# The Implicit Relational Assessment Procedure is not suitable for individual use

# due to very wide confidence intervals around *D* scores

# Ian Hussey

Author note: Ian Hussey, Ghent University, Henri Dunantlaan 2, Gent 9000, Belgium. [ian.hussey@rub.de](mailto:ian.hussey@rub.de). IH was supported by Ghent University grant 01P05517 and the META-REP Priority Program of the German Research Foundation (#464488178).

# Abstract

Vahey et al.’s (2015) meta-analysis suggested that the Implicit Relational Assessment Procedure has potential “as a tool for clinical assessment”. Here I present evidence to the contrary. Using a large open dataset, 95% Confidence Intervals were calculated for each participant’s *D* scores via bootstrapping. Results demonstrate that Confidence Intervals are extremely wide; only a small fraction of individuals were shown to have non-zero IRAP effects; only a small fraction of individuals’ *D* scores were discriminable from other individuals’ *D* scores; and individuals’ confidence intervals spanned roughly half the total observed range. An alternative scoring algorithm, the Probabilistic Index (PI), did not meaningfully improve these metrics. Lastly, the IRAP was also shown to be greatly inferior to the most popular implicit measure, the Implicit Association Test, on all metrics of individual level utility. The IRAP is therefore very unsuitable for individual level use or assessment.

# The Implicit Relational Assessment Procedure is not suitable for individual use

# due to very wide confidence intervals around *D* scores

The Implicit Relational Assessment Procedure (IRAP) is a reaction-time based task that has seen significant use as both a measure of implicit attitudes and within Relational Frame Theory research as a measure of relational responding (Barnes-Holmes & Harte, 2022; Hussey, Barnes-Holmes, et al., 2015). There is increasing interest in using the IRAP at the individual level. For example, in their meta-analysis of clinically relevant IRAP studies, Vahey et al. (2015) concluded that the IRAP has potential “as a tool for clinical assessment” (p.64). Subsequently, a recent study has reported individual level analyses of IRAP data (Finn et al., 2019).

However, for the IRAP to have individual-level utility, whether that be for clinical assessment, research use or otherwise, scores produced on the task by a single individual would need to be well estimated. Unfortunately, there is good *a priori* reason to believe that the IRAP effects – typically quantified using the *D* scoring algorithm (Barnes-Holmes et al., 2010; Greenwald et al., 2003) – are likely to be poorly estimated. In a typical IRAP, *D* scores are calculated from only 36 reaction times. This small number of trials is also in contrast to the use of reaction time based tasks elsewhere in psychology. For example, the Implicit Association Test calculates scores from 120 reaction times (Greenwald et al., 1998), and the Stroop effect is frequently calculated from several hundred reaction times (Liefooghe et al., 2019). Given the high degree of variability and skew associated with reaction time data (Ratcliff, 1993; Whelan, 2008), this means that any given individual’s IRAP effect is likely to be poorly estimated.

Surprisingly, no research to date has quantified how well IRAP effects are estimated at the individual level. This study therefore calculates confidence intervals around individual participants’ IRAP effects. This is done using a large open dataset containing many different domains. These intervals are then used to estimate (1) the typical width of confidence intervals around IRAP *D* scores, (2) the proportion of *D* scores that can be inferred to differ from zero (i.e., where evidence of an IRAP effect was obtained), (3) the proportion of *D* scores that can be said to differ from one another (i.e., agnostic to the zero point), and (4) the proportion of the observed range of all *D* scores that is covered by an individual participant’s *D* score. Following a recent call to consider alternative scoring algorithms for IRAP data, the performance of the *D* score is then compared with the a more robust method: the Probabilistic Index (PI: De Schryver et al., 2018). Finally, the IRAP’s utility for individual assessment is compared to the most common popular measure of the implicit biases, the Implicit Association Test (IAT: Greenwald et al., 1998), again using a large open dataset of IATs assessing many different domains.

