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 $\bar{r} = .45$) and therefore "the potential of the IRAP as a tool for clinical assessment" (p. 64). 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 every stage of the data extraction and analysis with errors generally biased towards inflating the effect size. The reported meta-analysis results were found to be mathematically implausible and could not be reproduced despite numerous attempts. The reproduction attempt with the closest compressive set of results required making two serious errors: using the wrong data set and mislabelling Confidence Intervals as Credibility Intervals and vice versa. Multiple internal discrepancies were found in the effect sizes such as between the forest plot and funnel plot, and between the forest plot and the supplementary data. 23 of the 56 (41.1%) individual effect sizes were not actually criterion effects and did not meet the original inclusion criteria. Inspection of the original articles revealed 360 additional effect sizes that met inclusion criteria that should have been included. A new meta-analysis was calculated to understand the compound impact of these errors. The effect size was half the size of the original ($\bar{r} = .22$), 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. A list of suggestions for error detection in meta-analyses is provided. All code and data available at osf.io/jg8td.

There is now a growing literature on post-publication scientific error detection in meta-analyses. Much of this has focused on quantifying the prevalence of issues that undermine the validity of meta-analysis results, such as errors in effect size extraction (e.g., Gøtzsche et al., 2007; Lakens et al., 2017; Maassen et al., 2020), or indicators of reproducibility such as the reproducibility of the systematic search strategy, specification of the exact method to compute effect sizes, choice of weightings and estimator function, and sharing of data and code (López-Nicolás et al., 2022). More recently, this has been supplemented with work that is explicitly focused on error detection that has the goal of examining what features of a meta-analysis can be checked and how, and where meta-analyses tend to make errors (Kadlec et al., 2023). This article continues in this vein: following the logic of error detection tools for original research articles (e.g., Heathers et al., 2018), it focuses on features of meta-analyses that are either informative but often overlooked or that repeat information. Both provide vectors for error detection. Indeed, these principles of examining overlooked repeated information to assess the trustworthiness of published work are now being integrated into Cochrane's systematic review process (Wilkinson et al., 2023). The intended meta-scientific utility of this manuscript is therefore to provide a relatively fine-grain description of what information was inspected for errors and how, in the hope that some of these methods of verification allow other meta-analyses to be more efficiently and effectively inspected for errors.

Briefly, Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2006, 2010) is a reaction-time based measure that has been used in over 150 publications (Hussey, 2023). Typical implementations of the IRAP involve presenting "sample" words or images at the top of the screen and "target" words or images in the middle of the screen. Participants must respond with one of two response options that involve opposing relational terms, such as True/False or (more rarely) similar/different, which are assigned to a left vs. right response key on the keyboard. Participants complete pairs of blocks of trials, most commonly three pairs of blocks of 'consistent' vs. 'inconsistent' trials, with 24 trials per block. Each trial requires the participant to provide a specific response to advance to the next trial. The other incorrect response causes corrective feedback to be presented on screen, most commonly a red X. The required response swaps between blocks. For example, a disgust IRAP could employ disgusting vs. pleasant images as sample stimuli and positive vs. negative words as target stimuli. On the 'consistent' blocks, when presented with a disgusting image and the words "I feel sick", the required response would be "True". On 'inconsistent' blocks, the required response would instead be "False". Participants are instructed and trained to maintain accuracy and speed criteria in practice blocks (e.g., median reaction time $<$ 2000ms and percentage accuracy $>$ 80%) before being presented with a fixed number of test blocks. Reaction time data from the test blocks are typically scored using a version of the Greenwald D metric developed for the Implicit Association Test (Greenwald et al., 2003; Hussey et al., 2015; although for issues with D and a more robust alterantive see De Schryver et al., 2018) to quantify the IRAP effect: the relative speed at which participants emit one pattern of relational responses relative to the other. This effect is sometimes used as a metric of (relational) implicit attitudes or beliefs and at other times is used in the study of the dynamics of relational responding.

Vahey et al.'s (2015) meta-analysis concluded that the IRAP possesses good clinical criterion validity and that results "demonstrates the potential of the IRAP as a tool for clinical assessment" (p. 64). Based on a non-systematic review followed by a meta-analysis, those authors (a) provided an estimate of the association between IRAP effects and clinically relevant criterion variables, (b) reported that the IRAP compares favorably to other popular implicit measures, including the Implicit Association Test (Greenwald et al., 1998), and (c) used their meta-analyzed 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

In addition to the meta-scientific utility of doing so discussed previously, there are at least four rationales for performing a verification of Vahey et al.'s (2015) results. 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 the individual effect sizes reported in meta-analyses of psychology research could not be reproduced from the original articles. This was attributed to 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 the 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 in lieu of a power analysis for that study (e.g., Bast & Barnes-Holmes, 2015; Farrell & McHugh, 2017; Leech et al., 2018; Maloney & Barnes-Holmes, 2016; Power et al., 2017; see supplementary materials for supporting quotes from each). 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) provided 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, Vahey et al. (2015) may have been modest when they reported that their results imply the IRAP compares "favorably" with other implicit measures. In fact, Vahey et al.'s (2015) reported meta-analytic effect size of $\bar{r}$ = .46 would place it in >90th percentile of all meta-analytic effect sizes reported in psychology (Richard et al., 2003). Given that the IRAP is a reaction-time-based measure, which and such measures are inherently prone to noise and therefore poor reliability (as I discuss in the next section), the original result implies that the IRAP is a truly remarkable measure to be able to correlate so highly with a range of clinical criterion measures. Or, something is amiss with Vahey et al. (2015).

