The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis

Ian Hussey & Chad E. Drake

Evidence for the Implicit Relational Assessment Procedures’ (IRAP) reliability and validity is mixed, with one meta-analysis concluding it has good criterion validity and potential for clinical assessment, and two others concluding that it demonstrates low reliability. Here, we extend this evidence base through meta-analyses of published and unpublished studies. Individual participant data was used to estimate both internal consistency and test-retest reliability across a large number of IRAPs (*k* = 44) and participants (*N* = 1839). Results suggest that internal consistency for the task as a whole is poor (α = .49) and its test-retest reliability is very poor (ICC2 = .10). If scores are instead calculated for individual trial types, both forms of reliability are very poor (α = .27, ICC2 = .18). Ways to improve reliability were explored in the data: using stricter exclusion criteria, lengthening the task, using an alternative scoring algorithm, and using only one block order were not associated with increased reliability. Using fixed rather than moving response options was associated with higher, yet still poor, reliability. Implications for statistical power implications are considered. We conclude that researchers should be very cautious about choosing to employ the IRAP or when interpreting its results.

The study of implicit social cognition has become a mainstay of psychological research in many domains over the past twenty-five years (Greenwald & Banaji, 1995; Greenwald & Lai, 2020). In addition to the most popular measure, the Implicit Association Test (IAT: Greenwald et al., 1998), many other measures have been developed, each with unique features or benefits in mind (Nosek et al., 2011). Among them, the Implicit Relational Assessment Procedure (IRAP: Barnes-Holmes et al., 2010) is one of few implicit measures designed to capture implicit beliefs or automatic relational responding (Gawronski & De Houwer, 2011). That is, it can capture not only the strength of association between concepts, but also the nature of the relationship among them: for example, the distinction between “I am good” and “I want to be good” (Remue et al., 2013, 2014).

Although over seventeen years old and having been used in at least 155 published articles (Hussey, 2022a), the IRAP’s utility remains a matter of ongoing debate. On the one hand, in their meta-analysis of criterion validity of clinically relevant IRAP studies, Vahey et al. (2015) argued that the IRAP has potential as a tool for clinical assessment. However, on the other hand, concerns have been expressed about the IRAP’s low reliability (Golijani-Moghaddam et al., 2013; Greenwald & Lai, 2020) and poor individual-level estimation (Hussey, 2020). This tension between reliability and validity, and the importance of precise measurement more generally, has received renewed attention within psychology in recent years due to concerns about the replicability and validity of our findings – a so-called ‘Measurement Crisis’ (Flake & Fried, 2020). Importantly, recent research has emphasized that poor reliability can result in statistical effects that are highly replicable which nonetheless lead to false or invalid conclusions (Devezer et al., 2020; Hussey & Hughes, 2020), and that laboratory tasks like the IRAP have been highlighted as particularly susceptible to low reliability, understudied reliability, and low validity as a result of this (Lilienfeld & Strother, 2020). Vasey et al. (2003) argued that such tasks have “been granted psychometric free rein that would probably never be extended to researchers using other measures, such as questionnaires” (p.84). Quantification of the IRAP’s measurement properties is therefore important to interpreting the results of existing research, and its utility in future work, particularly in light of recent debates around what the IRAP’s utility is has claimed to be (Barnes-Holmes & Harte, 2022; Hussey, 2022b).

## 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 7 published studies containing 9 independent samples, including 318 participants for the meta-analysis of internal consistency and one study of 23 participants assessing test-retest reliability. Meta estimates of internal consistency (i.e., split-half reliability via Pearson’s *r* correlations with Spearman-Brown correlations) were = .65, 95% CI [.54, .74]. Just one study was found that reported test-retest reliability: Pearson’s *r* = .49, 95% CI [.10, .75] (NB confidence intervals were calculated here using the total sample size).

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, providing a total of 1016 participants for the meta-analysis of internal consistency and 74 participants assessing test-retest reliability in two different studies. We were able to computationally reproduce Greenwald & Lai’s (2020) estimates from their data and code for both internal consistency (Cronbach’s α = .60, 95% CI [.47, .69]) and for test-retest reliability (*r* = .43, 95% CI [.17, .63]).

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 conclusion that the IRAP’s reliability is problematically low, as both are below that typically accepted for assessment measures in psychology (Nunnally & Bernstein, 1994). 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. Nonetheless, given the very low number of studies (*k* = 2) and sample sizes (*N* = 74 total) used to estimate the IRAP’s test-retest reliability, this aspect of reliability in particular requires further research.

## The current research

Two factors suggest that there is need for additional assessment 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 revealing 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’ should be reported instead such as Intraclass Correlation Coefficients (i.e., ICC2: Parsons et al., 2019; Shrout & Fleiss, 1979)

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 the mean of this distribution of reliabilities. Importantly, this method approximates Cronbach’s α 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 may be addressed by conducting a file drawer meta-analysis. That is, where all studies – both published and unpublished – originating from an individual or group are used. We therefore pooled data from studies that we have been involved in.

# Method

## Data

### Data availability

All code and data needed to reproduce our analyses or reuse our data for new purposes is available on the Open Science Framework along with all details fo the measures ([osf.io/v3twe?view\_only=b19a0fc6d72845ac88917d5b003fc446](https://osf.io/v3twe?view_only=b19a0fc6d72845ac88917d5b003fc446)).

