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

Ian Hussey

Ghent University

Author Note

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

Abstract

Vahey et al.,

Keywords: [Click here to add keywords.]

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

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

## The IRAP

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

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

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

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

## Replicability and reproducibility

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

Concerns about reproducibility have been raised about not only original research articles but also the results of meta-analyses. Lakens et al. (2017) recently demonstrated that the results of the majority of a random sample of meta-analyses published in psychology cannot be reproduced [in what way?]. Maassen et al. (2020) found that almost half of effect-sizes reported in meta-analyses of psychology research could not be reproduced. This was attributed to due to a variety of issues such as errors in the extraction of effect sizes from original studies, insufficient details regarding data processing and transformation of effect sizes, insufficient details of the specific meta-analytic approach employed, or failures to adhere to meta-analysis reporting guidelines. In this article, I therefore sought to assess the reproducibility of Vahey et al.’s (2015) data and results.

# Methods and results

In the following sections, I work backwards through Vahey et al.’s (2015) results and data and attempt to reproduce each step. Then, in order to assess the compound impact of the reproducibility issues I found in each step, I conduct a new meta-analysis and power analysis.

All data was taken from Vahey et al.’s (2015) article and supplementary materials, from the original articles that they extracted their effect sizes from, or from data sent to me upon request by the authors of the original studies. All analyses were written in R (REF) using the packages pwr (REF) and metafor (REF). All data and code to reproduce my analyses can be found on the Open Science Framework (osf.io/XXXX).

Prior to attempting to reproduce Vahey et al.’s (2015) data and results or implementing the data processing and analyses in R myself, I contracted the first author of Vahey et al. (2015) and requested copies of their data, code and/or analysis scripts. However, he declined to share these materials. Preliminary findings of this paper were presented at a conference in 2019. All my current data and code was made publicly available for comment and also sent to the first author of Vahey et al. (2015). No feedback was received in the 12 months between sharing my materials and writing this manuscript. Finally, I contacted all three authors of Vahey et al. (2015) before submitting this article for publication or to a preprint server. The draft manuscript, along with all data and code was supplied to them, and they were encouraged to find any possible mistakes in my reanalysis and provide comments. [comment on whether Vahey et al. (2015) responded or found any issues]

## Power analyses

Vahey et al.’s (2015) reported meta-analytic effect size estimate for the association between the IRAP and clinically relevant criterion variables was *r* = .45, 95% CI [.40, .54], 95% CR [.23, .67]. Using this effect size, they conducted power analyses for sample size planning. They reported that, to detect a zero order correlation with 80% power, 29 participants would be required when using the meta-analytic effect size, or 37 if using the lower bound of the CI (alpha = .05, one-tailed; following recommendations by Perugini, Gallucci, & Costantini, 2014). These sample size recommendations were computationally reproducible.

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

## Meta-analytic effect size

Vahey et al.’s (2015) aforementioned power analyses relied on the accuracy of the meta-analytic effect size. I attempted to computationally reproduce the meta-analytic effect size from the weighted-mean effect sizes, 95% Confidence Intervals and sample sizes reported in Vahey et al.’s (2015) forest plot (p.XX). I noted that some of the confidence intervals in Vahey et al.’s (2015) forest plot were asymmetrical around the point estimate. This is uncommon for Pearson’s *r* effect sizes, and was not accounted for by Vahey et al. detailing of how they calculated the effect sizes and their confidence intervals. However, I took them at face value as they are the most detailed data available to work from. 95% Confidence Intervals around the effect sizes were extracted from the forest plot and converted to variances using the following formula:

Vahey et al. (2015) reported employing a Hunter and Schmidt style meta-analysis: the Hunter & Schmidt estimator was used, and the effect sizes and their variances were weighted by sample size. Results demonstrated a meta-analytic effect size of *r* = .47, 95% CI [.40, .54], 95% CR [.40, .54], *p* < .0001. No heterogeneity was observed, *Q*(*df* = 14) = 7.05, *p* = .933, 𝜏2 = 0.0, *I*2 = 0.0, *H*2 = 1.0. Vahey et al.’s (2015) meta-analysis results could therefore not be precisely computationally reproduced using the data they reported in their forest plot and their descriptions of their analytic approach. Estimate of the meta-analytic effect size differed by only a small amount (Δ*r* = .02), confidence intervals were identical, and credibility intervals differed substantially.