Fourth, in light of 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 $\bar{r}$ = .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 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 x and y ($r_{xy}^{observed}$) is a function of the true correlation between the variables (r_{xy}^{true}) and the reliability of both measures (their self-correlation r_{xx} and r_{yy}). This can be quantified via the Attenuation Formula derived from classical test theory (Revelle, 2009, equation 7.3):

$$r_{xy}^{true} = \frac{r_{xy}^{observed}}{\sqrt{r_{xx} r_{yy}}} \quad (1)$$

Two of the variables in Equation 1 already have empirical estimates. First, Vahey et al.'s (2015) estimate of the observed correlation between the IRAP and criterion variables was r_{xy}^{true} = .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 (α = .27) and test-retest (ICC$_2$ = .18) are extremely low (Hussey & Drake, 2020).[1] This leaves two remaining variables, the IRAP's criterion validity after adjusting for measurement error (r_{xy}^{true}) and the criterion tasks' mean reliability (r_{yy}). Both of these variables share the same constraint: as correlations, their value cannot be below -1 or above 1. For the moment, if we assume that the criterion tasks' mean reliability is very good (r_{yy} = 0.90). This would imply that the lower limit of the IRAP's true criterion validity after adjusting for measurement error is somewhere between (a) implausibly high, r_{xy}^{true} = .91 (when using the estimate of internal consistency), and (b) mathematically impossible, r_{xy}^{true} = 1.12 (when using the estimate of test-retest reliability). 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 r_{yy} = .70, r_{xy}^{true} = 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 the IRAP's 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 the IRAP's criterion validity). Ultimately it will be up to the research community to determine whether our analyses in Hussey & Drake (2020) are sound, and we provided open data and code to aid others in inspecting our work for errors. Because I am confident in our results reported there, I am instead motivated 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 the 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.

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). In the process of conducting this verification attempt, I contacted the corresponding

[1] Two other reviews of the IRAP's test-retest reliability have also been conducted (Golijani-Moghaddam et al., 2013: $\bar{r}$ = .49; Greenwald & Lai, 2020: $\bar{r}$ = .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 ICC$_2$ rather than Pearson's *r*. See Hussey & Drake (2020) for further discussion.

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

			Vahey et al. (2015)		Verified	New meta-analysis	
Test	Tails	Estimated using*	$\bar{r}$	N	N	$\bar{r}$	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 the result reported by Vahey et al. (2015) and the recalculated result

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 the 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 five years after I initially raised these concerns, the senior author of Vahey et al. (2015) has continued to cite the article favorably 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), the 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).

Meta-analysis

Issues with the meta-analysis results

Vahey et al. (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 $\bar{r}$ and CR) and the text on pages 62-63 (for the CI): $\bar{r}$ = .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 $\bar{r}$, 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 and they 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 $\bar{r}$ and Credibility Intervals (reported in Figure 1) were obtained from different meta-analyses, employing different data and/or different modeling approaches.

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

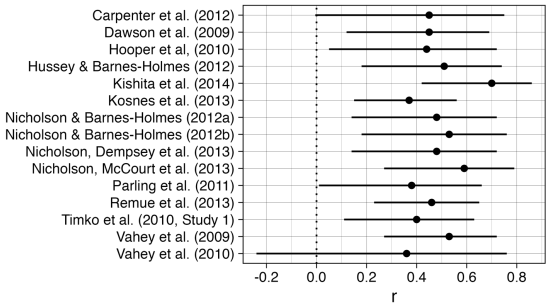
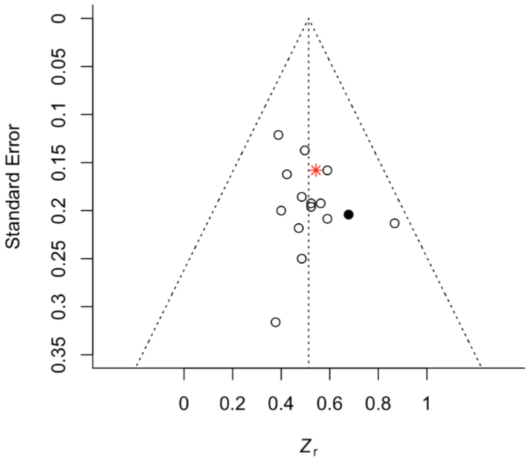


Figure 2. Discrepancies between the data in Vahey et al.'s (2015) forest plot (their Figure 1) vs. funnel plot (their Figure 2).



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 match between the two plots.

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 labeled 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 a transformation having been applied (e.g., Fisher's r-to-z). As I discuss later 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 as weighted average r values.

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). This duplication of data between two plots provided a vector for error detection. 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 original funnel plot and forest plot were created from slightly different data sets. It is unclear which one represents the 'correct' data set (especially in light of the 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) and its accompanying scripts that are maintained by Fields on his website (https://www.discoveringstatistics.com/repository/fieldgillett/how_to_do_a_meta_analysis.html). 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); (b) within Vahey et al. (2015) itself; (c) between Field and Gillett's (2010) descriptions and Field's actual code implementations, and (d) between the different implementations of the Hunter & Schmidt style meta-analysis between Field's two different scripts that are associated with Field & Gillett (2010). I will discuss each of these. Trying to unravel this was extremely challenging, to a degree that is difficult to convey.

Field & Gillett (2010) describe two different ways of conducting meta-analyses: a Hedges and colleagues 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. This becomes more apparent when examining the metrics returned by Field's accompanying SPSS scripts for Field & Gillett (2010): "Meta_Basic_r.sps" and "h_s syntax.sps". To complicate things, both scripts contain code to produce a Hunter and Schmidt style meta-analysis, with the former also producing a Hedges and colleagues style 'basic' meta-analysis. Table 2 catalogs the metrics reported in Vahey et al. (2015) and those nominally calculated by the scripts, based on an inspection of their code. Table 2 illustrates that neither script's features (e.g., use of corrections, transformations, and reliability estimates) nor outputs (point estimates and types of intervals, which I discuss in detail later) correspond with the results reported in Vahey et al. (2015).