### Sources

Data was pooled from two sources. First, we took data from IRAP studies we the authors have been involved in and had access to. Inclusion criteria were (1) use of at least one IRAP, (2) access to raw data and the task parameters used in the study, and (3) necessary adherence to all standard operating procedures in the study’s design and data collection, as we have employed in all our other published IRAP research. This resulted in the omission of several projects that we felt did not meet publishable standards. Although the datasets employed here are not used to answer the research questions they were designed to address, we have included only studies that we considered publishable in principle (agnostic to their results). Exclusion criteria were embargos on data that are soon to be published, whose data could therefore not be made open for this meta-analysis. 3 studies excluded on this basis, 2 of which were in domains already present in other included datasets.

The second source of data was from published IRAP articles. This was done after reviewer feedback on a previous version of this manuscript raised the concern that data from the first source may not be sufficiently representative of the broader IRAP literature. We therefore contacted both the corresponding and senior authors of each of the 52 IRAP manuscript published in the last five years (i.e., 2018 to 2022) and asked them to contribute their data to a public repository where it could be re-used for other projects, including this one ([osf.io/nugzb](https://osf.io/nugzb/)). This list of articles was obtained from an existing, publicly available systematic search of the IRAP literature (Hussey, 2023b), and contained all still-active researchers listed as frequently authors of IRAP studies by a recent authorship analysis (Hussey, 2022b). This therefore represents a best effort to maximize the size and representativeness of the dataset available for these analyses. Unfortunately, only eight of 52 datasets were shared, and only two of those shared met the inclusion criteria (access to trial-level data and the task parameters used in the study).

### Prior publication

As a leading authority on best practices in meta-analyses, the Cochrane Handbook is explicit that meta-analyses should include data from sources other than published literature, as publication bias does not merely exclude low quality studies but instead is biased towards studies with non-significant results (Cochrane collaboration, 2022). Nonetheless, reviewers raised questions about whether data from unpublished studies may be of lower quality or drive down the meta-estimates of reliability. The publication status of each dataset is therefore listed in Table 1. 45.5% of the datasets employed here have already been published at time of writing. Data from 42.7% of participants came from already published studies.

Additionally, the results section contains analyses of whether reliability differed between published and unpublished studies (as will be discussed later, no significant differences were found).

## Participants

Data from 44 IRAPs across 19 different content domains (see Figure 1, upper panel) included a total of 1839 participants. Test-retest data was available for a subset of 8 domains with two different follow-up periods: immediate (7 domains) and 1-week (1 domain; see Figure 1, lower panel). 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 (64.4% female, 35.4% male, 0.2% non-binary), young adults (*M*age = 19.9, *SD* = 3.8), White (50.0%; 32.1% Black, 9.8% Hispanic, 3.0% Asian, 5.1% other), and heterosexual (88.4%; 7.8% bisexual, 3.9% homosexual).

## Measures and procedure

General description of the IRAP. Like many implicit measures, the IRAP is a computer-based task that uses reaction time differentials to calculate scores. Participants are instructed to respond as quickly and accurately as possible. On each trial, category stimuli are presented at the top of the screen and attribute stimuli are presented in the middle of the screen. Response option are presented at the bottom left and right hand sides of the screen, and are mapped to the left and right response keys (typically the ‘D’ and ‘K’ keys). Correct responses alternate between blocks of trials. For example, a race IRAP might employ “White people” and “Black people” as category stimuli and positive and negative words as attribute stimuli, with the response options “True” and “False”. Correct responses are required to proceed to the next trial. Incorrect responses result in a red X being presented on screen. As such, participants would typically be required to respond to “White people” and “dangerous” with “True” on one block and “False” on the subsequent block. Blocks typically consist of 24, 36, or 48 trials depending on the number of stimuli exemplars employed, and use an equal number of combinations of the category and attribute stimuli (e.g., White people – positive, White people – negative, Black people – positive, and Black people – negative). Participants typically complete between one and four pairs of practice blocks until they meet performance criteria (e.g., median reaction time ≤ 2000 ms and percentage accuracy ≥ 78% in both block types), followed by three pairs of test blocks from which scores are calculated. The IRAP’s procedural details and variations have been discussed in detail elsewhere (Barnes-Holmes et al., 2010; Hussey et al., 2015).

Specifics of each study. All data collection was performed in individual experimental cubicles. Data collection took place at sites in Ireland, the USA, the United Kingdom, Belgium, and Brazil. This is broadly representative of where data collection takes place in other published IRAP studies (i.e., because this study makes use of published data from published studies, and unpublished data came from research groups with ongoing IRAP research interests).

Details of each dataset are provided in a single spreadsheet in the supplementary materials (<https://osf.io/7nz5u?view_only=b19a0fc6d72845ac88917d5b003fc446>). This includes details of use of the data in a prior publication (i.e., references for the publication in which data was already used, references for the publications that used an identical IRAP, if available), all details needed to reconstruct each IRAP (i.e., all word and image stimuli, responding rules, response options, response option locations, accuracy feedback messages, latency feedback messages, number of trials per block, number of pairs of test blocks, practice block accuracy criterion, practice block speed criterion, and block order), and details of the sample and context (i.e., sampling strategy, remuneration, location, year of data collection). In addition to this, software implementations of the IRAP used for data collection in the different studies and an experimenters’ script used to train experimenters can also be found in the supplementary materials ([osf.io/v3twe?view\_only=b19a0fc6d72845ac88917d5b003fc446](https://osf.io/v3twe?view_only=b19a0fc6d72845ac88917d5b003fc446)).

