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”

Ian Hussey

Vahey et al.’s (2015) meta-analysis concluded that the Implicit Relational Assessment Procedure (IRAP) has high clinical criterion validity (meta-analytic = .45) and therefore potential “as a tool for clinical assessment”. Vahey et al.’s (2015) power analyses are frequently cited for sample size determination, especially their heuristic of *N* > 37. This article attempts to verify those results. Results were found to have very poor reproducibility at almost stage of their analytic strategy, with errors biased towards inflating the effect size. 23 of the 56 (41.1%) originally included effect sizes were not criterion effects that were erroneously included. Inspection of the original articles revealed 360 additional effect-sizes that met inclusion criteria and should have been included. The corrected meta-analytic effect size was half the size of the original ( = .22), and the power analyses recommended sample sizes nearly 10 times larger than the original (*N* > 346), which no published original study using the IRAP has met. In aggregate, this seriously undermines the credibility and utility of the original article’s conclusions and recommendations. Vahey et al. (2015) appears to need substantial correction. In particular, researchers should not rely on its results for sample size justification. All code and data available at [osf.io/jg8td](https://osf.io/jg8td/).

Vahey et al.’s (2015) meta-analysis of the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2010) concluded that it possesses good criterion validity and that the IRAP ”demonstrates the potential of the IRAP as a tool for clinical assessment”. 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. While there has been a subsequent debate about the degree to which the IRAP is or is not an “implicit” measure (Barnes-Holmes & Harte, 2022a; Hussey, 2022), and indeed what the term even means (Corneille & Hütter, 2020), these debates are secondary to the fact that the IRAP, and tasks like it, are claimed to be valid measures of individual differences based on sources of evidence such as Vahey et al. (2015).

## 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 (e.g., Kadlec et al., 2023; Lakens et al., 2017).

Second, Vahey et al.’s (2015) article has been well-cited and used to guide subsequent work. At time of writing in July 2024, it has been cited 143 times on Google Scholar, 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 = .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 & 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 reported here in reverse order. Subsequently, I report a new meta-analysis and power analysis using the re-extracted individual effect sizes.

**Table 1. Verifications of power analyses for 80% power.**

|  |  |  | Vahey et al. (2015) | | Verified | New meta-analysis | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test | Tails | Estimated using\* |  | *N* | *N* |  | *N* |
| Pearson’s *r* | One | Point estimate | 0.45 | 29 | 29 | .22 | 126 |
| Pearson’s *r* | One | Lower bound of 95% CI | 0.40 | 37 | 37 | .15 | 273 |
| Pearson’s *r* | Two | Point estimate | 0.45 | 36 | 36 | .22 | 160 |
| Pearson’s *r* | Two | Lower bound of 95% CI | 0.40 | - | 46 | .15 | 346 |
| Independent *t*-test (Cohen’s *d*)\*\* | One | Point estimate | 1.01 | 26 | 26 | .45 | 124 |
| Independent *t*-test (Cohen’s *d*)\*\* | One | Lower bound of 95% CI | 0.87 | 36 | 34\*\*\* | .30 | 270 |
| Dependent *t*-test (Cohen’s *d*) \*\* | One | Point estimate | 1.01 | 8 | 8 | .45 | 32 |
| Dependent *t*-test (Cohen’s *d*) \*\* | One | Lower bound of 95% CI | 0.87 | 10 | 10 | .30 | 69 |
| *Notes:*  \* Researchers often use the point estimate of the meta-effect size. Perugini et al. (2014) recommended the lower bound of the 95% CI instead. Vahey et al. (2015) used both for power analyses. \*\* Necessary conversions from *d* to *r* were not reported in Vahey et al. (2015), but are recalculated here using the effectsize R package’s ‘r\_to\_d’ function.  \*\*\* Discrepancy between result reported in Vahey et al. (2015) and recalculated result | | | | | | | |