Specifically, Vahey et al. (2015) likely (but not definitely, or perhaps not consistently across analyses) used Fisher's r-to-z transformations (e.g., due to asymmetric Confidence Intervals in the weighted mean effect sizes in their Figure 1, and the reference to this transformation in Figure 2), and reported both Confidence Intervals and Credibility Intervals. In

Table 2. Alignment between the results reported in Vahey et al. (2015) and Field's SPSS scripts accompanying Field & Gillett (2010)

Source	$\bar{r}$	CIs	CRs	Adjustments & Transformations	Employs reliability estimates
Vahey et al. (2015)	Yes	Yes	Yes	Likely Fisher's r-to-z*	No
Hunter & Schmidt meta-analysis via Field's "h_s_syntax.sps"	Yes	Yes	Yes	No	Yes
Hunter & Schmidt meta-analysis via Field's "Meta_Basic_r.sps"	Yes	No	Yes	Overton corrections, Fisher's r-to-z for $\bar{r}$ but not CRs	No
Hedges and colleagues meta-analysis via Field's "Meta_Basic_r.sps"	Yes	Yes	No	Overton corrections, Fisher's r-to-z	No
Hunter & Schmidt meta-analysis via a modification of Field's "Meta_Basic_r.sps" to remove apparently erroneous Overton correction**	Yes	No	Yes	Fisher's r-to-z	No

Notes: Shaded cells match the requirements to be capable of producing the same type of output as reported in Vahey et al. (2015), agnostic to whether the numerical results match those reported in Vahey et al. (2015).

* Vahey et al. (2015) did not explicitly state using any transformations. However, their forest plot's (their Figure 1) individual effect sizes have asymmetric confidence intervals implying a transformation; their funnel plot (their Figure 2) is labeled as employing Fisher's r-to-z transformed values; and the method they state they followed employs Fisher's r-to-z transformations.

** Field & Gillett (2010) describe the Hedges and colleagues style meta-analysis as involving an Overton correction but not the Hunter and Schmidt style meta-analysis. However, their script applies the correction to both, misaligning the code with the article. In order to correct this apparent issue and attempt to more closely align the code with the described method, I therefore removed the Overton correction from this version.

contrast, “h_s syntax.sps” script’s Hunter and Schmidt style meta-analysis does not calculate Confidence Intervals; the “Meta_Basic_R.sps” script’s Hunter & Schmidt style meta-analysis does not report Confidence Intervals; and its Hedges and colleagues style meta-analysis does not use Fisher’s *r*-to-*z* transformations or report Credibility Intervals. In addition to this, the “h_s syntax.sps” script requires the researcher to provide reliability estimates for both variables in each correlation (i.e., the reliabilities r_x and r_y for the correlation r_{xy}) in order to correct the effect sizes for attenuation. Vahey et al. (2015) did not report extracting or using reliability estimates in this way in their article or supplementary materials.

Based on these facts, we could conclude that Vahey et al. (2015) did not in fact employ the Hunter & Schmidt style meta-analysis specified in Field & Gillett (2010), as stated. It is possible they ran more than one type or implementation of the meta-analyses implemented in these scripts and reported them as one, or perhaps they modified the analytic strategy in an undisclosed way.

In light of this, I therefore altered the implementations in multiple ways in order to attempt to reproduce Vahey et al.’s (2015) results. The code used to implement each verification attempt, notes on what was modified from the default original code, and the results of the meta-analyses are reported in Table 3. Copies of all original and modified scripts are available in the supplementary materials.

Definitions of different types of 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, $\hat{\sigma}^2$), whereas Credibility Intervals are a function of between-study variance ($\hat{\tau}^2$: 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, however, this is not because they are more 'conservative' than Confidence Intervals but because they quantify a different property under different assumptions. It is unclear whether this discrepancy in Vahey et al. (2015) was due to (a) a misinterpretation 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:

$$95\%\ \mathrm{CI} = \bar{r}\ \pm 1.96\sqrt{\hat{\sigma}^2} \tag{2}$$

$$95\%\ \mathrm{CR} = \bar{r}\ \pm 1.96\sqrt{\hat{\tau}^2} \tag{3}$$

$$95\%\ \mathrm{PI} = \bar{r}\ \pm 1.96\sqrt{\hat{\tau}^2 + \hat{\sigma}^2} \tag{4}$$

Where $\bar{r}$ is the weighted average effect size, $\hat{\sigma}^2$ is the estimated within-study variance (i.e., the square of the Standard Error of $\bar{r}$), and $\hat{\tau}^2$ is the estimated between-study variance (heterogeneity).

An important point to appreciate regarding Credibility Intervals is that when between-study heterogeneity is zero ($\hat{\tau}^2 = 0$), the CR interval width will also be zero, as is the case in the results of the verification attempts reported later.

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 s_r^2) 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 $\hat{\sigma}_p^2$ 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 $\hat{\sigma}_p^2$ 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., $95\%\ \mathrm{CR} = \bar{r} \pm 1.96 \times 0 = [\bar{r},\ \bar{r}]$). 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).

Verification attempt 1

The first verification attempt employed Field’s “h_s syntax.sps” SPSS script. The default 80% Credibility Interval widths were changed to 95% to match what was reported by Vahey et al. (2015).