## Data processing

IRAP studies typically use the *D* scoring method to convert each participant’s reaction times into analyzable scores (see Barnes-Holmes et al., 2010; Hussey et al., 2015). 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. 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.

Participants were excluded if their median reaction times in the IRAP test blocks was > 2000 ms or if their accuracy was < 78% in either of the block types (i.e., the typical strategy employed in the IRAP literature: Barnes-Holmes et al., 2010; Hussey et al., 2015). A total of 180 participants (8.7%) were excluded on this basis leaving 1839 participants in the internal consistency sample and 318 in the test-retest sample.

# Results

## Meta-analytic strategy

All data processing and analyses were done in R (R Core Team, 2022). Intraclass Correlation Coefficients were calculated using the psych package (Revelle, 2016). Meta-analyses were conducted using the metafor package (Viechtbauer, 2010, version 3.8-1) 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 for reporting. Analyses of test-retest reliability using Pearson’s *r* correlations involved Fisher’s *r*-to-*z* transformations and inverse transformations. Heterogeneity metrics refer to heterogeneity in the transformed estimates.

## 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 and with other implicit measures and previously published work, and to provide a more 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 reported in most published research, 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: = .54, 95% CI [.47, .60], 95% PI [.18, .74], 𝜏2 = 0.08, *I*2 = 44.6%, *H*2 = 1.8. Similar results were found when excluding outliers (which are discussed later), = .50, 95% CI [.44, .56], 95% PI [.30, .65], 𝜏2 = 0.03, *I*2 = 20.2%, *H*2 = 1.3.

**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 of internal consistency was found to be very poor: = .48, 95% CI [.41, .54], 95% PI [.15, .68], 𝜏2 = 0.06, *I*2 = 36.3%, *H*2 = 1.6. Similar results were found when excluding outliers (which are discussed later), = .46, 95% CI [.39, .52], 95% PI [.22, .62], 𝜏2 = 0.03, *I*2 = 21.7%, *H*2 = 1.3. In contrast, a recent meta-analysis reported that the IAT’s internal consistency, when calculated using this method, was substantively better (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 splits should be computed (i.e., 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). Using the permutation method, the meta-analytic estimate of internal consistency was found to be poor, α = .53, 95% CI [.47, .58], 95% PI [.29, .69]. A moderate degree of heterogeneity was found between estimates, *Q*(*df* = 43) = 62.47, *p* = .028, 𝜏2 = 0.04, *I*2 = 28.6%, *H*2 = 1.4.

Metafor’s ‘influence’ function was then used to detect effect sizes with undue influence on the meta-effect size. This produces multiple metrics which were used to define outliers, specifically Cook’s distance if deleted, 𝜏2 if deleted, and DFFITS if deleted (i.e., how many standard deviations the meta-effect size changes). Results indicated that three effect sizes exerted undue influence on the meta-effect size and were excluded as outliers: Sexuality IRAP 1: α = .85, 95% CI [.66, .93], Sexuality IRAP 2: α = .94, 95% CI [.84, .97]), and Gender IRAP 3: α = .64, 95% CI [.52, .73].

When outliers were excluded, the meta-analytic estimate of internal consistency was found to be poor, α = .49, 95% CI [.43, .54], 95% PI [.43, .54], with no heterogeneity, *Q*(*df* = 40) = 29.69, *p* = .884, 𝜏2 = 0.00, *I*2 = 0.0%, *H*2 = 1.0. The exclusion out outliers therefore successfully removed observed heterogeneity. See Figure 1 (upper panel) for Forest plot.

Due to the combination of the permutation-based split-half method and the exclusions of outliers, this represents the most appropriate estimate of the IRAP’s internal consistency among those we have reported here. Subsequent calculations and core conclusions are therefore based on this estimate, and this model and the included effect sizes are used as the basis of other variations (e.g., multilevel and moderator meta-analyses reported later).

**Figure 1.** Forest plots.





**Reliability of individual trial types.** The preceding analyses followed the convention set by previous meta-analyses of the IRAP’s reliability by quantifying the reliability of the task as a whole (Golijani-Moghaddam et al., 2013; Greenwald & Lai, 2020). That is, a single estimate was calculated for each IRAP. However, IRAP data is often analyzed separately for each of its four trial types, which are akin to subscales. For example, a death IRAP assesses overall evaluations of death versus life, but its individual trial types assess the positivity of death, the negativity of death, the positivity of life, and the negativity of life. Insofar as some researchers argue that IRAP data should be scored and analyzed at the trial type level, so too are they likely to argue that the IRAP’s reliability should be estimated at the trial type level. However, it is worth noting that there is actually little evidence that the IRAP trial types are indeed independent in this way (Hussey, 2023c). Additionally, calculating each reliability estimate from fewer trials is likely to lower reliability. Nonetheless, given that this argument is frequently repeated, and on the basis of a reviewer request, we therefore estimated reliability at the trial type level too.