**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/jg8td](https://osf.io/jg8td/)). 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. In July 2019, I shared a copy of an earlier version of these verification attempts with the corresponding author, including code, data, and a set of slides outlining my concerns about the credibility of their findings. I have received no contact from the corresponding author since then. No corrections of Vahey et al. (2015) have been issued at time of writing (July 2024), and to the best of my knowledge the authors of Vahey et al. (2015) have made no public statements about these concerns about the credibility of their findings. Even three years after I initially raised these concerns, the senior author of Vahey et al. (2015) has continued to cite the article favourably as evidence for the IRAP’s validity (e.g., Barnes-Holmes & Harte, 2022a, 2022b).

## 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 (Champely, 2016). Table 1 contains the results of both the power analyses reported by Vahey et al. (2015) and those 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, with one exception (difference *N* = 36 vs. 34).

**Figure 1. Weighted-mean effect-sizes and their 95% Confidence Intervals extracted from Vahey et al.’s (2015) forest plot (Figure 1).**

**A graph with numbers and lines

Description automatically generated**

## Meta-analysis

### Issues with the meta-analysis results

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 in their Figure 1 (for and CR) and the text on pages 62-63 (for the CI): = .45, 95% CI [.40, .54], 95% CR [.23, .67].

Prior to any attempt to reproduce these results, it is important to note that the point estimate and confidence intervals are not possible: the upper bound Confidence Interval is +.09 larger than the point estimate , whereas the lower bound Confidence Interval is -.05 smaller. While asymmetric intervals are indeed possible (e.g., when using a transformation such as Fisher’s *r*-to-*z*), such transformations would create an asymmetry in the opposite direction (i.e., a smaller upper interval and larger lower interval). I know of no legitimate way to produce a meta-analytic effect size with asymmetric Confidence Intervals of the type that Vahey et al. (2015) report which are, to the best of my knowledge, mathematically impossible. One plausible explanation is that one or more values are the result of typos. Another plausible explanation is that the Confidence Intervals (reported in text) and the and Credibility Intervals (reported in Figure 1) were obtained from different meta-analyses, employing different data and/or different modelling approaches.

### Issues with the original effect sizes’ Confidence Intervals

Separately, it is also worth inspecting the weighted average effect sizes for each of the component studies reported in Vahey et al.’s (2015) Figure 1. These numerical values were extracted and are reproduced in Figure 1 of this manuscript, and serve as the data for the verification attempts below. The individual effect sizes are labelled as representing weighted *r* values and its Confidence Intervals. As such, the confidence intervals should be symmetrical, but they are not. For example, the effect size reported for Vahey et al. (2010) is *r* = .36, 95% CI [-.24, .76]. This makes the upper bound +.40 larger than the point estimate whereas the lower bound is -.60 from the point estimate. The asymmetry is therefore in the opposite direction to that in the reported meta-analytic effect size’s Confidence Interval, and is in principle compatible with the a transformation such as Fisher’s *r*-to-*z* having been applied. As I discuss later on in meta-analysis reproduction attempt 4, it is likely that the original figure mislabels what are weighted average Fisher’s *r*-to-*z* transformed values asweighted average *r* values.

**Figure 2. Discrepancies between the data in the original funnel plot vs. forest plot.**

A diagram of a point

Description automatically generated

Note: This plot was created from Vahey et al.’s (2015) results reported in their forest plot (their Figure 1) vs. their funnel plot (their Figure 2). The black dot refers to the location of one weighted mean effect size according to their forest plot. The red dot asterisk refers to its approximate location in their funnel plot (their Figure 2). Circles refer to data points that matched between the two plots.

### Data in the funnel plot does not match the forest plot

Vahey et al. (2015) also reported their weighted average effect size estimates in a funnel plot (see their Figure 2). However, when I created a funnel plot from the results reported in their forest plot, one of the data points did not match their funnel plot. See Figure 2 which illustrates this discrepancy. This suggests that the funnel plot and the forest plot were created from slightly different data sets. It is unclear which one represents the ‘correct’ data set (especially in light of section on ‘average effect sizes’ that I discuss later), but this speaks to the broader pattern of non-reproducibility and internal inconsistencies in Vahey et al.’s (2015) results.