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). 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. Partially missing values can be imputed via the mean, but at least some reliability values must be provided. However, Vahey et al. (2015) do not report any extracting or

Table 3. Verification attempts for the meta-analysis

				95% CI		95% CR		95% PI	
Source	Implementation	Modifications from the original code	$\bar{r}$	Lower	Upper	Lower	Upper	Lower	Upper
Vahey et al. (2015)	Vahey et al. (2015) state they followed Field & Gillett's (2010) description of a Hunter and Schmidt style meta-analysis	Unknown.	.45	.40	.54	.23	.67	-	-
Verification attempt 1	Hunter & Schmidt method using Field & Gillett's (2010) "h_s_syntax.sps"	All reliabilities were 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 so that CRs must be non-negative.	-	-	-	-	-	-	-
Verification attempt 3	Hunter & Schmidt method using a reimplementation of Field & Gillett's (2010) "Meta_Basic_r.sps" in R	Set variance in population correlations to zero if it is negative so that CRs must be non-negative.	.46	-	-	.46	.46	-	-
Verification attempt 4	Hunter & Schmidt method using a converstion of Field & Gillett's (2010) "Meta_Basic_r.sps" to R	Set variance in population correlations to zero if it is negative so that CRs must be non-negative. Removed erroneous Overton transformations.	.47	-	-	.47	.47	-	-
Verification attempt 5	Hunter & Schmidt method using Viechtbauer's (2022) implementation in R and metafor.	Credibility intervals were implemented using Field & Gillett's (2010) equations 2 to 5.	.47	.40	.54	.47	.47	.40	.54
Verification attempt 6	A mix of Hunter & Schmidt and Hedges methods using Viechtbauer's (2022) implementation in R and metafor.	Credibility intervals were 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
Verification attempt 7	Hunter & Schmidt method using Field & Gillett's (2010) "h_s_syntax.sps"	All reliabilities were set to 0. Data were the 56 individual weighted effect sizes rather than 15 weighted average effect sizes.**	.48	.20	.76	.39	.57	-	-

Notes: CI = Confidence Interval. CR = Credibility Interval. PI = Prediction Interval. Although PIs were not reported in Vahey et al. (2015), where possible they were calculated in the verification attempts to see if they corresponded 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 what can be accounted for by rounding or truncation (±.01).

* This SPSS script contains multiple issues that prevent it from running. See main text for discussion.

** This verification attempt was performed on the speculative basis that perhaps Vahey et al. (2015) ran their meta-analysis on the 56 individual weighted effect sizes rather than than the 15 weighted average effect sizes. There is some mild alignment between the results of this analysis, but only on the assumption that Vahey et al. (2015) also confused their CIs with their CRs. Alignment is closer but still imperfect, i.e., reported $\bar{r}$ = .45, reclaculated $\bar{r}$ = .48; reported CI = [.40, .54], reclaculated CR = [.39, .57]; reported CR = [.23, .67[,reclaculated CI = [.20, 74].

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 the reliability for all variables to 1.0 in order to allow the script to run.

This verification attempt did not reproduce the original results for the point estimate, Confidence Interval, or Credibility Interval (see

Verification attempt 3

I then reimplemented the math specified in the "Meta_Basic_r.sps" and "h_s syntax.sps" in R. I obtained identical results for the SPSS and R versions of the latter, providing some confidence that the reimplementation of the former was also accurate.

One necessary alteration was made to the code: if $\hat{\sigma}_p^2$ was negative it was set to zero to produce a Credibility Interval width of 0. This correction was specified in "h_s syntax.sps" but not "Meta_Basic_r.sps" – I merely applied it in both. Without this alternation, if $\hat{\sigma}_p^2$ was negative the script would fail to run.

This verification attempt of the R implementation of the Hunter and Schimdt style meta-analysis implemented in "Meta_Basic_r.sps" also did not reproduce the original results. The point estimate was off by only a small amount ($\bar{r}$ = 0.01), although this is more than can be accounted for by common methods of rounding, although it could be obtained via (erroneous) truncation. 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 $\hat{\sigma}_p^2$ was negative and interval width was therefore set to zero) and therefore greatly differed from the original results.

Verification attempt 4

A close reading of Field & Gillet (2010) and "Meta_Basic_r.sps" revealed an inconsistency between them: Field & Gillett state that Overton corrections should be applied to the individual correlations in the Hedges and colleagues approach but not the Hunter and Schmidt approach. However, the SPSS script applies Overton corrections in both. I therefore removed this correction from my R implementation for attempt 4. This changed the results very little from attempt 3, and did not reproduce Vahey et al.'s (2015) results.

Verification attempt 5

In order to try to obtain the original results, I then switched from using manual implementations of the equations reported in Field & Gillett (2010) (i.e., their SPSS code or my translations into R) to instead using an established R package for meta-analyses: Viechtbauer's (2022) implementation of a Hunter & Schmidt style meta-analysis written using the metafor package (Viechtbauer, 2010, 2024). This provided new avenues to attempt to reproduce the original results in a programming language and package I was more familiar with, allowing me to try a variety of variations on a given attempt more efficiently. Field & Gillett's (2010) equations 2-5 were used to implement Credibility Intervals. In this attempt, the Confidence Intervals reported by 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 4, as well as being very close to 3.

This verification attempt also attempted to reproduce the original forest plot (Vahey et al., 2015, Figure 1), which was more feasible in R and metafor. 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.