Estimates of the reliability of individual trial types were calculated by applying the split-half via many permutations method again, this time applied to *D* scores calculated for each trial type in each domain. This maximized the ability to compare the results of this analysis with the previous one.

Results were again meta-analyzed, this time with a three-level meta-analytic model that included a random intercept for domain. This acknowledged the non-independence of the four estimates derived from each domain, one for each IRAP trial type. The meta-analytic estimate of internal consistency at the trial type level was very poor, α = .27, 95% CI [.23, .30], 95% PI [.23, .30]. No heterogeneity was observed, *Q*(df = 163) = 79.94, *p* = .999, suggesting that it was appropriate to meta-analyse across trial types (i.e., zero heterogeneity suggests that there are not unmodelled differences between the trial types). See Figure 2 for a caterpillar plot of the trial type level reliability estimates.

**Figure 2.** Caterpillar plot of internal consistency scored at the level of the individual trial types.



## 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., Intraclass Correlation Coefficients) than simple correlations between timepoints, on the basis that correlations capture preservation of rank but not absolute changes in scores. Meta-analyses of both Pearson’s *r* correlations and ICCs are reported here, based on their relevance to making direct comparisons with previously published work and to provide a more accurate estimate of test-retest reliability, respectively.

**Test-retest via Pearson’s *r*.** Results suggested that test-retest reliability was very poor and with substantial heterogeneity, *r* = .12, 95% CI [-.11, .34], 95% PI [-.45, .62], 𝜏2 = 0.08, *I*2 = 73.7%, *H*2 = 3.8. Test-retest correlations were negative for three IRAPs (i.e., gender, body image, and race). This result may be most useful when attempting to directly compare against previously published research, which has typically used Pearson’s *r* correlations, although it does not necessarily represent the best estimate of the IRAP’s true internal consistency.

The IRAP’s test-retest reliability therefore appears to be significantly lower than the IAT’s (*r* = .50) according to the recent review by Greenwald and Lai (2020).

**Test-retest via ICC.** When using ICCs, results also suggested that test-retest reliability was very poor and with substantial heterogeneity, ICC2 = .10, 95% CI [-.11, .32], 95% PI [-.48, .69], 𝜏2 = 0.08, *I*2 = 78.8%, *H*2 = 4.7. Test-retest was near zero or negative for half of the IRAPs (see Figure 1, lower panel).

Outlier metrics (Cook’s distance if deleted, 𝜏2 if deleted, and DFFITS if deleted) did not flag any IRAPs as likely outliers. As such, this heterogeneity may be attributable to other unmodeled factors, such as the domain, follow-up period, features of the stimulus set or task parameters, or others.

Due to the combination of ICC and outlier analysis, this represents the most appropriate estimate of the IRAP’s test-retest reliability among the two we have reported here. Subsequent calculations and conclusions are therefore based on this estimate, and this model and the included effect sizes are used as the basis of other variations (e.g., multilevel and moderator meta-analyses reported later).

**Reliability of individual trial types.** Similar to internal consistency, the test-retest reliability was also estimated for IRAP data scored at the individual trial type level. This was done using the ICC method and a comparable three-level meta-analytic model with random intercept for domain. Results suggested that test-retest reliability was very poor, ICC2 = .18, 95% CI [.09, .27], 95% PI [-.05, .41]. Heterogeneity was detected, *Q*(df = 31) = 47.14, *p* = .032.

## Differences between published and unpublished studies

In order to assess whether reliability systematically differed between published and unpublished studies, we therefore conducted a moderator meta-analysis that compared estimates from these two subgroups.

No differences in internal consistency were found between published (α = .44, 95% CI [.33, .53]) and unpublished studies published (α = .51, 95% CI [.40, .60]), *QM*(df = 1) = 1.740, *p* = .187. Similarly, no differences in test-retest reliability were found between published (ICC2 = .21, 95% CI [.00, .42]) and unpublished studies published (ICC2 = .00, 95% CI [.00, .32]), *QM*(df = 1) = 1.603, *p* = .201.

## The IRAP’s individual-level utility

An under-appreciated implication Test-Retest Reliability is that it can be used to determine a measure’s utility at the level of individuals. In combination with the Standard Deviation of participants’ scores on that measure, the measure’s Test-Retest Reliability can be used to calculate the Standard Error of Measurement (SEM; not to be confused with the standard error of the mean) using the following equation:

SEM = SD ×

The SEM can then be used to construct confidence intervals on individuals’ scores:

95% CI = score SEM × 1.96

Given an observed weighted average SD for overall *D* scores was SD = 0.22, and test-retest reliability for overall *D* scores was ICC2 = 0.10, the precision of an individual’s score can be estiamted to be their overall *D* score ± 0.41. To understand the implications of this, we can use it to interpret the average participant’s overall *D* score (weighted mean across domains = 0.10, i.e., slightly faster on the consistent blocks than the inconsistent blocks). However, due to the task’s poor reliability, their score can actually only be said to be somewhere interval *D* = 95% [-0.31, 0.51]. That is, anywhere between much faster on the consistent blocks and much faster on the inconsistent blocks. This is particularly problematic when we place it in the context of trying to differentiate this average participant’s score from other participants. 95% of all the 1839 participants’ observed overall *D* scores fell within the range -0.37 to 0.57. As such, this means that estimation of the average participant’s score covers most of the range (87.23%) of all observed participants. Therefore the average individual IRAP participant cannot be distinguished from the *D* scores demonstrated by the *great majority* of other participants’ scores, from extremely positive to extremely negative.