However, credibility intervals and therefore estimated heterogeneity differed by a large amount (large heterogeneity in Vahey et al. 2015, no heterogeneity in my reanalysis).

* Definition of credibility interval, interpretation of heterogeneity.

## Weighted-average effect sizes

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

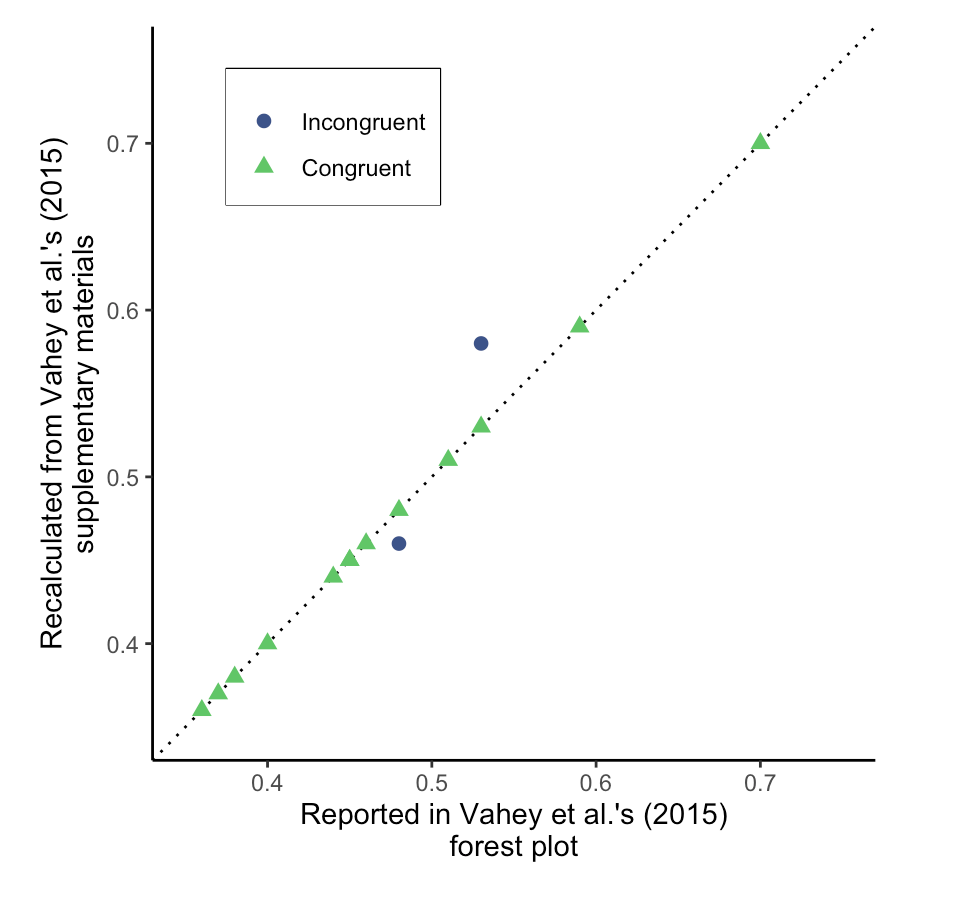


Figure XX. XXXX

## Extraction and conversion of effect sizes

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

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

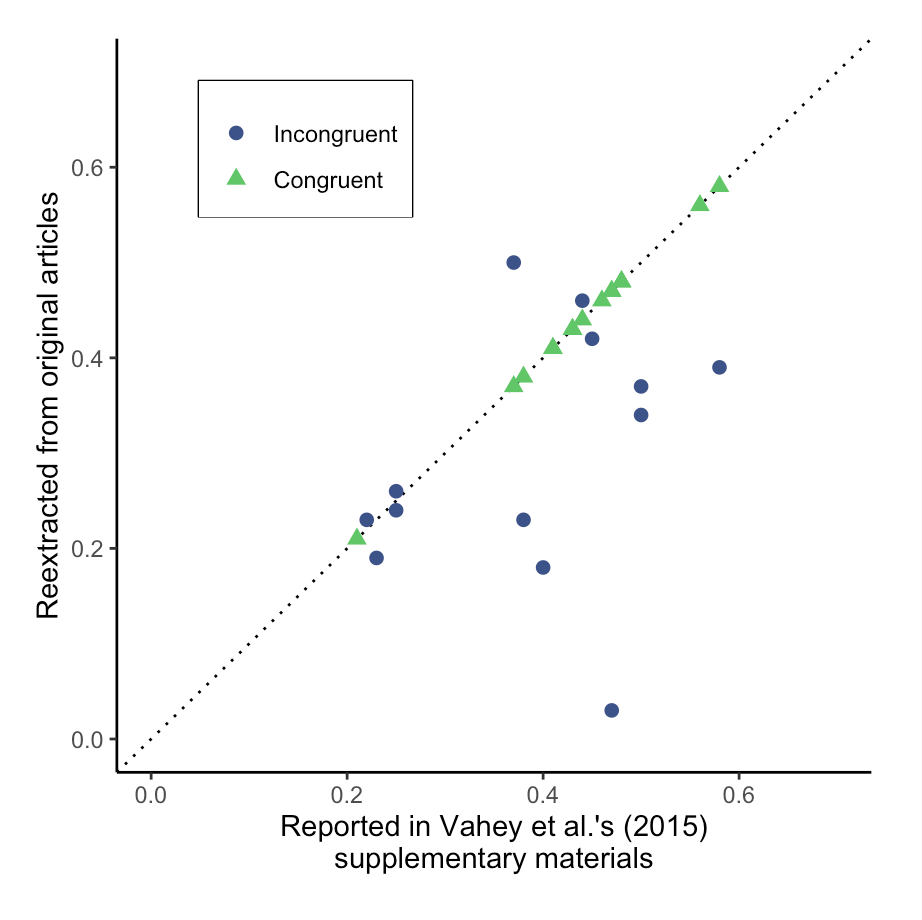


Figure XX. XX

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

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

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

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

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

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

## Selection of effect sizes

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

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

Vahey et al. extracted 56 effect sizes from 15 articles, but provided no information about the number of effects that were not included or details of these excluded effects. I re-extracted all effect sizes reported in these 15 articles, resulting in 334 effect sizes. Some additional effect sizes were found that were non-independent with the extracted ones (e.g., follow-up *t* tests after ANOVA, correlations with the overall IRAP score when its component trial types were also correlated, or correlations with a scale’s sum score when correlations with its subscale sum scores were also available).

Two independent raters then rated each effect (both the IRAP domain and the criterion) for clinical relevance using Vahey et al.’s (2015) definition. If both raters scored the effect as clinically relevant it was included in the meta-analysis. Agreement was found in 90% of cases (Cohen’s = 0.88, *p* < .0001).

