# 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. (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 psychological research, 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 the one implicit measure, the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2010), concluded that it possesses good criterion validity and furthermore ”demonstrates the potential of the IRAP as a tool for clinical assessment” (Vahey et al., 2015). Based on a non-systematic review, 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.

## Rationale for verification

There are at least three 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. Comparable errors in meta analyses have also been reported by others (Gøtzsche et al., 2007).

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 122 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. 528) 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 (2015) results.

Third, in light of recent estimates of the IRAP’s low reliability (Hussey & Drake, 2020), there is good reason to believe that Vahey et al.’s (2015) meta-analytic estimate of *r* = .45 is implausibly large. According to classical test theory, a measure’s reliability refers to the proportion of the variance that is caused by the construct measure rather than noise (Allen & Yen, 2002, p.73). As such, reliability places a limit on the mean observable associations between scores on any two measures: the less reliably the two variables are measured, the lower the observable correlation between the two variables. The observed correlation between two measures and () is a function of the true correlation between the variables () and the reliability of both measures (their self-correlation and ). This can be quantified via the Attenuation Formula derived from classical test theory (Revelle, 2009, equation 7.3):

Two of these variables have already empirical estimates. First, Vahey et al.’s (2015) estimate of the observed correlation between the IRAP and criterion variables was = .45. Second, estimates of the IRAP’s reliability have been provided by a recent meta-analysis. At the trial-type level (i.e., the method of scoring IRAP as four scores that proponents of the task typically recommend), both internal consistency ( = .27) and test-retest (ICC2 = .18) are extremely low (Hussey & Drake, 2020).[[1]](#footnote-1) This leaves two remaining variables, the IRAP’s criterion validity after adjusting for measurement error () and the criterion tasks’ mean reliability (). Both of these variables share the same constraint: as correlations, their value cannot be above 1. For the moment, if we assume that the criterion tasks’ mean reliability is very good ( = 0.90). This would imply that the lower limit of the IRAP’s true criterion validity after adjusting for measurement error is between implausibly high, = .91 (when using the estimate of internal consistency), or mathematically impossible as it is beyond the [0,1] bounds for a correlation, = 1.12 (when using the estimate of test-retest). Using lower and arguably more plausible values for the mean reliability of the criterion tasks produces even higher estimates for the true correlation, making both values impossible (i.e., when = .70, = 1.04 or 1.27 respectively).

Given that these estimates range between highly implausible and impossible, something appears to be amiss. Either Vahey et al.'s (2015) estimate of average criterion associations is somewhere between highly implausible and mathematically impossible given the IRAP's reliability (i.e., assuming Hussey & Drake 2020 are right about reliability), or Hussey & Drake's (2020) estimates of the IRAP's average reliability is implausibly low given the IRAP's high criterion validity (i.e., assuming Vahey et al. 2015 are right about validity). Ultimately it will be up to the community to determine whether our analyses in Hussey & Drake (2020) are sound, and we provide open data and code to aid others in inspecting our work for errors. Because we are confident in the numbers, this motivates us to inspect Vahey et al.'s (2015) data and analyses to determine if the issue lies there instead.

# Method & Findings

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

*Provide a comprehensive overview of the findings of the analyses as preregistered. In the case of results which were not preregistered but result from justified deviations from the protocol should clearly be reported under a separate ‘Exploratory Findings’ heading.*

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 reported here in reverse order. Subsequently, I report a new meta-analysis and power analysis using the reextracted individual effect sizes.

## 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). In the process of conducting this verification attempt, I contacted the corresponding author of Vahey et al. (2015) and requested that they share their code or further details of their analytic approach, who declined. I also shared a copy of an earlier version of these verification attempts with the corresponding author in July 2019, including code, data, and a set of slides outlining my concerns about Vahey et al.’s (2015) computational reproducibility. I have received no contact from the corresponding author since then.

## 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 (2015) sample size recommendations were found to be computationally reproducible when their meta-analytic effect size was used. [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 how they implemented their analyses or 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

Vahey et al.’s (2015) 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 meta-analysis 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 (2015) 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.

