# The IRAP demonstrates very poor internal consistency and test-retest reliability:

# A file-drawer meta-analysis

# Ian Hussey

# Abstract

Vahey et al.’s (2015) meta-analysis argued that the Implicit Relational Assessment Procedure (IRAP) has potential for clinical assessment. However, Here I present evidence to the contrary through meta-analyses of file drawer data. Individual participant data from all published and unpublished studies I have been involved with was used to estimate the IRAP’s internal consistency and test-retest reliability across a large number of domains (*k* = XX) and participants (*N* = XX). Results suggest that both internal consistency (alpha = XX) and test-retest reliability (ICC = XX) are poor. This severely limits the IRAP’s ability to function as a useful measure, for clinical use or otherwise, especially at the individual level. I conclude that the IRAP is not currently suitable for clinical use, and that we should be cautious about choosing to employ it or when interpreting its results.

# The IRAP demonstrates very poor internal consistency and test-retest reliability:

# A file-drawer meta-analysis

The Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2010) is a computer-based reaction time task designed to capture automatic relational responding. It has been used extensively as a measure of implicit attitudes (Gawronski & De Houwer, 2011) and within Contextual Behavioral Science due to its historic ties to Relational Frame Theory (Hussey et al., 2015). In their meta-analysis, Vahey et al. (2015) argued that the IRAP has potential as a tool for clinical assessment. This aspiration that the IRAP might be used in applied contexts is long-standing. However, concerns have been expressed about the IRAP’s low reliability (Golijani-Moghaddam et al., 2013; Greenwald & Lai, 2020) and poor measurement properties (Hussey, 2020). Together, these suggest that the IRAP’s utility in both research and applied settings is likely to be lower than has sometimes been claimed.

The importance of precise measurement has received renewed attention within psychology in recent years. For example, multiple authors have recently noted that poor reliability can result in experimental effects that are highly replicable that nonetheless lead to false or invalid conclusions (Devezer et al., 2020; Hussey & Hughes, 2020). There is mutual consensus, even field of research that have at times been skeptical of the utility of psychometric methods (e.g., behaviorism), that measurement is a cornerstone of the scientific method. For example, even an animal-behaviorist working with rats in Skinner boxes must be concerned with whether the levers function well as measures of the animal’s emitted behaviour: if the lever is too heavy or too stiff, a recorded acquisition curve will not accurately reflect the animal’s lever-pressing behaviour, and will hinder the researcher’s analysis. As such, although the IRAP has often been employed by behaviorally orientated researchers interested who identify as engaging in function analytic-abstractive research (see Barnes-Holmes & Hussey, 2016; Hayes & Brownstein, 1986), issues of psychometric reliability cannot be ignored.

## Previous meta-analyses of the IRAP’s reliability

The IRAP’s reliability has been examined in two previous meta-analyses of published articles. Golijani-Moghaddam et al. (2013) extracted data from 31 published studies (systematic search?), including 1301 participants for the meta-analysis of internal consistency and one study of 23 participants assessing test-retest reliability. Meta estimates of internal consistency were Pearson/Spearman-Brown *r* = .65, 95% CI [.54, .74] (is this SB corrected or not?). Just one study was found that reported test-retest reliability: *r* = .49, 95% CI [.10, .75].

More recently, Greenwald & Lai (2020) conducted a large scale review and meta-analyses of multiple implicit measures including the IRAP. Thanks to making their data and code openly available, it was possible to computationally reproduce their meta-analyses of IRAP data (see supplementary materials for data and code: [osf.io/v3twe](https://osf.io/v3twe/)). They note in their data that many estimates were sourced from other meta-analyses – presumably Golijani-Moghaddam et al.’s (2013). Data was included from 13 published studies (systematic search?), including 1207 participants for the meta-analysis of internal consistency and two studies including 124 participants assessing test-retest reliability. Given that Greenwald and Lai’s (2020) meta-analysis was conducted 7 years after Golijani-Moghaddam et al.’s (2013) but includes significantly fewer articles, this suggests that Greenwald and Lai (2020) employed a less systematic search strategy. A recent systematic review of all published IRAP research suggested that XXX articles were published before the end of 2018 (REF). This suggests that Greenwald and Lai’s (2020) analysis had relatively poor coverage of the published literature. Meta-estimates of internal consistency using Greenwald & Lai’s (2020) data were Cronbach’s α = .56, 95% CI [.46, .65], 95% CR [.03, .85]. Meta-estimates of test-retest reliability were Pearson’s *r* = .45, 95% CI [.33, .55].