### Implementation of the meta-analysis

Vahey et al. (2015) stated that they employed a Hunter & Schmidt style meta-analysis and cited Field & Gillett (2010). In personal correspondence with Vahey, he stated it should in principle be possible to reconstruct their analytic strategy and results from the code provided by Field & Gillett (2010), but he declined to share the actual code used in Vahey et al. (2015). In practice, it was surprisingly difficult to reconstruct what was done because of multiple discrepancies both (a) between Field & Gillett (2010) and Vahey et al. (2015) and (b) within Vahey et al. (2015) itself, as I will now discuss.

Field & Gillett (2010) describe and provide code for two different ways of conducting meta-analyses: a Hedges style “basic” meta-analysis and a Hunter and Schmidt style psychometric meta-analysis. Despite Vahey et al. (2015) stating that they applied the Hunter and Schmidt approach, multiple features of this approach are missing from their results (e.g., no correction for the reliability of the measures used to generate the correlation), and other features that are unique to the Hedges’ style meta-analysis were reported (e.g., calculation of Confidence Intervals, e.g., p. 63; apparent use of Fisher’s *r*-to-*z* transformation, e.g., Figure 2).

In order to attempt to verify the original results, I downloaded the accompanying materials for Field & Gillett (2010) from Field’s website. This contained two different SPSS scripts, “Meta\_Basic\_r.sps” and “h\_s syntax.sps”. Confusingly, both scripts contain code to produce Hunter and Schmidt style meta-analyses, with the former also producing a Hedges style ‘basic’ meta-analysis.

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 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. Additionally, and more problematically, 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) state 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 and 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 article 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.

### Definitions of different intervals

Vahey et al. (2015) report both Confidence Intervals (CI) and Credibility Intervals (CR which attempt to estimate the generalizability of the meta-effect size (Field & Gillett, 2010; Hunter & Schmidt, 2004).

Vahey et al. (2015) state that such “Credibility Intervals are generally wider and thus more conservative than corresponding Confidence Intervals” (p.61), however this is not the case: Confidence Intervals and Credibility Intervals have different estimands and therefore the two have no correlation. Confidence Intervals quantify the precision of the estimate given sampling error (i.e., within-study variance, ), whereas Credibility Intervals are a function of between-study variance (: see Field & Gillett, 2010, equations 2, 3, 4, and 5). A third type of interval, Prediction Intervals (PI) take both into account, and are often reported for meta-analyses (e.g., within the metafor R package). It is true that PIs are at least as wide as Confidence Intervals, and however this is not because they are more ‘conservative’ that Confidence Intervals but because they quantify a different properties under different assumptions. It is unclear whether this discrepancy in Vahey et al. (2015) was due to (a) a misinterpretations of Credibility Intervals, or (b) whether they actually calculated PIs but mislabelled them as Credibility Intervals.

In order to attempt to resolve this for the purpose of verification, it is useful to define all three to highlight the differences between them:

Where is the weighted average effect size, is the estimated within-study variance (i.e., the square of the Standard Error of ), and is the estimated between-study variance (heterogeneity).

An important point to appreciate regarding Credibility Intervals is that when between study heterogeneity is zero ( = 0), the CR interval width will also be zero, as is the case in the results of the verification attempts reported later on. This also follows from Field & Gillet’s (2010) equations 2, 3, 4 and 5 (note that they use slightly different notation), which define the variance in the estimate of population correlations as the variance of sample effect sizes (which Vahey et al. 2015 denote as ) minus the sampling error variance. As such, if the sampling error variance is found to be larger than the variance in the sample effect sizes, then will be negative, and 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 code 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 both wider than Credibility Intervals, contrary to Vahey et al.’s (2015) claim, and indeed where the Credibility Intervals are implausibly narrow (i.e., 0).