## Average effect sizes

Vahey reported that the 15 weighted average effect sizes they used in their meta-analysis were calculated from the 46 individual effect sizes and degrees of freedom they reported in their supplementary materials. I attempted to verify this by recalculating weighted averages from the individual effect sizes, using Vahey’s strategy of weighting by degrees of freedom. Results were not computationally reproducible: 2 of 15 (13%) recomputed weighted averages differed from those reported in Vahey’s forest plot. On the one hand, the magnitudes of the differences were small (Δ*r* = -.02 and .05). On the other hand, given the simplicity of these calculations, they should be expected to be reproducible.

## Individual effect sizes

### Assessment of incorrect inclusions

Lakens et al. (2016) listed incorrect inclusion as a common type of error in meta-analyses. These were described as inclusion of effect sizes that do not meet the inclusion criteria. 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). Their inclusion criterion was 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). References to neither the specific clinical condition that was targeted by the IRAP and the criterion variable nor the specific empirical literature that Vahey et al. (2015) used to justify the inclusion of each criterion were provided in the original article or supplementary materials. Nonetheless, Vahey’s own inclusion criterion required that effects referred to covariation between an IRAP and an external clinically-relevant criterion variable, consistent with the APA dictionary of psychology definition of criterion validity (REF).

Using the descriptions in their supplementary materials, and with reference to the original papers, the individual effect sizes were re-evaluated against Vahey’s inclusion criterion of covariance between an IRAP and a second external variable. 23 of the 56 effect sizes (41% ) were found to involve no external variable (i.e., they described a reaction time differential between the IRAP block types), and were therefore not suitable to be included in a meta-analysis of the IRAP’s criterion validity.

### Assessment of incorrect exclusions

In addition to incorrection inclusions, it is equally plausible that effect sizes that met inclusion criteria were erroneously not included. Effect sizes were reextracted from the same 15 articles as Vahey et al. (2015). Vahey et al. did not report how many effect sizes were excluded (e.g., because they did not meet the inclusion criteria of being a clinically relevant criterion effect size). Following Vahey et al.’s (2015) method, extractions were not limited to effect sizes reported in the articles, but also considered ones implied by the reported analyses. Where necessary, original authors were contacted to obtain additional information. For example, if non-significant correlations were reported as merely “other correlations were non-significant”, these effect sizes were obtained where possible. XX effect sizes were excluded as non-criterion effects (e.g., quantifying the average IRAP effect within a group). Two independent raters the rated each effect for clinical relevance using Vahey’s definition. Agreement was found in 90% of cases (Cohen’s Kappa = 0.88, *p* < .0001). As in Vahey et al. (2015), if either rater originally rated the effect as clinically relevant then it was included.

308 effect sizes were originally extracted. 53 were excluded as non-criterion effect sizes. 99 more were excluded as non-clinically relevant. This left 156 effect sizes for meta-analysis, compared to Vahey et al.’s (2015) 56. Note that these extractions are not exhaustive: some authors of original studies who replied to Vahey and colleagues’ requests for additional information did not reply to mine. Vahey et al.’s (2015) extraction of effect sizes from the original articles was therefore not reproduced, as many effect sizes were overlooked in the original study. These effect sizes were converted to Pearson’s *r* for meta-analysis. The specific methods of conversion are described in the supplementary materials.

### Assessment of erroneous calculation

Erroneous calculation refers to errors made in the transposition, conversion, or reporting of effect sizes. This can involve using the incorrect formula to convert effect sizes, treating standard errors if they are standard deviations, and other errors. Previous work has shown that such errors are unfortunately common in published meta-analyses (e.g., XX of all effect sizes: Gøtzsche REF). In their supplementary materials, Vahey et al. (2015) provided explanations and references for how individual effect sizes were converted to Pearsons *r*. Inspection of the explanations revealed at least one error: 2 of the effect sizes were effect sizes taken from ANOVAs, which Vahey et al. (2015) stated that they “equated the relevant statistic [] with *r*2 therefore obtaining *r* using the square root function. However, this conflates with , which is not the case. As a partialized effect size, cannot be converted to *r*, and therefore these conversions are erroneous.

A comprehensive assessment of the reproducibility of the conversions of the individual effects sizes to Pearson’s *r* was not performed on the basis that the above assessments had already determined these effect sizes to contain many errors: 33 effect sizes were common to both Vahey and the reextractions, 23 of Vahey’s effect sizes actually met exclusion criteria, and 123 additional effect sizes meeting inclusion criteria were found.