Both meta analyses half strengths and weaknesses: Golijani-Moghaddam et al. (2013) benefited from a more comprehensive search strategy, whereas Greenwald & Lai (2020) benefited from greater computational reproducibility. In one sense, the results of the two meta-analyses show a significant degree of variation, with Greenwald & Lai (2020) reporting a substantively lower estimate of internal consistency than Golijani-Moghaddam et al. (2013). However, both meta-analyses support the idea that the IRAP’s reliability is problematically low, both below typically accepted cut-offs for assessment measures in psychology (REF), and also lower than other implicit measures such as the Implicit Association Test (IAT: Greenwald et al., 1998). This poses a significant threat to the task’s basic and applied utility, both in relation to other assessment methods more generally but also compared to alternative implicit measures more specifically. Contrary to the conclusions of Vahey et al.’s (2015) meta-analysis, the IRAP’s poor measurement properties suggest that it does not currently demonstrate potential for clinical assessment.

## The current research

Nonetheless, two factors suggest that there is need for additional assessments of the IRAP’s reliability. First, meta-analyses of published literature are susceptible to publication bias. Given the relationship between internal consistency and statistical power (Parsons, 2018), it is quite possible that IRAP studies that demonstrated poor measurement properties were less likely to obtain significant results, and therefore were unfortunately less likely to be published.

Second, published articles have used a range of different metrics when reporting reliability, and have frequently not reported gold-standard metrics. For example, published studies on test-retest reliability have reported Pearson’s *r* correlations. However, Parsons et al. (2019) recently highlighted that Pearson’s *r* captures one specific aspect of stability (i.e., the preservation of rank among participants between time-points) but neglects others (e.g., the absolute change in scores between timepoints). This can be illustrated using a simple example: imagine if at time-point 2 all participants scored exactly 10 points higher on an IQ scale than they did at time-point 1. A Pearson’s *r* correlation would suggest that test-retest reliability was perfect (*r* = 1.0) because rank among participants was preserved, despite there being clear and large changes in responses between the timepoints. In order to capture both aspects (preservation of rank and lack of absolute change), a measure of ‘Absolute Agreement’ such as Intraclass Correlation Coefficients (ICC) should instead be reported (specifically ICC[2,1]: see Parsons et al., 2019).

To take another example, the calculation of internal consistency via split-half reliability involves a somewhat arbitrary decision regarding how the data is split. While most IRAP studies have split by odd versus even trials by order of presentation, other common implicit measures such as the IAT instead split by first versus second half of the task by order of presentation. Parsons et al. (2019) note that both choices are arbitrary, and that internal consistency should instead be estimated by a permutation resampling approach. This involves creating a large number of random splits of the data and calculating reliability for each, then taking this mean of this distribution of reliabilities. Importantly, this method approximates Cronbach’s alpha where others frequently do not. However, in order to calculate both ICCs and permutation based estimates of internal consistency, access to trial-level data is needed.

Both of the above factors were addressed by conducting a file drawer meta-analysis. That is, where all studies both published or unpublished originating from an individual or group are used. Data was provided by two active IRAP researchers.

# Method

## Data

All code for data processing and analyses and all processed data is available on the Open Science Framework ([osf.io/v3twe](https://osf.io/v3twe/)).