# Method

## Data

### IRAP data

Data were taken from a large open dataset of all IRAP research conducted by two labs and their coauthors (Hussey & Drake, 2020a). This dataset contains both trial level IRAP data, as well as all word and image stimuli, instructions, responding rules, and task parameters used in each of the IRAPs (osf.io/v3twe). Inclusion criteria used for the curation of that dataset were listed by Hussey and Drake (2020) to be as follows: (1) study used at least one IRAP task, excluding variants such as MT-IRAP (Levin et al., 2010), NL-IRAP (Kavanagh et al., 2016), or training-IRAP (Murphy et al., 2019); if a given study employed more than one IRAP, only data from the first IRAP each participant completed was used; and (3) trial-level reaction time data was available. The dataset contains both published and (previously) unpublished data (for prior publications see Drake et al., 2016, 2018; Finn et al., 2016; Hussey, Daly, et al., 2015). Data was included from both published and unpublished papers following Cochrane guidelines (Higgins & Thomas, 2022: section 4.3.2). Data came from 18 different substantive domains: body shape, Clinton-Trump, Christianity-Islam, suffering and development between countries, disgust, gender stereotypes, ideographic evaluations of friends and enemies, life and death, personality, race, religion, rich-poor, sexuality and arousal, shapes and colors, stigma, and valenced words. Some domains involved more than one stimulus set within the IRAP, for example there were two variants of the race IRAP.

**IAT data**

Inclusions criteria were self-reported fluent English, complete trial-level IAT data, …. Due to a combination of long bootstrapping calculation runtimes and a very large IAT dataset, a subset of 100 participants per domain after exclusions were randomly sampled from the full public dataset, to make 9500 participants total.

## Participants

The IRAP dataset included 1571 participants prior to exclusions. Individual participants were excluded on the basis of outlier reaction times (i.e., deviation of ± 2 median absolute deviations). 109 participants were excluded on this basis, leaving 1462 in the analytic sample. This sample size is therefore roughly three times larger than the total sample size studied in Vahey et al.’s (2015) meta-analysis of clinically relevant IRAP research (*N* = 494), and is roughly 40 time larger than the median published IRAP study. Where demographic data was available, the sample was 62.4% women, 37.3% male, and 0.2% identified using another label; *M*age = 20.1, *SD* = 4.3. Sample sizes for each domain ranged from *N* = 11 to 137, *M* = 44.9, *SD* = 30.1. The included IAT dataset included 9500 participants after exclusions and random sampling of 100 participants in each of the 95 domains. Where demographic data was available, the sample was 62.4% women, 37.3% male, and 0.2% identified using another label; *M*age = 20.1, *SD* = 4.3. All participants in both datasets provided informed consent and individual studies were approved by the local institutional review board.

## Measures

### Implicit Relational Assessment Procedure

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. On each block of trials, participants are presented 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 only 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). Differential reaction times between the two block types are used to quantify the IRAP effect.

### Implicit Association Test

Likewise, the IAT had been described in great detail elsewhere (Greenwald et al., 1998; Nosek et al., 2005), and so only a brief overview will be provided here. On each block of trials, participants are presented with images or words in the middle of the screen from four different categories (e.g., images of white face, images of black faces, positive words, and negative words). Two pairs of category labels for these stimuli are presented on the top left and top right hand sides of the screen (e.g., White people, Black people, Positive, and Negative) and are mapped to the left and right response keys so that each key shares both a target and a attribute category. Participants must categorize the stimuli in the middle of the screen using the two response options which are mapped to the four categories. In order to progress to the next trial, the correct response must typically be given. Incorrect responses result in a red X being presented on screen. Between blocks of trials, the mapping of the category labels and the required response changes so that, for example, on one block type participants White people and positive words share the left response key and Black people and negative words share the right response key. On the other block type the opposite is the case, for example White people and negative words share the left response key and Black people and positive words share the right response key. Participants are instructed to sort the stimuli as quickly and accurately as possible. Differential reaction times between the two block types are used to quantify the IAT effect.