Verification attempt 6

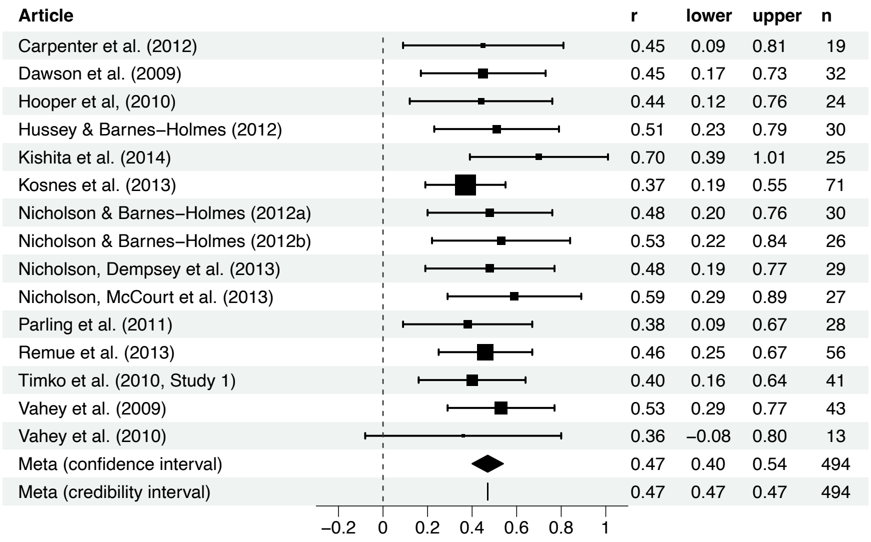
Next, I applied Fisher's *r*-to-*z* transformations to the individual effect sizes prior to meta-analysis and back transformations prior to reporting and plotting. The analysis was otherwise identical to the previous attempt. All estimated values were identical to attempt 5, therefore the original meta-analysis 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.

Verification attempt 7

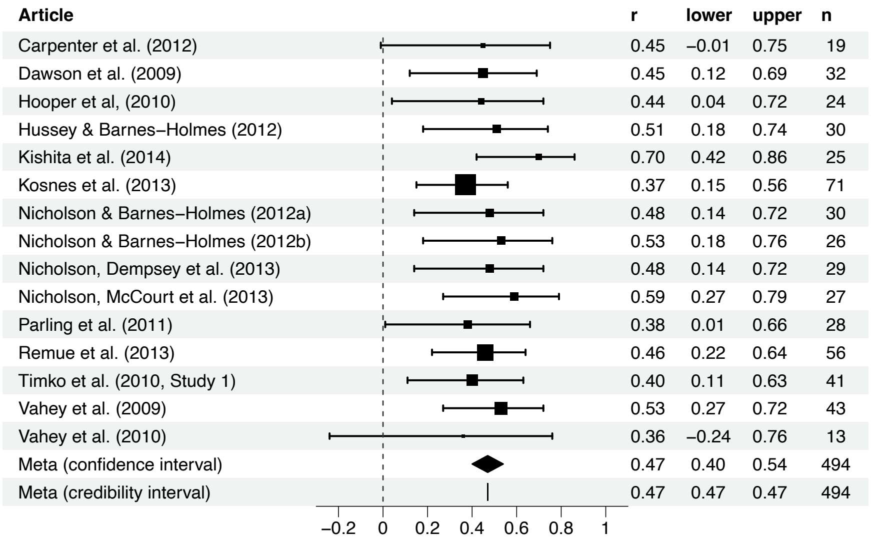
Lastly, I made an attempt that purposefully made statistical errors and went against Vahey et al.'s (2015) descriptions of their analytic strategy to see if it could allow me to reproduce the original results. Although Vahey et al. (2015) are explicit that they meta-analyzed the 15 weighted average effect sizes, in this attempt I instead used the 56 individual weighted effect sizes, weighted by the sample sizes reported in the original forest plot (their Figure 1).

Figure 3. Forest plot for meta-analysis verification attempt 5.



Article	r	lower	upper	n
Carpenter et al. (2012)	0.45	0.09	0.81	19
Dawson et al. (2009)	0.45	0.17	0.73	32
Hooper et al, (2010)	0.44	0.12	0.76	24
Hussey & Barnes–Holmes (2012)	0.51	0.23	0.79	30
Kishita et al. (2014)	0.70	0.39	1.01	25
Kosnes et al. (2013)	0.37	0.19	0.55	71
Nicholson & Barnes–Holmes (2012a)	0.48	0.20	0.76	30
Nicholson & Barnes–Holmes (2012b)	0.53	0.22	0.84	26
Nicholson, Dempsey et al. (2013)	0.48	0.19	0.77	29
Nicholson, McCourt et al. (2013)	0.59	0.29	0.89	27
Parling et al. (2011)	0.38	0.09	0.67	28
Remue et al. (2013)	0.46	0.25	0.67	56
Timko et al. (2010, Study 1)	0.40	0.16	0.64	41
Vahey et al. (2009)	0.53	0.29	0.77	43
Vahey et al. (2010)	0.36	–0.08	0.80	13
Meta (confidence interval)	0.47	0.40	0.54	494
Meta (credibility interval)	0.47	0.47	0.47	494

Figure 4. Forest plot for meta-analysis verification attempt 6.



Article	r	lower	upper	n
Carpenter et al. (2012)	0.45	–0.01	0.75	19
Dawson et al. (2009)	0.45	0.12	0.69	32
Hooper et al, (2010)	0.44	0.04	0.72	24
Hussey & Barnes–Holmes (2012)	0.51	0.18	0.74	30
Kishita et al. (2014)	0.70	0.42	0.86	25
Kosnes et al. (2013)	0.37	0.15	0.56	71
Nicholson & Barnes–Holmes (2012a)	0.48	0.14	0.72	30
Nicholson & Barnes–Holmes (2012b)	0.53	0.18	0.76	26
Nicholson, Dempsey et al. (2013)	0.48	0.14	0.72	29
Nicholson, McCourt et al. (2013)	0.59	0.27	0.79	27
Parling et al. (2011)	0.38	0.01	0.66	28
Remue et al. (2013)	0.46	0.22	0.64	56
Timko et al. (2010, Study 1)	0.40	0.11	0.63	41
Vahey et al. (2009)	0.53	0.27	0.72	43
Vahey et al. (2010)	0.36	–0.24	0.76	13
Meta (confidence interval)	0.47	0.40	0.54	494
Meta (credibility interval)	0.47	0.47	0.47	494

This analysis therefore purposefully ignored the dependencies among the effect sizes, which Vahey et al. (2015) acknowledged is problematic (see their Footnote 2). For this, I returned to using the original "h_s syntax.sps" SPSS script, with all reliabilities still set to 0.0.