Internal consistency data came from 25 different IRAPs in 13 different domains. In total, this meta-analysis included 886 participants (see Figure 1 for domains and sample sizes). Data from several of these IRAPs has been published (i.e., XX% of participants; gender stereotypes: REF; friend-enemy: REF; one of the life-death IRAPs: REF; Lincoln-Hitler: REF; race: REF; and shapes and colors IRAPs: REF). However, none of these studies were included in either previous meta-analyses of the IRAP’s reliability.

Test-retest data came from two different IRAP studies employing two different follow-up periods (immediate and 1 week). In total, the test-retest meta-analysis included 67 participants (see Figure 1 for domains and sample sizes). Both studies were unpublished, and neither were included in either previous meta-analyses of the IRAP’s reliability.

## Participants

All participants provided informed consent prior to participation, and studies were approved by the local institutional review boards. Where demographics data was available, a majority of participants were women (63% female, 37% male, 0.2% non-binary; *M*age = 21.0, *SD* = 5.7).

**Measures**

The IRAP is a computer-based reaction time task. Its procedural parameters have been discussed in great detail in many other papers (Barnes-Holmes et al., 2010; Hussey, Thompson, et al., 2015), and so only a brief overview will be provided here (see Hussey, 2020). On each block of trials, participants are presents with images or words at the top of the screen and in the middle of the screen. Response options are presented on the bottom left and bottom right hand sides of the screen, and are mapped to the left and right response keys. In order to progress to the next trial, the correct response must be given. Incorrect responses result in a red X being presented on screen. Between blocks of trials, this correct response changes so that, for example, participants must respond to “white people” and “dangerous” with “True” on one block and “False” on the other block. Participants complete pairs of these blocks in two phases: practice and testing. In order to progress from practice to testing, the participant must respond quickly and accurately on both blocks within the pair (typically with median reaction time < 2000 ms and percentage accuracy > 80%). Should they fail to meet this criteria, the participant completes another pair of practice blocks. Should they meet the criteria, they progress to the testing phase where they complete three pairs of blocks in a row. Following standard practice, only reaction time data from the test blocks is used in the analyses (Hussey, Thompson, et al., 2015).

## Data processing

IRAP studies typically using the *D* scoring method to convert each participant’s reaction times into analyzable values. The *D* score has some similarities to Cohen’s *d*, insofar as it is a trimmed and standardized difference in mean reaction time between the two block types. The specifics of the *D* score have been discussed in precise detail in other publications (Barnes-Holmes et al., 2010; Hussey, Thompson, et al., 2015) and therefore will only be summarized here. Its key points are that reaction times > 10,000 ms are trimmed, a mean reaction time is calculated for the trials in each block type, and a standard deviation is calculated for the pooled trials in both blocks. The difference between the means is then divided by the standard deviation, resulting in a *D* score.

# Results

## Meta-analytic strategy

All data processing and analyses were done in R (R Core Team, 2020). Interclass Correlation Coefficients were calculated using the psych package (Revelle, 2016). Meta-analyses were conducted using the metafor package (Viechtbauer, 2010, version 2.4-0) and Restricted Maximum Likelihood (REML) estimation. Meta-analysis of internal consistency estimates involved Bartlett transformations prior to analysis and inverse Bartlett transformations of meta-estimates. Analyses of test-retest reliability involved Fisher’s *r*-to-*z* transformations and inverse transformations. Heterogeneity metrics refer to transformed data, following standard practice.

## Internal consistency

As noted in the introduction, the IRAP’s internal consistency can be estimated by split-half reliability; however, multiple ways of splitting the data exist. Three ways were computed and are reported here, based on their relevance to making comparisons with the output of common software implementations of the IRAP, comparisons with other implicit measures, and to provide the most accurate estimate of internal consistency.