## Scoring methods

IRAP and IAT 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; Greenwald et al., 2003; 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 block types. The difference between the means is then divided by the standard deviation, resulting in a *D* score. *D* scores have a maximum possible range of -2 to +2, with 0 representing the neutral point. In the IRAP literature, this zero point is often employed as a meaningful reference point from which comparisons are made, such differences from zero *t*-tests (Hussey, Daly, et al., 2015) or testing the proportion of scores above vs. below zero (Finn et al., 2019).

De Schryver et al. (2018) discuss some of the limitations of the *D* score, whose assumptions are typically violated by IRAP and IAT data, and argued that more robust scoring and interpretable scoring methods should be employed, specifically the Probabilistic Index (PI). This effect size has also been employed under several other names including the probability of superiority or Ruscio’s A (Ruscio, 2008). The PI can be interpreted as “the probability that a randomly selected inconsistent trial has a larger RT than a randomly selected consistent trial” (De Schryver et al., 2018, p.100). As a probability, PI scores have a maximum possible range of 0 to 1, with 0.50 representing the neutral point of no IRAP/IAT effect. PI scores can also be implemented using exactly this definition, as an exhaustive comparison of ordinal rank between block types. Computationally efficient R code to do this was supplied in Ruscio’s (2008) supplementary materials, which was used to calculate PI scores as well as the more typical *D* scores.

Note that the IRAP literature has historically described the neutral point of equal speed of responding between the two block types as the “zero point”, on the basis that the neutral point equals the zero point when using the *D* score. For the sake of computability with the existing IRAP literature, this article employs the term “zero point” for both *D* = 0 and PI = 0.50 (i.e., “zero point” refers to the point of zero bias rather than a score of zero).

## Bootstrapped 95% Confidence Intervals

Participants are typically described as demonstrating a positive effect if its value is descriptively above the neutral point (i.e., *D* = 0, PI = 0.50), and a negative effectif it is descriptively below zero. This can be a useful description of how to interpret the direction of an effect description of an effect (Hussey, Thompson, et al., 2015). However, when dealing with data from individual participants, this practice moves from mere description or interpretation to necessitating an inference method. That is, if we wish to consider whether a given participant demonstrated a positive IRAP effect on a given trial type, it is not sufficient that their score merely be descriptively greater than the neutral point (e.g., *D* > 0), rather it must be possible to show their score is greater than the neutral point via an inference test (e.g., the lower bound confidence interval of the *D* score > 0). Depending on the width of the confidence intervals, it may be the case that even descriptively large *D* scores do not allow us to infer a deflection from zero. In order to quantify the uncertainty around individual *D* scores and allow us to make inferences about individuals, I therefore calculated confidence intervals around individual scores (i.e., one *D* score for each of the four IRAP trial-types for each participant; a single *D* score for the IAT following standard practice). To the best of my knowledge, no published study using the IRAP or IAT has calculated or reported confidence intervals on individuals’ scores before now. In order to provide a comparison for the performance of the *D* score and following De Schryver et al.’s (2018) recommendations, PI scores and their 95% confidence intervals were also calculated.

A common method for calculating confidence intervals the arithmetic method (e.g., CI = mean ± SEM\*z, where z for 95% interval = 1.96; Swinscow & Campbell, 1997). However, this requires the standard error of the mean of the effect size, or a derivative of it such as its variance, to be specified. To the best of my knowledge, the SEM of the *D* score effect size has not yet been defined. This is possibly based on its somewhat odd properties such as finite range and correlation between numerator and denominator, in contrast to other forms of standardized mean difference effect sizes on which it was nominally based (see De Schryver et al., 2018; Greenwald et al., 2003).

An accessible alternative method for calculating confidence intervals is bootstrapping. Briefly, bootstrapping, or random sampling with replacement, is a resampling method that is often used as an alternative to mathematical statistical inference in cases where parametric assumptions might be violated or parameters are not trivial to calculate, such as with the *D* score. In this case, bootstrapping involved calculating scores using random samples from the data for each participant, with replacement, a large number of times. The resulting distribution of bootstrapped scores was then parameterized to obtain confidence intervals. For a book length introduction to bootstrapping see for example the classical text by Mooney et al. (1993). This was accomplished for both *D* and PI scores via bootstrapping using the R package boot (Canty, 2002) using 5000 resamples. The Bias Corrected and Accelerated (BCA) method was used to minimise bias relative to other bootstrapping methods (see Albright, 2019 for discussion and simulation study). Confidence intervals were bootstrapped, but the point estimate *D* and PI score were computed as normal. All data and R code to reproduce the analyses or reuse for other purposes are available on the Open Science Framework ([osf.io/mb4ph](https://osf.io/mb4ph)).