**Table 2. Verification attempts for the meta-analysis**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | 95% CI | | 95% CR | | 95% PI | |
| Label | Implementation | Modifications from original code |  | Lower | Upper | Lower | Upper | Lower | Upper |
| Vahey et al. (2015) | Vahey stated in personal corrispondance that they employed Field & Gillett’s code. | Unkown. | .45 | .40 | .54 | .23 | .67 | - | - |
| Verification attempt 1 | Hunter & Schmidt method using Field & Gillett’s (2010) "h\_s\_syntax.sps" | All reliabilities set to 0. | .47 | .20 | .74 | .47 | .47 | - | - |
| Verification attempt 2 | Hunter & Schmidt method using Field & Gillett’s (2010) “Meta\_Basic\_r.sps” | Set variance in population correlations to zero if it is negative, as in "h\_s\_syntax.sps" | .46 | - | - | .46 | .46 | - | - |
| Verification attempt 3 | Hunter & Schmidt method using Viechtbauer’s (2022) implementation in R and metafor. | Credibility intervals implemented using Field & Gillett’s (2010) equations 2 to 5. | .47 | .40 | .54 | .47 | .47 | .40 | .54 |
| Verification attempt 4 | Mix of Hunter & Schmidt and Hedges methods using Viechtbauer’s (2022) implementation in R and metafor. | Credibility intervals implemented using Field & Gillett’s (2010) equations 2 to 5. Fisher’s *r*-to-*z* transformations and *z*-to-*r* back transformations. | .47 | .40 | .54 | .47 | .47 | .40 | .54 |
| *Notes:* CI = Confidence Interval. CR = Credibility Interval. PI = Prediction Interval. Although PIs were not reported in Vahey et al. (2015), they are were calculated where possible in the verification attempts to see if they corrisponded with the original CRs on the basis that the CRs could have been mislabelled. Cells shaded in grey match those reported in Vahey et al. (2015) within ±.01. | | | | | | | | | |

### Verification attempt 1

The first verification attempt employed Field’s “h\_s syntax.sps” script which implements the Hunter & Schmidt style meta-analysis that Vahey et al. (2015) cite and which Vahey stated in personal correspondence that they used. Copies of all original and altered scripts are available in the supplementary materials. The default 80% Credibility Interval widths were changed to 95%, as Vahey et al. (2015) reported using the latter width.

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 (Field & Gillett, 2010; Hunter & Schmidt, 2004). Although Hunter & Schmidt (2004) did describe what they referred to as a “bare-bones” meta-analysis that did not perform this deattentuation based on reliability, for Field’s “h\_s syntax.sps” script to run it requires the researcher to provide reliability values for both of the measures that produced each effect size. Vahey et al. (2015) do not report any extracting or estimating reliabilities or deattenuating the effect sizes 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 2 presents the meta-analysis effect sizes estimates reported by Vahey et al. (2015) as well as the results of all verification analyses conducted here.

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 ( = 0.02), but more than could be accounted for by rounding or other erroneous methods such as truncation. Confidence Intervals were nearly four times wider in the verification analysis than the original results. Credibility Intervals were had zero width (i.e., because was negative, interval width was therefore set to zero), whereas they were wider than the Confidence Intervals in the original analysis. Note that setting the reliabilities to other values (e.g., the common reliability thresholds such as .70 or otherwise), would serve to inflate both the meta-analysis effect size and intervals. As such, additional verifications using alternative reliability values were not useful to explore.

### Verification attempt 2

The second verification attempt employed Field’s “Meta\_Basic\_r.sps” script, which implements a Hedges’ style “basic” meta-anlysis. Note that, unlike the “h\_s syntax.sps” script, this script does not contain code to calculate Confidence Intervals. One necessary alteration was made to the code: if was negative it was set to zero to produce a Credibility Interval width of 0, as in the “h\_s syntax.sps” script. Without this alternation, if was negative the script would fail to run.