**Split-half via odd vs. even trials.** The modal strategy used in the IRAP literature is to use an odd-even split-half, where separate *D* scores are calculated for odd- and even-numbered trials by order of presentation, Pearson’s *r* correlations between these two sets of *D* scores are calculated, and then the Spearman-Brown correction is applied to adjust for test shortening (i.e., ). Multiple software implementations of the IRAP report this form of split-half *D* scores in their output. This result may be most useful when attempting to directly compare against results collected using the most common software implementations of the IRAP, although it does not necessarily represent the best estimate of the IRAP’s true internal consistency. When internal consistency was calculated using this method for each IRAP, the meta-analytic estimate of internal consistency was found to be poor: = .53, 95% CI [.42, .62].

**Split-half via first vs. second half.** Other popular implicit measures typically employ a different splitting method: the IAT’s split-half reliability is usually calculated by dividing the trials into the first- versus second-half of trials by order of presentation. Again, Pearson’s *r* correlations were then calculated between these two sets of *D* scores, and a Spearman-Brown correction was applied. This method is useful to calculate in order to directly compare the IRAP’s internal consistency to the IAT’s. Using this method, the meta-analytic estimate internal consistency was found to be very poor: = .33, 95% CI [.19, .45]. In contrast, a recent meta-analysis reported that the IAT’s internal consistency was substantively better and comparable with many self-report measures, α = .80 (Greenwald & Lai, 2020).

**Split-half via many permutations.** The large differences in the results found between these two methods (odd vs. even, first vs. second half) serves to highlight that the choice of splitting method is simultaneously arbitrary and yet has a significant impact on conclusions. Which method, if any, should researchers accept as providing more accurate results? Parsons et al. (2019) argued that no single decision need be made: instead of employing a single splitting method, a very large number of permutations of spits should be computed (e.g., 2000). In each permutation, the data is split into two randomly determined halves, *D* scores are calculated for each, Pearson’s *r* correlations are calculated from these two sets of *D* scores, and then a Spearman-Brown correlation is applied. A distribution of estimates is therefore obtained across permutations. This distribution is then parameterized: the mean value is used as the estimate, and the quantile method is used to find 95% Confidence Intervals. Parsons et al. (2019) noted that this method approximates Cronbach’s α, and remove assumptions associated with specific split strategies (e.g., regarding learning occurring with the task between the first vs. second half). The permutation method was therefore deemed to be the most appropriate strategy to accurately estimate the IRAP’s internal consistency, and its results were used for conclusions. Using the permutation method, the meta-analytic estimate of internal consistency was found to be poor, α = .53, 95% CI [.45, .61]. See Figure 1 (upper panel) for forest plot. A moderate degree of heterogeneity was also found between IRAPs, *Q*(*df* = 24) = 51.29, *p* = .001, 𝜏2 = 0.04, *I*2 = 56.3%, *H*2 = 2.3. This may suggest that internal consistency is moderated by unmodeled factors such as study domain, features of the stimulus set, task parameters, or other variables.

## Test-retest reliability

As noted in the introduction, Parson’s (2019) argues that test-retest reliability is better captured by the calculation of metrics of ‘Absolute Agreement’ (i.e., Interclass Correlation Coefficients) than simple correlations between timepoints, on the basis that correlations capture preservation of rank but not absolute changes in scores. Results suggested that test-retest reliability was poor, ICC = .40, 95% CI [.24, .57]. No heterogeneity was found between the two studies, *Q*(*df* = 1) = 0.00, *p* = .95, 𝜏2 = 0.00, *I*2 = 0.0%, *H*2 = 1. See Figure 1 (lower panel) for forest plot. The IRAP’s test-retest reliability therefore appears to be lower than the IAT’s (*r* = .50), according to the recent meta-analysis by Greenwald and Lai (2020). Of course, these results should be interpreted with caution given the limited range of studies, domains, follow-up periods, and sample sizes.



**Figure 1.** Forest plots of results.

# Discussion

Results summary.