Effect sizes were excluded if they were not rated as clinically relevant. No exclusions were made on the basis of ‘retrospective a priori predictions’ for the reasons discussed above. 144 effect sizes remained were selected for inclusion in the meta-analysis.

## Revised meta-analysis

One could argue that while some of the above steps were not found to be computationally reproducible, the differences between the results reported by Vahey et al. (2015) and the reproduced results here are small in many cases (e.g., the meta-analytic effect size estimate). However, it is important to appreciate that the individual steps cannot be viewed in isolation. For example, the impact of issues early on the analytic process (e.g., the extraction and conversion of effect sizes) may have implications for later steps (e.g., the meta-analytic effect size estimate) that are not visible when attempting to reproduce each step separately. In order to assess the compound impact of these reproducibility issues on Vahey et al.’s (2015) conclusions, a revised meta-analysis was conducted using the reextracted effect sizes.

The meta-analytic strategy was updated in the following ways relative to Vahey et al. (2015) in order to employ contemporary standards. Effect sizes were weighting by inverse variances rather than sample size, on the basis that sample size provides a poorer proxy of measurement error. Effect sizes were transformed using Fishers *r*-to-*z* transformations prior to meta-analysis, in order to deal with ceiling effects in Confidence Intervals, and results were back transformed for reporting. A Restricted Maximum Likelihood estimator function was used. Recent results from simulation studies suggests that the method employed by Vahey et al. (2015) to deal with the non-independence of multiple effect sizes estimates from a given study (i.e., weighted-mean effect sizes) provides poor statistical power. The more powerful and recommended approach of employ a multi-level meta-analysis model was employed instead (REF). The individual effect sizes were entered in the model without averaging, and their source article was entered as a random effect (i.e., random intercept).