This situation is substantially worsened if one prefers to employ trial-type level *D* scores, whose precision implied by SD = 0.36 and ICC2 = 0.18, was trial-type level *D* score ± 0.64. 95% of all observed trial-type *D* scores fell within the range -0.71 to 0.89. As such, this means that estimation of the average participant’s score covers *more* than the range (136.17%) of all observed participants. Therefore the average individual IRAP participant cannot be distinguished from the *D* scores demonstrated by the *any* of the other participants scores, from extremely positive to extremely negative.

This uncertainty around individual’s scores strongly limits the IRAP’s utility to make inferences about individuals rather than groups. Similar arguments have been made using other individual level estimation methods (i.e., using bootstrapping approaches, see Hussey, 2020).

## Possible ways to improve reliability

Given the low reliability estimates observed, it seems important to explore ways in which the IRAP’s reliability could be improved. This list is by no means exhaustive: it represents analyses and suggestions that were possible with the existing data.

**Use stricter performance exclusion criteria.** Published IRAP studies have employed a variety of different performance exclusion criteria. For example, differences in cut-offs (e.g., median reaction time < 2000 ms in some studies and 2500 in others), or applying these criteria differently (e.g., applying this criterion to both block types across the test blocks vs. block types and in each individual block). The main analyses here employed the most common strategy of requiring each participant to meet the criteria (accuracy ≥ 80% and median reaction time ≤ 2000 ms) in both block types (consistent vs. inconsistent) across the test blocks. A more stringent criteria would be to require them to meet these criteria in every one of the individual blocks and exclude them if they failed to maintain them in one or more blocks. This represents quite a strict criteria relative to what has typically been used in published IRAP research.

In order to assess whether these stricter criteria influence reliability estimates, we constructed a multi-level moderator meta-analysis model that directly compared estimates calculated using the typical versus stricter method. IRAP type was used as a random intercept in the model to acknowledge the non-independence of the pair of estimates produced from each IRAP. Exclusion strategy was entered as a moderator. Results demonstrated that no significant differences were detectable between the typical and the stricter exclusion strategies, *QM*(df = 1) = 0.68, *p* = .412. The stricter strategy produced a numerically lower reliability (α = .46, 95% CI [.36, .54]) than the typical strategy (α = .49, 95% CI [.43, .55]). As such, there is no evidence that tightening the exclusion strategy increases the IRAP’s reliability.

**Lengthen the task.** One possible and commonly recommended way of improving a tasks’ reliability is to increase its length. In this case this would involve adding additional trials to the IRAP. 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 (a rearrangmenet of Revelle, 2009, equation 7.12):

Using the meta-analytic estimate of the IRAP’s internal consistency (α = .49), in order to increase internal consistency to α = .70, the task would need to contain 2.4 times the number of trials it currently does. Using the meta-analytic estimate of test-retest reliability (ICC2 = .10), in order to increase internal consistency to ICC2 = .70, the task would need to contain 21 times the number of trials it currently does. 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 24 minutes and 5.5 hours to complete, depending on the type of reliability desired. While technically possible, this may either put an unreasonable burden on participants or lower the tasks utility relative to information that could be collected via alternative methodologies. This estimation also assumes no fatigue effects, which may lower the reliability. It therefore seemed useful to explore alternative ways to improve reliability.

**Use a more robust scoring method.** Recent research has argued that the *D* score is overly sensitive to the outliers that are frequently observed in reaction time data (De Schryver et al., 2018), and has suggested a more robust scoring method as an alternative. This method has been referred to by several names, including the Probabilistic Index, the Probability of Superiority and Ruscio’s A (Ruscio, 2008). This non-parametric scoring method has a straightforward interpretation and method of calculation: it is the probability that a randomly selected reaction time in one block type is longer than a randomly selected reaction time in the other block type. We therefore calculated A scores for each IRAP using code provided in the RProbSup R package (Ruscio, 2019). We then assessed whether internal consistency was different between *D* and A scores (NB changes in test-retest reliability were not calculated due to much lower sample size and therefore statistical power). This was done using a multilevel moderator meta-analysis model. A random intercept was used to acknowledge the non-independence of the scores produced using data from each IRAP. Scoring method was entered as a moderator. No differences were observed in internal consistency between the two scoring methods, *D* scores: α = .50, 95% CI [.43, .56], A scores: α = .53, 95% CI [.46, .59], *Q*M(*df* = 1) = 0.734, *p* = .392.