In summary, *D* scores were calculated following standard practice for each participant, at the trial type level for the IRAP (but not the IAT which does not separate trial types), and in each domain. For each score, 95% Confidence Intervals were also calculated. The same data used to calculate *D* scores and their corresponding confidence intervals was then used to also calculate PI scores and their confidence intervals.

**Figure 1.** Caterpillar plot of IRAP *D* scores and 95% Confidence Intervals calculated for each domain and trial type and displayed split by domain



# Results

## Utility of individual level IRAP *D* scores

Participants’ IRAP *D* scores and 95% Confidence Intervals are presented in Figure 1. These are clustered by domain, and arranged by ranking the participants by their *D* score. This type of plot is sometimes referred to as a caterpillar plot in the meta-analysis literature (e.g., Fernández-Castilla et al., 2020), and a similar form of plot – albeit without 95% CIs – has been used in a recent IRAP publication that analyzed data at the individual level (Finn et al., 2019, figure 3 p.433). Individual estimates and their intervals have been colored based on whether the interval excludes the zero point or not (*D* = 0). It is split by domain but not trial type on the basis that a further split by trial type would make each individual plot so small as to be uninterpretable. Note however that while the plot does not separate the trial types within each domain, the estimates and bootstraps were indeed calculated at the trial type level. The trial type level data is available in the supplementary materials.

### 95% Confidence Intervals widths

The distribution of confidence interval widths demonstrated very strong skew. As such, it was not appropriate to meta-analyze the widths or describe their distribution using means or even medians. Instead, I report the Maximum A Posteriori (MAP) estimate (Makowski et al., 2019), which represents the most probable value in a distribution of continuous values (i.e., is akin to the mode for continuous data). Across all domains and trial types, the most probable value (MAP) for the width of an individual’s *D* score’s 95% Confidence Interval was *D* = 1.31. Within domains and trial types, the smallest most probable value (MAP) was *D* = 0.75 and the largest was *D* = 1.35. Figure 1S in the supplementary materials illustrates the MAP 95% Confidence Interval widths by domain and trial type, and suggests that widths are fairly consistent, but with exceptions. However, these exceptions cannot be diagnosed as a pattern between trial types or the domains being assessed (e.g., both high and low MAP 95% CI widths for different race IRAPs and shapes-colors IRAPs). Luckily, the assessment of the utility of the IRAP data at the individual level does not rely on quantifying the width of confidence intervals directly, but instead on three properties of these intervals. The following sections discuss these each in turn.

**Figure 2.** Proportion of IRAP *D* and PI scores that exclude the zero point within each domain and trial type



### Proportion of non-zero scores

The color of the point estimates and intervals in Figure 1 are determined by whether the interval excludes the zero point. That is, they are colored by whether an IRAP effect was detectable or not. Analyses assessing deviation from the zero point have been used throughout the IRAP literature to date (e.g., Finn et al., 2019). On this basis, if the IRAP has utility at the individual level, a large proportion of participants’ scores on the IRAP should also be detectably different from the zero point (i.e., there should be detectable IRAP effects). As can be seen from the plot, based on their confidence intervals the vast majority of *D* scores are not significantly different from zero, and as such only a small minority of participants can be inferred to have demonstrated an IRAP effect (within a given domain and trial type). Figure 2 illustrates the proportion of IRAP *D* scores that excluded the zero point, split by trial type and domain, for both *D* and PI scores.