## Comparisons with other meta-analyses

Results are consistent with Greenwald and Lai’s (2020) recent meta-analysis of the IRAP’s reliability, who reported a meta internal consistency of XXX and a test-retest reliability of XXX. Both point estimates and confidence intervals are comparable between these meta-analyses. Importantly however, inspection of Greenwald and Lai’s publicly available data demonstrated that my dataset and their contained no overlaps in the studies they considered. As such, this provides convergent evidence from two very large meta-analytic and individual-participant-data meta-analytic datasets that the IRAP’s reliability is poor.

## Implications for statistical power

A measures reliability has a direct relationship with its ability to detect true effects, and therefore the sample sizes needed for a given analysis. Parsons (REF) provides a primer on how reliability provides a ceiling for the associations among variables. if a criterion measure has a reliability of 1.0 (perfect), the max correlation that the IRAP can correlate with it is XX. True small/medium/large effects are observed as XX/XX/XX due to low IC. This also has power implications, bumping the required sample size to detect a small/medium/large true effect (assuming the criterion task has perfect reliability) from XX/XX/XX to XX/XX/XX.

## Ways to improve reliability

It seems important to consider ways in which the IRAP’s reliability could be improved. One possible and commonly recommended way of improving a tasks’ reliability is to increase its length, in this case by adding additional trials. The Spearman-Brown prediction formula can be rearranged to make a specific prediction about the relative change in task length that would be needed to obtain a given reliability estimate. Where refers to the goal reliability, refers to the current reliability, and refers to the multiple of current test length:

Using the meta-analytic estimate of the IRAP’s internal consistency (α = .54), in order to increase internal consistency to α = .70, .80, or .90, the task would need to be lengthened by a factor of 2.0, 3.4, or 7.6, respectively. Using the meta-analytic estimate of test-retest reliability (ICC = .38), in order to increase internal consistency to ICC = .70, .80, or .90, the task would need to be lengthened by a factor of 3.8, 6.5, or 14.7, respectively. In order to put these in context, the IRAP currently takes around 10 to 15 minutes to complete. These increases would therefore result in a task that would take between 20 minutes and two hours to complete, depending on the level of reliability desired. While technically possible, it is likely that lengthening the task in this manner would either put an unreasonable burden on participants or lower the tasks utility relative to information that could be collected via alternative methodologies.

[other methods, such as rethinking core features of the task in order to increase stimulus control, or using better scoring algorithms. Ultimately, reaction time based tasks are susceptible to noise, especially when they require participants to emit more complex relational responses.]

Consider other points that Lai or Nima include in their discussions

# 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.

Barnes-Holmes, D., & Hussey, I. (2016). The functional-cognitive meta-theoretical framework: Reflections, possible clarifications and how to move forward. *International Journal of Psychology*, *51*(1), 50–57. https://doi.org/10.1002/ijop.12166

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). *The case for formal methodology in scientific reform* [Preprint]. https://doi.org/10.1101/2020.04.26.048306

Gawronski, B., & De Houwer, J. (2011). Implicit measures in social and personality psychology. In C. M. Judd (Ed.), *Handbook of research methods in social and personality psychology* (Vol. 2). Cambridge University Press. 10.1017/CBO9780511996481.016

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

Hayes, S. C., & Brownstein, A. J. (1986). Mentalism, behavior-behavior relations and a behavior-analytic view of the purposes of science. *The Behavior Analyst*, *9*, 175–190.

Hussey, I. (2020). *The IRAP is not suitable for individual use due to very wide confidence intervals around D scores*. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I., Barnes-Holmes, D., & Barnes-Holmes, Y. (2015). From Relational Frame Theory to implicit attitudes and back again: Clarifying the link between RFT and IRAP research. *Current Opinion in Psychology*, *2*, 11–15. https://doi.org/10.1016/j.copsyc.2014.12.009

Hussey, I., & Hughes, S. (2020). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *In Press*. https://doi.org/10.31234/osf.io/7rbfp

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

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. http://CRAN.R-project.org/package=psych

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 59–65. https://doi.org/10.1016/j.jbtep.2015.01.004

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