**Use only one block order.** The IRAP presents pairs of blocks in which the required response switches between those blocks (e.g., responding to ‘White people’ and ‘positive’ with ‘True’ on one block and ‘False’ on the other). Which block each participant first encounters is often randomized between participants, on the basis that block order has sometimes been shown to have an influence on mean *D* scores. These blocks have in the past often been referred to as being assumed to be ‘consistent’ versus ‘inconsistent’ with participants’ learning histories. Although this terminology is common, we have avoided it in this article until now on the basis that we feel that it can confuse aspects of the procedure and results (i.e., consistency with learning history should be derived from the results rather than assumed). As such, it is important to note that the ‘consistent’ block order is an imposition of the researcher’s expectations rather than a conclusion based on the data. Nonetheless, this variable is commonly recorded and reported in articles, and it may be the case that internal consistency results differ based on block order. The data used for the internal consistency sensitivity meta-analysis was therefore split into two groups: participants who received the consistent-first vs. the inconsistent first block order. Permuted internal consistency estimates were again calculated, and then compared in a multilevel moderator meta-analysis, with IRAP type as random intercept and block order as moderator. Only IRAPs which contained both block type orders between participants were considered. No differences in internal consistency were observed between the block orders; consistent block first: α = .45, 95% [.33, .54], inconsistent block first: α = .50, 95% CI [.33, .62], *Q*M (*df* = 1) = 0.444, *p* = .506.

**Fix the location of the response options.** Finally, another commonly reported variation in the IRAP’s procedural features is whether the response options (e.g., True and False) were either static (e.g., True always on the left, False on the right) or whether they swapped sides pseudorandomly between trials. Roughly one third of the studies in our dataset used static response options, and two thirds used moving. Although it is not often discussed within published articles, informal discussion among IRAP researchers around the decision to use static or moving response options has often been that, on the one hand, static response options appear to make the task easier to complete and perhaps therefore reduces noise in reaction times. But, on the other hand, static response options may allow participants to privately recode the response options in order to make the task easier for themselves (e.g., treating the ‘True’ response as if it is labelled ‘False’ to make responding in the history-inconsistent blocks easier). This provided a testable hypothesis, that internal consistency would be higher when response options were static. The permutated estimates from the internal consistency meta-analysis were used in a moderator meta-analysis that added response option location as a moderator. A difference in internal consistency was observed between static (α = .59, 95% [.50, .66]) versus moving response options (α = .45, 95% CI [.30, .56]), *Q*M (*df* = 1) = 6.378, *p* = .012. This suggests that IRAPs with fixed response options may produce higher internal consistency.

# Discussion

Results demonstrate that the IRAP’s internal consistency is poor and its test-retest reliability is unacceptably low. In half of the domains, test-retest reliability was zero or negative. This work has several benefits compared to previous meta-analyses: (a) it is the largest analysis to date, (b) it used more optimal analytic methods, and (c) it is computationally reproducible due to sharing both data and code.

Our best estimate of internal consistency (α = .49, 95% CI [.43, .54]) was smaller than that those reported in both previously published meta-analyses ( = .65, 95% CI [.54, .74]: Golijani-Moghaddam et al., 2013; (α = .60, 95% CI [.47, .69]: Greenwald & Lai, 2020).

Our best estimate of test-retest reliability (ICC = .10, 95% CI [-.11, .32]) was much lower than those reported by either previously published meta-analysis (*r* = .49, 95% CI [.10, .75]: Golijani-Moghaddam et al., 2013; *r* = .43, 95% CI [.17, .63]: Greenwald & Lai, 2020). However, both these previous meta-analyses were based on a much lower number of studies (*k* = 2) and sample sizes (*N* = 73 total) than the current estimate (*k* = 8, *N* = 354). Differences in results may therefore be due to one or more features of our work relative to previous research, such as our larger sample size and variety of domains or our more advanced statistical methods (e.g., controlling for absolute change between timepoints, use of permutation-resampling to avoid arbitrary choices in split-half, or assessment of outliers). Results were not explained by the inclusion of data from both published and unpublished sources, as tests of moderation by publication status were non-significant for both forms of reliability. While our results differ from previous meta-analyses to some degree, the conclusions of all three, with increasingly large sample sizes over time, agree that the IRAP’s internal consistency and test-retest reliability is poor at best.

Our access to the trial level data also allowed us to score the data in other ways than previous meta analyses. Insofar as some researchers argue that IRAP data should be scored and analyzed at the trial type level, such researchers should also quantify its reliability from data scored at the trial type level (which are akin to subscales). When scored this way, both internal consistency (α = .27, 95% CI [.23, .30]) and test-retest reliability (ICC2 = .18, 95% CI [.09, .27]) were very poor.

## Implications of low reliability for statistical power

An underappreciated fact is that a measure’s reliability has a direct relationship with its ability to detect true effects (i.e., statistical power), and therefore the sample sizes needed for a given analysis. Parsons (2018) provides a useful discussion of how reliability provides a ceiling for the associations among variables. The average observable correlation among two measures and (i.e., ) is a function of the true correlation () and also the reliability of both measures (i.e., their self-correlation and (Revelle, 2009, equation 7.3)

We can imagine that one measure is the IRAP and the other is some external variable of interest, such as a disgust-related behavioral approach task (Nicholson & Barnes-Holmes, 2012). If we put aside the reliability of the external variable (e.g., imagine it is perfect, = 1), we can use our meta-analyzed estimates of the IRAP’s reliability to estimate the maximum average observable correlations that could be observed between the two (i.e., when = 1). No one form of reliability fully captures a measure of global reliability, so it is useful to calculate estimates using estimates for both test-retest reliability (ICC2 = .10) and internal consistency (α = .49). The maximum average observable correlation between the IRAP and a criterion task, assuming a true correlation between them of 1 and that the criterion task had perfect reliability, was estimated to be *r* = .32 and .70, respectively. That is to say, average correlations larger than this cannot be observed due to the limits of the IRAP’s reliability. In all other (realistic) circumstances, such as where the true correlation is lower than 1 and the reliability of the criterion task is lower than 1, the average observable correlations would be lower than this again.