The results of this verification attempt again did not reproduce the original results (nor did multiple small variations on it, e.g., weighting by degrees of freedom rather than *N*, or using no weights). However, one thing was perhaps notable: making one additional purposeful mistake provides the closest reproduction of the original results that I managed to obtain. Specifically, if one mislabels the Credibility Intervals as Confidence Intervals and vice versa, results show some alignment with the originally reported results: reported $\bar{r}$ = .45, recalculated $\bar{r}$ = .48; reported CI = [.40, .54], recalculated CR = [.39, .57]; reported CR = [.23, .67[, recalculated CI = [.20, 74].

Summary of attempts

A larger number of small variations on the attempts that are reported here were also tried. For example, alternative values for reliability estimates, and not back-transforming the *z* values back to *r* values. I also tried several other 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 6) and not when they weren't (verification attempts 1-3).

Meta-analysis Confidence Intervals were only reproduced when putting Field's SPSS scripts aside and reconstructing the analyses in R using the metafor package. This is difficult to account for.

Credibility Intervals could not be reproduced in any attempt. Indeed, all verification attempts in both SPSS and R, whether using Field's mathematical solutions or metafor's, returned CRs with widths of 0. The only exceptions to this were situations where I made purposeful errors. It remains totally unclear how Vahey et al. (2015) produced their reported values. The closest I came to reproducing them was attempt 7, which had to make two serious mistakes on purpose: using the 56 individual effect sizes rather than the 15 weighted averages, and mislabelling Confidence Intervals as Credibility Intervals and vice versa.

Lastly, with regard to the point estimate of the meta-analytic effect size, 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. Interestingly, 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 halfway point between the intervals). A point estimate of .47 combined with Confidence Intervals of [.40, .54] was reproduced in verification attempts 5 and 6 using metafor. 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 from 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. Therefore, it remains unclear how Vahey et al. (2015) analyzed their data or obtained all their results, or even which mistakes if any during the meta-analysis may have given rise to their reported results.

Weighted 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 Vahey et al's (2015) 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 et al.'s forest plot. On the one hand, the magnitudes of the differences were small ($\Delta\bar{r}$ = -.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 whose first authors were 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 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 et al.'s (2015) 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 et al.'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 incorrect 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 of 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 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 Pearson's r. However, inspection of those explanations revealed at least one error: 2 of the effect sizes were η_p^2 effect sizes taken from ANOVAs, which Vahey et al. (2015) stated that they "equated the relevant statistic [η_p^2] with r^2 therefore obtaining r using the square root function". However, this conflates η_p^2 with η^2: as a partialized effect size, η_p^2 cannot be converted to r, and therefore these conversions are erroneous.

A comprehensive assessment of the reproducibility of the conversions of the individual effect 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).

Issues in the publication bias analyses

Vahey et al. (2015) reported employing tests of funnel plot asymmetry, a sensitivity analysis based on selection models (Vevea & Woods, 2005), and Kendall's τ Rank Correlation Test (e.g., Egger et al., 1997). While I did not attempt to systematically verify the results of all of these, two points are worth highlighting here.

First, Vahey et al. (2015) conclude that the results of these tests suggest "that the current meta-analysis was not subject to publication bias" (p. 62). However, this falls into a statistical fallacy that is common in original research: non-significant p values should not be interpreted as evidence for the null hypothesis, only failure to reject the alternative hypothesis (Aczel et al., 2018; Greenland et al., 2016). Put differently, the absence of evidence is not the same as evidence of absence. This is especially important in the context of meta-analysis bias tests which frequently have very low power (Rücker et al., 2011; Sterne et al., 2000), as is the case here. The correct interpretation of such non-significant results is no evidence of bias was obtained rather than evidence of no bias. This difference in wording may seem subtle at first, but represents a fundamentally different and stronger claim. There are few areas of research where publication bias and p-hacking could reasonably be assumed to be completely absent. As such, direct evidence for this null effect would need to be strong to dismiss the presence of bias as a plausible default assumption.

Second, bias tests can allude to rigor or objectivity that might obscure other sources of information about whether bias is truly present. It is worth noting that 8 of the 11 (73%) articles used in the meta-analysis were co-authored by at least one author of Vahey et al. (2015). The authors of Vahey et al. (2015) therefore had direct knowledge of whether there was a file drawer of unpublished studies (or indeed any other source of bias), but they do not consider this in their estimation of bias. My own compilation of unpublished IRAP studies suggests that there are at least 6 unpublished PhD theses with clinically relevant IRAP studies, most of which came from Barnes-Holmes's research group (Hussey & Drake, 2022). Reporting quantitative tests of publication bias without also reporting *prima facie* evidence of publication bias from one's own research group ignores important evidence, and does so in a way that is biased towards enhancing the apparent criterion validity of the measure – a measure which was also created by the last author of Vahey et al. (2015).

