and those reported here are small (e.g., meta-analytic effect size estimate). However, no individual step can be viewed in isolation. For example, the large differences in individual effect sizes had an as-yet unknown impact on the meta-analytic effect size estimate. In order to assess the compound impact of the reproducibility at each step on Vahey et al.’s (2015) final results and conclusions, a new meta-analysis was conducted, followed by new power analyses using the meta effect size.

Recent results from simulation studies suggests that the weighted-mean approach method employed by Vahey et al. (2015) to deal with non-independence of effect sizes estimates provides poor statistical power, and that the alternative approach of employing a multi-level meta-analysis model should instead be employed (REF). 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 Likelihood estimator function

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. 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]).

Given the large number of effect sizes being meta-analysed, 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).

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.

## [old points]

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.

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.

Assessment of bias

One or more authors of Vahey et al. (2105) 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.

## 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

# References

Please enter references in the APA style and include a DOI where available.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, *60*, 527–542.

Corneille, O., & Hütter, M. (2020). Implicit? What Do You Mean? A Comprehensive Review of the Delusive Implicitness Construct in Attitude Research. *Personality and Social Psychology Review*, 1088868320911325. https://doi.org/10.1177/1088868320911325

De Houwer, J., & Moors, A. (2010). Implicit measures: Similarities and differences. In *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176–193). Guildford Press.

Fried, E. I., & Kievit, R. A. (2016). The volumes of subcortical regions in depressed and healthy individuals are strikingly similar: A reinterpretation of the results by Schmaal et al. *Molecular Psychiatry*, *21*(6), 724–725. https://doi.org/10.1038/mp.2015.199

Greenwald, A. G., & Lai, C. K. (2020). Implicit Social Cognition. *Annual Review of Psychology*, *71*(1), 419–445. https://doi.org/10.1146/annurev-psych-010419-050837

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Heo, M., Kim, N., & Faith, M. S. (2015). Statistical power as a function of Cronbach alpha of instrument questionnaire items. *BMC Medical Research Methodology*, *15*. https://doi.org/10.1186/s12874-015-0070-6

Hussey, I., & Drake, C. E. (2020). The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis. *Preprint*. https://doi.org/10.31234/osf.io/ge3k7

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D., Page-Gould, E., van Assen, M. A. L. M., Spellman, B., Schönbrodt, F. D., Hasselman, F., Corker, K. S., Grange, J., Sharples, A., Cavender, C., Augusteijn, H., Augusteijn, H., Gerger, H., Locher, C., Miller, I. D., Anvari, F., & Scheel, A. M. (2017). *Examining the Reproducibility of Meta-Analyses in Psychology: A Preliminary Report* [Preprint]. BITSS. https://doi.org/10.31222/osf.io/xfbjf

Maassen, E., Assen, M. A. L. M. van, Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107

Parsons, S. (2018). Visualising two approaches to explore reliability-power relationships. https://doi.org/10.31234/osf.io/qh5mf

Roefs, A., Huijding, J., Smulders, F. T. Y., MacLeod, C. M., de Jong, P. J., Wiers, R. W., & Jansen, A. T. M. (2011). Implicit measures of association in psychopathology research. *Psychological Bulletin*, *137*(1), 149–193. https://doi.org/10.1037/a0021729

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (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. https://doi.org/10.1016/j.jbtep.2015.01.004

# Other information required for submission, not for review

Contribution Statement

Please list all contributions towards this manuscript, including their roles and affiliations at the time of data collection.

Ian Hussey was solely responsible for all contributions to this manuscript. I was affiliated with Ghent University, Belgium, when I began this project. I am now affiliated with Ruhr University Bochum, Germany.

Acknowledgements

Many thanks to Jamie Cummins for feedback on earlier versions of this manuscript.

Conflict of Interest

I acknowledge that one of the authors of the original article being verified (Prof Dermot Barnes-Holmes) was my PhD supervisor (2010-2015). I have not actively collaborated with Prof Barnes-Holmes since 2015. Articles lead by third parties of which we were both co-authors were published up to 2018. The author declares no other conflict of interest associated with the publication of this manuscript.

Funding statement

IH was supported by Ghent University grant 01P05517 (awarded in 2017) and the META-REP Priority Program of the German Research Foundation (#464488178) (awarded in 2021).