Maximum observable correlations could also be calculated for other true correlations; these would be also be scaled downward to a comparable degree as those for perfect true correlations. For example, using Cohen’s (1988) rules of thumb, a large true correlation ( = .50) would imply maximum observable correlations of *r* = .15 and .33, respectively. These large reductions in the actual observed correlation among variables must then be considered when choosing sample sizes – loosely speaking, in order to detect what is in reality a ‘large’ effect size, the researcher may have to power the study to detect ‘small’ effect sizes. Tasks with low reliability, such as the IRAP, therefore place studies under increased data collection burdens or lower statistical power to detect true effects.

This also raises important questions about the credibility of correlations in the literature which are larger than this. For example, Vahey et al. (2015) reported a meta-analyzed observed average correlation of *r* = .45 between the IRAP and clinical criterion variables (i.e., not corrected for measurement error). Taking if we take our estimate of the IRAP’s internal reliability (α = .49), and (somewhat implausibly) assume that the criterion tasks all had perfect reliability ( = 1), this would imply that the true average correlation between the IRAP and clinical criterion tasks is *r* = .67. This would imply that either (a) the IRAP has truly exceptional criterion validity despite its low reliability, and that it outperforms the vast majority of other classes of measures across psychological science, or (b) Vahey et al.’s (2015) estimate of criterion validity is overestimated in some way. Future research may therefore wish to consider assessing the computational reproducibility of Vahey et al.’s (2015) results in order to ensure that they were not obtained in error. Existing unpublished research has already suggested that their results are not computationally replicable and several errors are present (Hussey, 2023a).

## Possible ways to improve reliability

We also considered multiple ways in which reliability could be improved. Lengthening the task to increase reliability is a common recommendation. However, depending on the type and degree of reliability that is sought, this may be less feasible in this case. Results suggest that the IRAP would need to be nearly five and a half hours long for it to provide high test-retest reliability. This is likely to be at odds with the goals and pragmatic considerations of many forms of research. We also used moderator meta-analyses to explore whether four factors might increase internal consistency. First, based on a reviewer suggestion, we assessed whether using more stringent performance exclusion criteria would increase reliability. However, no significant improvement was found. Numerical differences favored the typical exclusion strategy over the stricter one. Second, based on the recommendations of De Schryver et al. (2018), we implemented a robust scoring algorithm as an alternative to the *D* score. However, no significant improvement in internal consistency was found. Numerical differences slightly favored PI scores over *D* scores. Prior work by De Schryver et al. (2018) suggests that the PI should be more robust to outliers, and may therefore warrant use as it is at worst non-inferior to the *D* score. Third and fourth, we assessed whether two commonly manipulated procedural parameters might increase internal consistency. No significant differences were found for the order in which participants completed the blocks. Results suggested that higher reliability was observed when response option locations were fixed rather than moving. All other things being equal, future IRAP research should therefore consider using fixed response option locations. Nonetheless, even the meta-estimate of reliability when using fixed locations (α = .59, 95% CI [.50, .66]) remained to be much lower than in both the most popular implicit measure, the IAT (α = .80: Greenwald & Lai, 2020), and the typically recommended minimum cut-off values for psychological measures (e.g., α > .7, .8, or .9: Nunnally & Bernstein, 1994).

Of course, other approaches to improving the IRAP’s reliability are possible and may be more effective, and could be explored in future research. Lessons could be learned from existing literature using similar tasks. For example, some versions of the Brief IAT have discarded data from the first few trials in each block as they tend to be slower and noisier than subsequent trials (Nosek et al., 2013). Other avenues of work would be to consider how to exert better stimulus control over responding within responding IRAP-like tasks such as which practice performance criteria are employed; or features of the stimuli employed (e.g., their complexity or readability). Research has already shown that many more task features serve as important sources of stimulus control over behavior within the task than was initially thought. For example, the dimension along which the two category stimuli are related factor into IRAP performance (even though the task never requires the participant to emit this relational response, see Hussey et al., 2016); or the instructions presented before each block that specify the responding rules for that block (Finn et al., 2016). While these and other sources of stimulus control over behavior within the task have been demonstrated, no work has used these to increase the reliability of behavior within the IRAP.

## Conclusions

Measurement is a cornerstone of the scientific method, even in fields that do not always explicate this importance. For example, even the behaviorist working with rats in Skinner boxes must be concerned with whether the lever functions well to capture the animal’s lever-pressing behavior: if the lever is too heavy or its action is too stiff, the acquisition curve recorded would not accurately reflect the animal’s behavior. Even fields of research that have at times been skeptical of the utility of psychometric methods (e.g., behaviorism, from which the IRAP emerged) are therefore negatively impacted by low reliability and poor measurement. As such, the IRAP’s poor reliability has implications for all IRAP research, regardless of whether the task is being used as a measure of implicit social cognition or relational responding (Hughes et al., 2012; Barnes-Holmes & Harte, 2022; although see Hussey, 2022b).