Corrected meta-analysis and power analyses

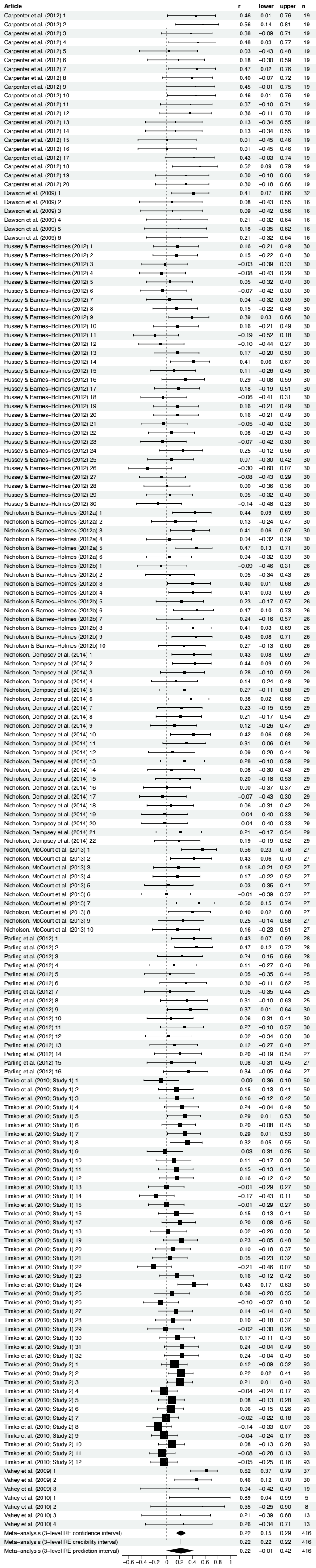
In order to understand the compound impact of the various errors on the conclusions of the meta-analysis, I fitted a new meta-analysis to the 156 effect sizes re-extracted from the original articles. I then used the meta-analysis effect results to calculate new power analyses. Whereas Vahey et al.'s (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 $\bar{r}$ = .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 by Vahey et al. (2015), i.e., $\bar{r}$ = .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 several different stages of the data extraction and analysis, and multiple errors and internal 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 were biased towards including larger effect sizes and omitting smaller ones. The averaging of these effect sizes for each article was not computationally reproducible in 13% of cases. The results of the meta-analysis could not be reproduced despite numerous different attempts and approaches, with the closest reproducing requiring two serious errors to be made on purpose (using the wrong dataset and mislabelling Confidence Intervals as Credibility Intervals and vice versa). The original power analyses were mostly reproducible, however, given the lack of reproducibility of the meta-analysis itself, the validity of those power analyses' results based on that meta-analysis estimate was fundamentally undermined.

This lack of reproducibility is 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 of $\bar{r}$ = .22, less than half that reported in the original article ($\bar{r}$ = .45). Vahey et al. (2015) stated that, according to their results, the IRAP's criterion validity compares "favorably" to the other popular implicit measures such as the Implicit Association Test ($\bar{r}$ = .22 for addiction and $\bar{r}$ = .30 for non-addiction psychopathologies: Greenwald et al., 2009) and evaluative priming methods ($\bar{r}$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 is required in future IRAP studies than recommended by Vahey et al. (2015). For example, although Vahey et al. (2015) make 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-analyzed 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-analyzed 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 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.

Future research on error detection in meta-analyses

The recommendations of much of the previous meta-science research on errors in meta-analyses have been recommendations to the authors of meta-analyses themselves on how to prevent errors (e.g., Lakens et al., 2016; López-Nicolás et al., 2022). Relatively fewer recommendations, or indeed general strategies, have been made for researchers engaged in error detection. Kadlec et al. (2023) provide an excellent example of this, with both descriptions of their general error detection strategies and concrete recommendations such as regarding Standardized Mean Difference effect sizes (e.g., Cohen's *d*, Hedges' *g*) that are larger than 3 with great suspicion. The current research offers some additional suggestions and guiding principles for error detection in meta-analyses:

1. Check whether reported intervals are symmetrical around the point estimate, including both intervals around estimates from original studies and meta-analysis results. The asymmetry of intervals can provide a clue that something may be amiss, depending on their compatibility with the reported model and transformations.
2. Plots that appear to have been created using software other than commonly employed meta-analysis software (e.g., common R packages for this, Cochrane's RevMan, Comprehensive Meta-Analysis, etc) may be more likely to contain or expose errors.
3. Information is sometimes repeated between plots and tables, e.g., between forest and funnel plots. This provides a vector for error-checking that data points which should be identical are indeed so. For example, forest plots often present effect sizes both graphically and numerically, and forest plots and funnel plots present both effect sizes and (repressions of) their precision (i.e., Confidence Intervals vs. Standard Errors, which can be calculated from one another: SE = [95% CI upper – 95% CI lower] / [1.96 × 2]).
4. Data can be extracted from plots for error checking using free and Open Source tools such as WebPlotDigitizer (Marin et al., 2017).
5. Systemic checking of effect size extractions and conversion can be time-consuming. However, initial spot checks can easily be performed on the most extreme effect sizes, which are most likely to have involved an extraction or conversion error (e.g., confusing SE with SD: see Kadlec et al., 2023).
6. The accurate application of the inclusion criteria can also be checked, whether systematically or using spot checks. This can include checks for both incorrect inclusions and incorrect omissions.
7. The normative plausibility and mathematical possibility of correlations can be assessed by deattenuating them for the reliability of the measures that produced them. This can be done using empirical estimates for those measures or plausible values (which can themselves be informed by data from the literature: Hussey et al., 2023). Correlations larger than the reliability of the component measures are implausible.
8. The normative plausibility of effect sizes, both original and meta-analytic, can also be compared to large-scale analyses of this in the literature (e.g., Hemphill, 2003; Plessen et al., 2023; Richard et al., 2003).
9. Bias assessments can also be scrutinized. This can include an assessment of over-claiming via incorrect interpretations of non-significant tests for bias.
10. Additionally, it can be useful to assess the overlap in authorship between the meta-analysis and the original studies in order to understand potential sources of bias, including but not limited to contextualizing the results of any quantitative tests of publication bias or *p*-hacking.