## Corrected meta-analysis and power analyses

In order to update the core conclusions of Vahey et al. (2015) in light of the numerous errors and non-reproducible analyses observed, a new meta-analysis was conducted, and new power analyses using the meta-estimate were then calculated. Whereas Vahey et al.’s (2015) method of dealing with the non-independence of multiple effect sizes taken from the same study was to average them, research suggests that it is more appropriate to model these dependencies using three-level meta-analyses (i.e., multi-level meta-analyses: Van den Noortgate et al. 2013).

A multi-level random effect meta-analysis with random intercepts for study was therefore employed. I employed the metafor packages’ default settings of a Restricted Maximum Likelihood estimator function and weighting by inverse variance (i.e., rather than *N*, given that inverse variance is a better estimate of precision and represents the contemporary standard). A point estimate and confidence intervals were produced by default. Credibility intervals were calculated following Field REF’s definition (equations XX). In addition to this, prediction intervals were calculated in order to accomplish the stated goal of Vahey et al., i.e., to assess the generalizability of the effect, using a more contemporary method (REF). Prediction intervals share the same goals as Hunter and Schmidt style credibility intervals, as Vahey states them (i.e., they assess the generalisability of the effect). Whereas Hunter and Schmidt style credibility intervals are generally wider and therefore more conservative than confidence intervals – and, notably, they were not in the verification meta-analyses – prediction intervals are always at least as wide as confidence intervals. That is, where Hunter and Schmidt style credibility intervals are based solely on the population variance of the correlations (i.e., tau), prediction intervals as implemented in the metafor package are based on the SEM and tau. Specifically:

[equation]

Results demonstrated a meta effect size *r* = .20, 95% CI [.12, .29], 95% CR [-.04, .44] (see Figure XX for forest plot). 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]). Table XX contains the new power analyses based on this meta-effect size. As can be seen from the table, sample sizes are substantially larger than those recommended by Vahey et al. (ref).

# Discussion

## Summary of findings

The meta-analytic estimate reported by Vahey et al. (2015) was found to have very poor reproducibility on multiple fronts. The original article’s inclusion and exclusion criteria were not comprehensively applied. Many effects that met Vahey et al.’s (2015) inclusion criteria were not included. Conversely, many effects that were included did not meet inclusion criteria, e.g., X% were not criterion effects as they did not involve a second external variable. The averaging of these effect sizes for each article were not computationally reproducible. 13% of cases demonstrating disagreement between the weighted average effect sizes reported in Vahey et al.’s (2015) forest plot and those recalculated from the effect sizes reported in their supplementary materials.

The results of Vahey et al.’s (2015) meta-analysis could not be reproduced despite numerous attempts and approaches. The original power analyses were mostly but not completely reproducible from the original meta-analysis’s results. However, given the lack of reproducibility of the meta-analysis itself, this validity of those power analyses’ results was greatly undermined. After correcting these issues, a new meta-analysis was conducted. Results suggested a meta-effect size less than half that reported in Vahey et al. (2015). This suggested that the individual effect sizes that were missed by Vahey et al.’s (2015) extractions were potentially not missing at random, but were systematically biased towards including larger effect sizes and omitting smaller ones.

New power analyses mirroring the original ones were then conducted using this new meta-analytic effect size. These suggested that a much larger number of participants are required in future IRAP studies than recommended by Vahey et al. (2015). For example, although Vahey et al. (2015) makes sample size recommendations for several different analyses and designs, it is most frequently cited for the specific recommendation of *N* > 37 to detect a a first order correlation (alpha = 0.5, one-tailed, 80% power; e.g., McEnteggart REF, Kavanagh REF). The sample size recommendation based on the updated meta-analytic effect size is *N* > 617. It is worth noting that no published IRAP study to date included a sample size this large, according to a recent systematic review of published research using the IRAP (188 studies in 153 publications, median *N* = 40, range = 4 to 210: Hussey REF).