Vahey et al.’s (2015) meta-analysis of the IRAP’s criterion validity concluded that the task shows promise as a clinical assessment measure. However, a degree of reliability is a prerequisite for validity (Loevinger, 1957). The results of this and two previous meta-analyses suggest that the IRAP’s reliability is poor at best and unacceptably low at worst. This poor reliability has direct negative implications for statistical power in past and future studies. Elsewhere, recent research has also suggested that the IRAP demonstrates very poor individual level estimation (Hussey, 2020), which is coherent with the current results given the mathematical relationship between reliability and individual level estimation (e.g., the Standard Error of Measurement, see: Revelle & Condon, 2019). As such, in its current form, the IRAP likely has limited use as an assessment tool in either research or applied settings. Researchers should be very cautious when choosing to use the IRAP in their research or when interpreting the results of IRAP studies. The IRAP may represent another example of what Lilienfeld and Strother (2020) describe as cautionary tales in psychological measurement: a laboratory procedure that researchers used to make substantive conclusions without first ensuring that it had adequate measurement properties to do so.

# Author notes

## Funding

IH was supported by Ghent University grant 01P05517 and the META-REP Priority Program of the German Research Foundation (#464488178).

## Conflict of interest

None.

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

Barnes-Holmes, D., & Harte, C. (2022). 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

Cochrane collaboration. (2022). Chapter 4: Searching for and selecting studies. In *Cochrane Handbook for Systematic Reviews of Interventions* (Version 6.3). https://training.cochrane.org/handbook/current/chapter-04#section-4-3-2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

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

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

Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016). Exploring the behavioral dynamics of the Implicit Relational Assessment Procedure: The impact of three types of introductory rules. *The Psychological Record*, *2*, 309–321.

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4.

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

Hughes, S., Barnes-Holmes, D., & Vahey, N. A. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science*, *1*(1–2), 17–38. https://doi.org/10.1016/j.jcbs.2012.09.003

Hussey, I. (2020). The Implicit Relational Assessment Procedure is not suitable for individual use. *PsyArXiv*. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I. (2022a). *A reproducible systematic review of research using the Implicit Relational Assessment Procedure (IRAP)*. https://doi.org/10.17605/OSF.IO/TNA8S

Hussey, I. (2022b). 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. (2023a). *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.”* https://osf.io/jg8td/

Hussey, I. (2023b). *A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using the Implicit Relational Assessment Procedure*. PsyArXiv. https://doi.org/10.31234/osf.io/g2x9p

Hussey, I. (2023c). *The Implicit Relational Assessment Procedure’s trial-types are not independent*. PsyArXiv. https://doi.org/10.31234/osf.io/4gsxq

Hussey, I., & Hughes, S. (2020). Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology. *Advances in Methods and Practices in Psychological Science*, 2515245919882903. https://doi.org/10.1177/2515245919882903

Hussey, I., Ní Mhaoileoin, D., Barnes-Holmes, D., Ohtsuki, T., Kishita, N., Hughes, S., & Murphy, C. (2016). The IRAP Is Nonrelative but not Acontextual: Changes to the Contrast Category Influence Men’s Dehumanization of Women. *The Psychological Record*, *66*(2), 291–299. https://doi.org/10.1007/s40732-016-0171-6

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

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology / Psychologie Canadienne*, *61*, 281–288. https://doi.org/10.1037/cap0000236

Loevinger, J. (1957). Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Nicholson, E., & Barnes-Holmes, D. (2012). Developing an implicit measure of disgust propensity and disgust sensitivity: Examining the role of implicit disgust propensity and sensitivity in obsessive-compulsive tendencies. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(3), 922–930. https://doi.org/10.1016/j.jbtep.2012.02.001

Nosek, B. A., Bar-Anan, Y., Sriram, N., & Greenwald, A. G. (2013). Understanding and using the Brief Implicit Association Test: Recommended scoring procedures. *Preprint*. http://ssrn.com/abstract=2196002

Nosek, B. A., Hawkins, C. B., & Frazier, R. S. (2011). Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences*, *15*(4), 152–159. https://doi.org/10.1016/j.tics.2011.01.005

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd edition). McGraw-Hill.

Parsons, S. (2018). Visualising two approaches to explore reliability-power relationships. *PsyArXiv*. 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. (2022). *R: A language and environment for statistical computing* (4.2.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

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

Remue, J., Hughes, S., De Houwer, J., & De Raedt, R. (2014). To be or want to be: Disentangling the role of actual versus ideal self in implicit self-esteem. *PLoS ONE*, *9*(9), e108837. https://doi.org/10.1371/journal.pone.0108837

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

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

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*, 1395–1411. https://doi.org/10.1037/pas0000754

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Ruscio, J. (2019). *RProbSup: Calculates Probability of Superiority* (2.1) [Computer software]. https://CRAN.R-project.org/package=RProbSup

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

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

Vasey, M. W., Dalgleish, T., & Silverman, W. K. (2003). Research on Information-Processing Factors in Child and Adolescent Psychopathology: A Critical Commentary. *Journal of Clinical Child & Adolescent Psychology*, *32*(1), 81–93. https://doi.org/10.1207/S15374424JCCP3201\_08

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