This verification attempt also did not reproduce the original results. The point estimate was off by a small amount ( = 0.01). This is more than could be accounted for by common methods of rounding, but could be obtained via (erroneous) truncation. Nonetheless, this value was the closed I could achieve across attempts and, in the grand scheme of things, is close to the reported value. However, the Confidence Intervals were nearly four times wider than the original results. In addition, the Credibility Intervals again had zero width (i.e., because was negative and interval width was therefore set to zero) and therefore greatly differed from the original results.

### Verification attempt 3

In a third verification attempt, I switched from using Field and Gillett’s SPSS code to Vichtbauer’s (2022) implementation of a Hunter & Schmidt style meta-analysis that was written in R and the metafor package (Viechtbauer, 2010, 2024). This provided new avenues to attempt to reproduce the original results in a programming language I was more familiar with, allowing me to more quickly try a wide variety of variations on a given attempt. However, this also opened up the possibility that non-reproduction was due to implementational differences between languages. Field & Gillett’s (2010) equations 2-5 were used to implement Credibility Intervals. In this attempt, 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 3. Confidence Intervals around individual effect sizes were symmetric and therefore did not reproduce the original plot.

**Figure 3. Forest plot for meta-analysis verification attempt 3.**



**Figure 4. Forest plot for meta-analysis verification attempt 4.**

****

### Verification attempt 4

In the fourth meta-analysis verification attempt, I applied 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 attempt 3, therefore the original meta-analysis results were not reproduced.

However, the forest plot associated with this attempt did reproduce the Confidence Intervals around the individual effect sizes from Vahey et al.’s (2015) original forest plot (see their Figure 1 and this manuscript’s Figures 1 and 3), suggesting that Vahey et al. (2015) employed these transformations but did not report them. This under-reported data transformation also implies a second form of underreporting: Vahey et al. (2015) reported employing a Hunter & Schmidt style meta-analysis, but this implies that they diverged from this strategy by also applying Hedges style data transformations (in addition to not applying Hunter & Schmidt style corrections for reliability). While this reproduction of the original individual effect sizes and their Confidence Intervals gets us one step closer to understanding the original analytic strategy, it nonetheless does not reproduce the meta-analysis results.

### Summary of attempts

A large number of small variations on the attempts that are reported here were also tried. For example, alternative values for reliability estimates in attempt 2, and not back-transforming the *z* values in attempt 4. I also tried a number of purposeful mistakes, such as miscalculating Credibility Intervals based on plausible mathematical and coding errors. No attempt successfully reproduced the originally reported results.

Confidence Intervals around individual effect sizes in the original forest plot were only reproduced when Fisher’s *r*-to-*z* transformations were applied (verification attempt 4) and not when they weren’t (verification attempt 1-3). 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.

I noted previously in the “Issues with the meta-analysis results” section that the original meta-analysis point estimate is incompatible with the reported Confidence Intervals. Interesting, if we assume that (a) the originally reported point estimate is incorrectly reported but the Confidence Intervals are correctly reported, and (b) that the Confidence Intervals are symmetrical, this would imply that a correct point estimate of .47 (i.e., at the half way point between the intervals). An point estimate of .47 combined with Confidence Intervals of [.40, .54] was reproduced in verification attempts 3 and 4. However, this does not imply that the original results are merely the result of a typo in the point estimate, as (a) the Credibility Intervals in verification attempts 3 and 4 are very different to the original results, and (b) more confusingly, these results were produced only by Viechtbauer’s (2022) implementation of the analysis in R and metafor, but not using the scripts that Vahey et al. (2015) report having used. It therefore remains to be unclear how Vahey et al. (2015) analysed their data or obtained all their results.