# Stage 1 Checklist

Include a separate page, confirming explicit agreement of the following:

1. All necessary support (e.g., funding, facilities, etc.) and approvals (e.g. ethics) are in place for the proposed research
2. The cover letter includes an anticipated timeline for completing the work if the initial submission is accepted
3. The authors agree to share their raw data, materials and code as appropriate.
4. In the event of the submission achieving Stage 1 in-principle acceptance, authors confirm that they agree to the journal registering their approved protocol on their behalf on the Open Science Framework (OSF) using its dedicated Stage 1 VR registration mechanism https://osf.io/rr/ (please see the Verification Report author guidelines for further details). The journal will provide the corresponding author with the URL to this registered protocol in the Stage 1 editorial acceptance letter, and authors must later include this URL in the Stage 2 manuscript. Note that the journal will register the protocol ONLY once the Stage 1 manuscript is in-principle accepted, and not if it is rejected or withdrawn by authors prior to being awarded in-principle acceptance.

For each author who currently has an account on the OSF (https://osf.io/), please provide their name and the URL of their OSF home page. E.g. “Thomas Rhys Evans, osf.io/ydmcr”. In the event of the Stage 1 protocol receiving in-principle acceptance, journal staff will include these authors as contributors to the OSF registration. It is not required that all authors have an OSF account, but only authors with an OSF account will be included by journal staff as contributors to the registered protocol on the OSF. At least ONE author must have an OSF account to ensure that the registered protocol is linked to at least one member of the authoring team. In the event of Stage 2 acceptance, authors without an OSF account will still be named as authors on the published article.

If the submission achieves Stage 1 in-principle acceptance, authors can instruct the journal to either make the registered Stage 1 manuscript immediately public on the OSF or instead register it under a private embargo for up to 4 years from the date of registration. If authors choose a private embargo, the embargo will be released and the registered protocol made public when any one of the following conditions are met: (a) submission of the Stage 2 manuscript; (b) withdrawal of the submission after in-principle acceptance and consequent triggering of a Withdrawn Registration (see Q5); or (c) natural expiry of the embargo period. Please choose the authors’ preferred method of registration following Stage 1 in-principle acceptance: Made public immediately OR Under private embargo. If choosing a private embargo please enter the duration of the embargo following in-principle acceptance. This can be specified either as a duration (e.g. “2 years”) or as a specific future date. The embargo period must be less then 4 years. Any entries that exceed this permissible maximum will be treated by the journal as “4 years”.

1. The authors confirm that if they withdraw their paper following Stage 1 in-principle acceptance then they agree to the journal (a) lifting any applicable private embargo on the registered Stage 1 protocol, thus making the protocol public on the OSF; and (b) publishing a short summary of the pre-registered study under the journal section Withdrawn Registrations, which will include the abstract of the Stage 1 submission, the URL of the registered Stage 1 protocol on the OSF, and a stated reason for the withdrawal.
2. Should Stage 1 in-principle acceptance be forthcoming, authors will be asked to provide the journal office with an estimated submission date for the completed Stage 2 manuscript. This deadline can be readily altered in consultation with the editors (e.g., in case of delays requiring additional time to complete the research). However, in the event that the authors (a) fail to submit the Stage 2 manuscript within 6 months of the mutually agreed deadline, while also (b) becoming non-responsive during this period to editorial enquiries, then the manuscript will be considered by the journal to be withdrawn, triggering publication of a Withdrawn Registration as outlined in Q5. Please confirm the authors’ agreement to these conditions.
3. The authors confirm that both of the following statements are true: (a) none of the authors of the Verification Report were authors or co-authors of the original work, and (b) none of the authors of the Verification Report are active collaborators of any of the authors of the original work (including holding shared grant funding in the previous three years or any other close connections).
4. The authors confirm that the cover letter makes clear whether the authors are submitting their Stage 1 manuscript prior to acquiring the data (and if so, a confirmation that all necessary permissions to acquire and reanalyse the data have been granted by the original authors/data custodian) OR instead whether they have already acquired the data and completed the (re)analyses but with the results redacted.

I, the single author, Ian Hussey (osf.io/3kzh8), confirm my agreement to all of the above points.