Implications for users of the IRAP:

Comparison with the IAT. Vahey et al. (2015) used their meta-effect size to compare the IRAP’s clinical criterion validity with that of other implicit measures. Using the original meta-estimate of r = .45, they reported that the IRAP performs “favourably” compared to Implicit Association Test (r = .22 for addiction and r = .30 for non-addiction psychopathologies: Greenwald et al., 2009) and evaluative priming methods (*r*s = .18 to .28: Cameron et al. 2012; Herring et al., 2013; Rooke et al., 2008). Using the results from the corrected meta-analysis (*r* = .22), other implicit measures now appear to be approximately equal to or better than the IRAP.

Broader context: results are consistent with maasen and lakens, who have found that it is very difficult to reproduce the results of meta-analyses. It is important to note that the The failure to reproduce results at almost all stages of the research process underscores the need to

## Limitations

It is important to appreciate that these analyses are intended to highlight the consequences of the issues detected in Vahey et al. (2015) on their core conclusions (i.e., the meta-analysis and power analyses) more than they are intended to accomplish Vahey’s original stated goal of estimating the clinical criterion validity of the IRAP. No attempt was made to include effect sizes from articles other than the XX Vahey et al (REF) considered which were published prior to 2015. The IRAP’s criterion validity would be better assess following guidelines for reproducible and valid practices. This would include (but likely not be limited to) (a) a systematic review followed by a meta-analysis, (b) preregistration of the meta-analytic method, (c) providing all materials (code, data, formulae, item-level rationales for including or excluding effect sizes, etc) (see Lakens et al. 2016).

## Future work

At first glance, the sample sizes recommendations may seem surprising given that IRAP papers typically report sample sizes of around 40 and yet most find some significant results. However, results are not incompatible with this: First, the analysis of IRAP data involves a large ‘garden of forking paths’, or researcher degrees of freedom for the researcher (REF). This has been shown to greatly inflate the false positive rate (REF). 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 likely inflated. To take one concrete example, IRAP studies very frequently report results from a 4X2 mixed within-between ANOVA, but do not apply alpha corrections for the family-wise error rate. Cramer et al. (2015) demonstrate that simply interpreting all three effects from a similar ANOVA (e.g., two main effects and an interaction) serves to increase the false-positive rate by nearly a factor of 3. Future research should therefore (1) conduct meta-methods research (REF) on the IRAP, in order to document researcher flexibility in the presentation, scoring, and analysis of IRAP data; and (2) estimate the impact of these researcher degrees of freedom on the false-positive rate in the IRAP literature, likely via simulation studies.

# References

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

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

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*(3), 527–542. https://doi.org/10.1007/BF03395726

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.

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The Implicit Relational Assessment Procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science*, *2*(3–4), 105–119. https://doi.org/10.1016/j.jcbs.2013.05.002

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data Extraction Errors in Meta-analyses That Use Standardized Mean Differences. *JAMA*, *298*(4), 430–437. https://doi.org/10.1001/jama.298.4.430

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

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

Kavanagh, D., Barnes-Holmes, Y., & Barnes-Holmes, D. (2022). Attempting to Analyze Perspective-Taking with a False Belief Vignette Using the Implicit Relational Assessment Procedure. *The Psychological Record*, *72*(4), 525–549. https://doi.org/10.1007/s40732-021-00500-y

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

Revelle, W. (2009). *An introduction to psychometric theory with applications in R*. Springer Evanston, IL. https://www.personality-project.org/r/book/

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 the Department of Psychology, Ghent University, Belgium, when I began this project. I am now affiliated with the Faculty of Psychology, 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 was my PhD supervisor (Prof Dermot Barnes-Holmes: 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 2017) and the META-REP Priority Program of the German Research Foundation (#464488178) (awarded 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.

1. Two other reviews of the IRAP’s test-retest reliability have also been conducted (Golijani-Moghaddam et al., 2013: r = .49; Greenwald & Lai, 2020: r = .45). However, both estimated test-retest reliability from a very small number of studies (*k*s = 1 and 2, respectively) with very small sample sizes (*N*s = 12 and 25, respectively). Hussey & Drake’s (2020) estimates, which were derived from a larger number of studies and participants (k = 8, N = 318) therefore represent the more precise estimates. In addition, Hussey & Drake (2020) employ both more appropriate methods to estimate reliability (i.e., permutation based internal consistency rather than split-half reliability, and ICC2 rather than Pearson’s *r*. See Hussey & Drake (2020) for further discussion. [↑](#footnote-ref-1)