## Average effect sizes

In order to attempt to retrace the steps involved in the original analysis, I then noted that Vahey et al. (2015) reported that the 15 weighted average effect sizes they used in their meta-analysis were calculated from 46 individual effect sizes and degrees of freedom taken from 15 studies. Vahey et al. (2015) reported the individual effect sizes and degrees of freedom in their supplementary materials. I therefore attempted to verify the weighted averages by recalculating them using using Vahey’s strategy of weighting by degrees of freedom. Results were not fully 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 ( = -.02 and .05). On the other hand, given the simplicity of these calculations, the discrepancy is difficult to understand. Both instances with discrepancies came from articles that were first-authored by co-authors of Vahey et al. (2015), suggesting that they were not unfamiliar with the original studies.

## Individual effect sizes

Next, I attempted to retrace the next step involved in the original analysis: the extraction of effect sizes from original articles. This involves (a) the correct application of inclusion criteria in terms of correct inclusions and the absence of incorrect omissions, and (b) the extraction and conversion of effect sizes.

### Assessment of incorrect inclusions

Lakens et al. (2016) argued that “incorrect inclusion” is a common type of error in meta-analysis. That is, the 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). Unfortunately, neither the original article nor its supplementary materials provided data for each extracted effect size regarding which specific clinical condition that was targeted by the IRAP and the criterion variable, or the “specific empirical literature” that Vahey et al. (2015) used to justify the inclusion of each criterion.

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 (American Psychological Association, 2024). Using the descriptions in Vahey et al.’s (2015) 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. While the clinical relevance of specific effects might be more subjective, the involvement of a criterion variable other than the IRAP can be assessed objectively. Worryingly, 23 of the 56 effect sizes (41%) employed by Vahey et al. (2015) were found to involve no external variable (i.e., they refer only to a reaction time differential between the IRAP block types, i.e. from a one-sample *t*-test), and were therefore not suitable to be included in a meta-analysis of the IRAP’s criterion validity. A large degree of incorrect inclusion error was therefore detected in Vahey et al.’s (2015) effect sizes.

### Assessment of incorrect exclusions

In addition to incorrection inclusions, it is equally plausible that effect sizes that would have met inclusion criteria were erroneously not included. I therefore re-examined the same 15 articles as Vahey et al. (2015) drew their effect sizes and searched for other effect sizes that met their inclusion criteria. 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 (e.g., correlations where only the statistically significant estimates were reported). Where necessary, I contacted the authors for the individual articles to obtain additional estimates or data. For example, if non-significant correlations were reported as merely “other correlations were non-significant”, these effect sizes were obtained where possible. Two independent raters the rated each effect for clinical relevance using Vahey et al.’s (2015) criteria. Agreement was found in 90% of cases (Cohen’s Kappa = 0.87, *p* < .001). 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 the 33 included by Vahey et al. (2015), after I excluded the 23 non-criterion effects (as discussed previously). This suggests that Vahey may have failed to include 85.3% of the effect sizes that met their inclusion criteria, representing a potentially serious source of incorrect non-inclusion error.

Note that these extractions are not exhaustive: some authors of original studies who were reported as having replied to Vahey et al.’s (2015) requests for additional information did not reply to my requests, perhaps due to the passage of time the ‘half-life’ of data.

These effect sizes were converted to Pearson’s *r* for use in a new meta-analysis that I discuss later. The specific methods of conversion are documented 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., Gøtzsche et al., 2007; Maassen et al., 2020). In their supplementary materials, Vahey et al. (2015) provided explanations and references for how individual effect sizes were converted to Pearsons *r*. However, inspection of those 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 : 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 several errors (e.g. related to incorrect inclusion).

## Corrected meta-analysis and power analyses

In order to understand the compound impact of the various errors on the conclusion the conclusions of the meta-analysis, I fit a new meta-analysis to the 156 effect sizes re-extracted from the original articles. I then used the meta-analysis effect results to calculate new power analyses. 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).