Results demonstrated a meta effect size *r* = .18, 95% CI [.10, .26], 95% CR [-.04, .39], *p* = .00001. No evidence of heterogeneity was found, *Q*(df = 141) = 159.03, *p* = .142, 𝜏2 = 0.00. Based on the non-overlap of their confidence intervals, this estimate was significantly smaller than the effect size reported in Vahey et al. (2015: *r* = .45, 95% CI [.40, .54]). Given the large number of effect sizes involves, these results are illustrated in a caterpillar plot rather than a forest plot (see Figure XX; similar to a forest plot but does not include article labels, and sorts effect sizes by magnitude).



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

## Revised power analyses

As in Vahey et al. (2015), this estimate of meta-analytic effect size was in a power analysis for future sample size planning. To detect a zero order correlation with 80% power (alpha = .05, two-sided), the minimum sample size was 240 participants when using the revised meta-analytic estimate, or 782 when using the more conservative lower bound of the estimate’s 95% Confidence Interval. These revised recommendations are more than 8 times the sample sizes recommended by Vahey et al. (2015).

## [old points]

The numeric results reported in the forest plot were also compared against estimations of the values displayed in the plot. No discrepancies were found in either the estimates or the confidence intervals.

While the degrees of freedom used were reported in supplementary materials, it was less clear how the samples sizes used for weightings and reported in the forest plot were obtained, given that the individual effect sizes that were converted to mean effect sizes were in many cases calculated from different sample sizes, yet the reported sample sizes were even numbers.

Assessment of bias

* One or more authors of Vahey et al. (2015) was also an author of 12 of the 15 articles (80.0%) from which effect sizes were extracted, indicating that the authors of the original meta-analysis were familiar with the research they were meta-analysing.
* “our meta-analysis did not appear to prioritize consideration of statistically significant effects over non-significant effects”

According to the systematic review (see Supplementary Materials), both of these estimates are more than ten times larger than the mean sample sizes employed in IRAP research to date.

# Discussion

The majority of the steps in Vahey et al.’s (2015) meta-analysis were not found to be computationally reproducible (i.e., meta-analysis results, calculation of weighted-mean effect sizes, extraction and conversion of individual effect sizes, and selection of effect sizes). Where one step were found to be computationally reproducible (power analyses), it was found to be poorly justified.

Nearly half of the effect sizes included in the original meta-analysis did not match those reextracted from the original articles. In one third of cases, the effect sizes used in the original meta-analysis were biased upwards relative to the re-extractions done here.

Data processing was found to not be reproducible, with 13% of cases demonstrating disagreement between the weighted average effect sizes reported in the forest plot and those recalculated from the effect sizes reported in the supplementary materials.

More worryingly, when all effect sizes were reextracted from the original articles a large number of questionable inclusions and inclusions were highlighted. When all effect sizes were included that a) met Vahey et al.’s inclusion criterion of being clinically relevant and b) were not derived from types of analyses that were defined a priori as producing invalid or misleading results or conclusions, the meta effect size estimate reduced greatly (original *r* = .45, 95% CI [.40, .54], new: *r* = 0.13, 95% CI [0.03, 0.23]).

Power analyses calculations for future research using this updated effect size estimate suggest minimum sample sizes of more than 460 participants; an estimate that is 16 times larger than recommended by Vahey et al. and 10 times larger than the mean sample sizes employed in IRAP research to date.

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

## Improving the reproducibility of future meta-analyses

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

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

## Conclusion

XXX

References

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

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

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

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

Gemar, M. C., Segal, Z. V., Sagrati, S., & Kennedy, S. J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. *Journal of Abnormal Psychology*, *110*(2), 282–289. https://doi.org/10.1037//0021-843X.110.2.282

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

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

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

Hussey, I., & Barnes-Holmes, D. (2012). The implicit relational assessment procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice*, *19*(4), 573–582. https://doi.org/10.1016/j.cbpra.2012.03.002

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

Lakens, Daniël, 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

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science*, *21*(4), 511–517. https://doi.org/10.1177/0956797610364762

Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M. A., & De Raedt, R. (2013). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self-versus ideal self-related cognitions in dysphoria. *Cognition & Emotion*, *27*(8), 1441–1449. https://doi.org/10.1080/02699931.2013.786681

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

Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a Fatal Decision: Direct Replication of the Predictive Validity of the Suicide–Implicit Association Test. *Psychological Science*, *31*(1), 65–74. https://doi.org/10.1177/0956797619893062

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