Conclusions

The results of Vahey et al. (2015) were found to have poor reproducibility at almost every 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 ($\bar{r}$ = .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.

Conflict of Interest

Prof Dermot Barnes-Holmes, one of the authors of Vahey et al. (2015), was my PhD supervisor between 2010 and 2015. I have not actively collaborated with Prof Barnes-Holmes since 2015. Articles led by third parties of which we were both co-authors were published up to 2018. I declare that I have no other conflicts of interest associated with the publication of this manuscript.

Acknowledgments

I first presented elements of these analyses at the Association for Contextual Behavioral Science meeting in Dublin in 2019. That symposium was also the first time I met my now-wife, Sabrina. I would therefore like to acknowledge that although the IRAP literature has brought me much frustration over the years, it also brought me my greatest joy.

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances* *in Methods and Practices in Psychological Science*, *1*(3), 357–366. https://doi.org/10.1177/2515245918773742

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., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*(7), 169–177.

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

Bast, D. F., & Barnes-Holmes, D. (2015). Developing the Implicit Relational Assessment Procedure (IRAP) as a Measure of Self-Forgiveness Related to Failing and Succeeding Behaviors. *The Psychological Record*, *65*(1), 189–201. https://doi.org/10.1007/s40732-014-0100-5

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

De Schryver, M., Hussey, I., De Neve, J., Cartwright, A., & Barnes-Holmes, D. (2018). The PIIRAP: An alternative scoring algorithm for the IRAP using a probabilistic semiparametric effect size measure. *Journal of Contextual Behavioral Science*, *7*, 97–103. https://doi.org/10.1016/j.jcbs.2018.01.001

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). *Bias in meta-analysis detected by a simple, graphical test*. https://doi.org/10.1136/bmj.315.7109.629

Farrell, L., & McHugh, L. (2017). Examining gender-STEM bias among STEM and non-STEM students using the Implicit Relational Assessment Procedure (IRAP). *Journal of Contextual Behavioral Science*, *6*(1), 80–90. https://doi.org/10.1016/j.jcbs.2017.02.001

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

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

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

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Heathers, J. A., Anaya, J., Zee, T. van der, & Brown, N. J. (2018). *Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)* (e26968v1). PeerJ Inc. https://doi.org/10.7287/peerj.preprints.26968v1

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., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2023). *An aberrant abundance of Cronbach's alpha values at .70*. PsyArXiv. https://doi.org/10.31234/osf.io/dm8xn

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

Hussey, I., & Drake, C. E. (2022). *The IRAP File-Drawer: A repository of unpublished studies using the Implicit Relational Assessment Procedure*. https://osf.io/g4qsu/

Hussey, I., Thompson, M., McEnteggart, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). Interpreting and inverting with less cursing: A guide to interpreting IRAP data. *Journal of Contextual Behavioral Science*, *4*(3), 157–162. https://doi.org/10.1016/j.jcbs.2015.05.001

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

Leech, A., Bouyrden, J., Bruijsten, N., Barnes-Holmes, D., & McEnteggart, C. (2018). Training and testing for a transformation of fear and avoidance functions using the Implicit Relational Assessment Procedure: The first study. *Behavioural Processes*, *157*, 24–35. https://doi.org/10.1016/j.beproc.2018.08.012

López-Nicolás, R., López-López, J. A., Rubio-Aparicio, M., & Sánchez-Meca, J. (2022). A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020). *Behavior Research Methods*, *54*(1), 334–349. https://doi.org/10.3758/s13428-021-01644-z

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

Maloney, E., & Barnes-Holmes, D. (2016). Exploring the Behavioral Dynamics of the Implicit Relational Assessment Procedure: The Role of Relational Contextual Cues Versus Relational Coherence Indicators as Response Options. *The Psychological Record*, *66*(3), 395–403. https://doi.org/10.1007/s40732-016-0180-5

Marin, F., Rohatgi, A., & Charlot, S. (2017). *WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: Application to ultraviolet spectropolarimetry* (arXiv:1708.02025). arXiv. http://arxiv.org/abs/1708.02025

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

Plessen, C. Y., Karyotaki, E., Miguel, C., Ciharova, M., & Cuijpers, P. (2023). Exploring the efficacy of psychotherapies for depression: A multiverse meta-analysis. *BMJ Ment Health*, *26*(1). https://doi.org/10.1136/bmjment-2022-300626

Power, P. M., Harte, C., Barnes-Holmes, D., & Barnes-Holmes, Y. (2017). Exploring Racial Bias in a European Country with a Recent History of Immigration of Black Africans. *The Psychological Record*, *67*(3), 365–375. https://doi.org/10.1007/s40732-017-0223-6

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

Rücker, G., Carpenter, J. R., & Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*, *53*(2), 351–368. https://doi.org/10.1002/bimj.201000151

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. https://doi.org/10.1016/S0895-4356(00)00242-0

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

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

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

Wilkinson, J., Heal, C., Antoniou, G. A., Alfirevic, Z., Avenell, A., Barbour, V., Brown, N. J. L., Carlisle, J., Dicker, P., Dumville, J., Grey, A., Gurrin, L. C., Hayden, J. A., Heathers, J., Hunter, K. E., Lasserson, T., Lam, E., Lensen, S., Li, T., ... Kirkham, J. J. (2023). *Protocol for the development of a tool (INSPECT-SR) to identify problematic randomised controlled trials in systematic reviews of health interventions* (p. 2023.09.21.23295626). medRxiv. https://doi.org/10.1101/2023.09.21.23295626