Results demonstrated a meta effect size = .22, 95% CI [.15, .29], 95% CR [.22, .22], 95% PI [-.01, .42] (see Figure 4 for forest plot). Based on the non-overlap of their Confidence Intervals, this estimate is significantly smaller than the effect size reported in Vahey et al. (2015), i.e., = .45, 95% CI [.40, .54], 95% CR [.23, .67]).

Table 1 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. (2015). For example, whereas the original abstract includes the recommendation that IRAP studies should employ “*N* of [at least] 29 to 37”, the update numbers for the same tests are *N*s of at least 126 to 273. Power analyses for the more common and less liberal two-sided test for Pearson’s *r* correlations would require *N*s of at least 160 to 346.

**Figure 5. Forest plot for the new meta-analysis.**

****

# Discussion

The results of Vahey et al. (2015) could not be verified at multiple different stages in the analysis, and multiple errors and discrepancies were detected. The original article’s inclusion and exclusion criteria were not consistently 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., 41.1% were not criterion effects as they did not involve an external variable. These inconsistencies in the application of the inclusion and exclusion strategy biased towards including larger effect sizes and omitting smaller ones. The averaging of these effect sizes for each article were not computationally reproducible in 13% of cases. The results of the meta-analysis could not be reproduced despite numerous different attempts and approaches. The original power analyses were mostly reproducible, however, given the lack of reproducibility of the meta-analysis itself, this validity of those power analyses’ results based on that meta-analysis estimate was fundamentally undermined.

This lack of reproducibility are consistent with what has been found elsewhere for meta-analyses: errors in data extraction and conversion are common, results are frequently not reproducible, and this is hindered by the unavailability of data and code (Gøtzsche et al., 2007; Kadlec et al., 2023; Lakens et al., 2016, 2017; Maassen et al., 2020).

After correcting the above issues, a new meta-analysis was conducted in order to convey the combined impact of these issues on the conclusions. Results suggested a meta-effect size or = .22, less than half that reported in the original article ( = .45). Vahey et al. (2015) stated that, according to their results, the IRAP’s criterion validity compares “favourably” to the other popular implicit measures such as the Implicit Association Test ( = .22 for addiction and = .30 for non-addiction psychopathologies: Greenwald et al., 2009) and evaluative priming methods (s = .18 to .28: Cameron et al. 2012; Herring et al., 2013; Rooke et al., 2008). Without endorsing the updated meta-analysis, by Vahey et al.’s (2015) line of argument, the current results suggest that the IRAP is therefore on par with other such measures rather than superior to them. This also brings the average criterion association observed for the IRAP closely in line with the average correlation observed across social and personality psychology (i.e., around *r* = .2: Hemphill, 2003; Hussey, 2023; Richard et al., 2003).

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 first order correlation (alpha = 0.5, one-tailed, 80% power; e.g., Kavanagh et al. 2022). The sample size recommendation based on the updated meta-analytic effect size is *N* > 273 for a one-tailed correlation, and 346 for the much more commonly used and less liberal two-tailed correlation. It is worth noting that between 0% and 2.1% of published original research using the IRAP has included sample size meeting these criteria, according to a recent systematic review of IRAP research published between 2006 and 2022 (188 studies in 150 publications, median *N* = 41, range = 9 to 210: Hussey, 2023).

With that said, it is important to acknowledge that the primary reason to calculate a new meta-analysis was to illustrate the combined impact of the errors on the results rather than to endorse the results of this new meta-analysis or power analysis. In my opinion, Vahey et al.’s (2015) approach of taking different types of effects between the IRAP and other criterion tasks is fundamentally flawed as it combines apples with oranges on multiple different fronts. All the following effects were meta-analysed together: effect sizes representing the magnitude of the compatibility effect on the IRAP itself, interaction effects between IRAP trial types and group allocations, and correlations between IRAP trial types and criterion tasks. In doing so, different types of IRAP data were combined as one: data from single trial-types, overall effects for the whole task, and effects averaging the trial types in different ways. Lastly, effects treating the IRAP as the dependent variable, the independent variable, and purely associative effects were combined as one. It is exceptionally difficult to know what the resulting meta-analysed effect size is an estimate *of*, i.e what the estimand is, and whether it applies to the type of effect that a researcher may wish to observe in their own future work. For example, to what degree is the interaction effect between a depression IRAP’s trial types and high vs. low experiential-avoidance group informative to a separate study on the correlation between self-reported self-esteem and a self-esteem IRAP in a prisoner population? I argue that this apples-with-oranges approach is erroneous and leads to misleading conclusions, but the point of the verifications presented here is to highlight that Vahey et al.’s (2015) erroneous analytic approach was also erroneously implemented.

## Limitations

It is possible that these verification analyses themselves contain errors. The purpose of a verification report is to attempt to independently verify the results of presented in the original article, and do not represent the last word on error detection. Equally, perhaps there is some way to reproduce the original results (e.g., of the meta-analysis) in a way that I have not considered, and my not being able to reproduce them does not necessarily mean they are non-reproducible in some absolute manner. Verification attempts are in general enhanced with access to the original code. Unfortunately, however, the first author of the original article declined to share their code.

## Conclusions

The results of Vahey et al. (2015) were found to have poor reproducibility at almost stage of their analytic strategy. In aggregate, these seriously undermine the credibility and utility of the conclusions and recommendations of the original article. Recalculated results suggest that the meta-effect size was less than half the original result ( = .45 vs. .22), and sample size recommendations were more than 15 times as large (minimum *N* = 37 vs. 346). Vahey et al. (2015) may therefore require substantial correction, and researchers should not use it for sample size planning.

# References

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

American Psychological Association. (2024). *APA Dictionary of Psychology*. https://dictionary.apa.org/criterion-validity

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

Barnes-Holmes, D., & Harte, C. (2022a). The IRAP as a Measure of Implicit Cognition: A Case of Frankenstein’s Monster. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-022-00352-z

Barnes-Holmes, D., & Harte, C. (2022b). Relational frame theory 20 years on: The Odysseus voyage and beyond. In *JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR* (Vol. 117, Issue 2, pp. 240–266). WILEY. https://doi.org/10.1002/jeab.733

Champely, S. (2016). *pwr: Basic Functions for Power Analysis* [Computer software]. https://CRAN.R-project.org/package=pwr

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

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665–694. https://doi.org/10.1348/000711010X502733

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

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*(1), 78–79. https://doi.org/10.1037/0003-066X.58.1.78

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Hussey, I. (2022). Reply to Barnes-Holmes & Harte (2022) “The IRAP as a Measure of Implicit Cognition: A Case of Frankenstein’s Monster”. *PsyArXiv*. https://doi.org/10.31234/osf.io/qmg6s

Hussey, I. (2023). A systematic review of null hypothesis significance testing, sample sizes, and statistical power in research using the Implicit Relational Assessment Procedure. *Journal of Contextual Behavioral Science*, *29*, 86–97. https://doi.org/10.1016/j.jcbs.2023.06.008

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

Kadlec, D., Sainani, K. L., & Nimphius, S. (2023). With Great Power Comes Great Responsibility: Common Errors in Meta-Analyses and Meta-Regressions in Strength & Conditioning Research. *Sports Medicine*, *53*(2), 313–325. https://doi.org/10.1007/s40279-022-01766-0

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., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), 24. https://doi.org/10.1186/s40359-016-0126-3

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

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

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

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

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

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2022). *Hunter and Schmidt Method*. https://www.metafor-project.org/doku.php/tips:hunter\_schmidt\_method

Viechtbauer, W. (2024). *metafor: Meta-Analysis Package for R* (4.6-0) [Computer software]. https://CRAN.R-project.org/package=metafor

1. Two other reviews of the IRAP’s test-retest reliability have also been conducted (Golijani-Moghaddam et al., 2013: = .49; Greenwald & Lai, 2020: = .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)