In order to quantify the proportion of individual participants with detectable biases, results were meta-analyzed across trial types, participants, and domains. This and all subsequent analyses were implemented as (meta-analytic) linear mixed effects models using the R packages lme4 (Bates et al., 2015) and emmeans (Lenth et al., 2022). The proportion of scores that differ from zero was calculated for each trial type and domain and used as the dependent variable. Because the dependent variable was a probability on a 0-1 scale, it was logit transformed prior to analysis and results were inverse logit transformed for reporting. This ensured that the model returned predictions within the theoretical limits of the dependent variable (i.e., probabilities from 0 to 1). The variance of each proportion was estimated via bootstrapping using the same method as the intervals on IRAP *D* scores. Following routine practice in meta-analysis, inverse variance was used as weights in the meta-analytic model (e.g., Viechtbauer, 2022). The model’s random effect was specified as trial types nested within domains, to reflect the nested nature of the way the data is generated by the IRAP (i.e., there are multiple domains, and within each domain there are four trial types). Finally, the scoring method (*D* vs PI scores) was entered as a fixed effect. Only the estimate for *D* scores is interpreted in this section; the comparison between *D* and PI scores within this model is returned to later on. Full results of this and all models can be found in the supplementary materials. This and all subsequent models return point estimates and 95% Confidence Intervals (CI), and also a 95% Prediction Interval (PI; using the nomenclature for this interval employed by the metafor R package: Viechtbauer, 2010). Whereas confidence intervals represent a long run probability of the true (i.e., data generating) value across levels of the random effect (i.e., across trial types and domains), prediction intervals instead represent the long run probability of the point estimates that are likely to be observed given the observed heterogeneity in the random effect, and are used in meta-analyses in order to quantify the impact of heterogeneity on results. These have utility within the current analyses given that there is heterogeneity in the proportion of participants that demonstrate non-zero IRAP effects between trial types and domains (see Figure 2). Prediction intervals allow us to quantify this variation among the observed data – even allow for generalizations to as-yet unobserved new conditions (e.g., new domains or stimulus sets).

Results of model 1 demonstrated that, across domains and trial types, the meta-analytic proportion of *D* scores that were found to differ from the zero point was =0.08, 95% CI [0.05, 0.12], 95% PI [0.01, 0.47]. Results of this and the next five meta-analytic models are depicted in Figure 5 (upper panel, *D* scores). To put the prediction interval in simple terms: across a wide variety of domains, some assessed via multiple different stimulus sets, and even between different trial types, 1–47% of *D* scores were found to be different from the zero point, typically 8%. As only a small proportion of individuals’ scores on the IRAP were detectably different from the zero point, this line of evidence suggests the IRAP does not have utility at the individual level.

**Figure 3.** Proportion of IRAP *D* and PI scores that can be discriminated from one another (i.e., lie outside of the 95% Discrimination Interval) within each domain and trial type



### Proportion of scores that differ from one another

The previous analysis treats the zero point as a meaningful reference point, on the basis that this is a common practice throughout the IRAP literature (e.g., Finn et al., 2019). However, some authors have argued that the zero point is not actually a neutral reference point for IRAP scores. This has been described various as a positivity bias (O’Shea et al., 2016), the Single Trial Type Dominance Effect (Finn et al., 2017), or the generic pattern among IRAP effects (Hussey & Drake, 2020b). Regardless of what label is used, there appears to be consensus that deviation from the zero point is often not exclusively due to the relation among the category and attribute stimuli within the task. A necessary implication of this is that the zero point (*D* = 0) cannot be inferred to represent no bias (agnostic to whatever that bias may represent for a given community of researchers, e.g., implicit attitudes vs. the impact of a history of relational responding). As such, in addition to the previous analysis that estimated the proportion of *D* scores that are different from the zero point, it is useful to also estimate the proportion of *D* scores that can be shown to differ from one another, which is agnostic to any one specific point. That is, rather than comparing a given individual’s *D* score’s 95% Confidence Interval against zero, we can compare it against all other participants’ *D* scores (within the same trial type and domain). We can refer to this as an assessment of the discriminability of an individual’s *D* score from other participants scores. If the IRAP has utility at the individual level, participants scores on the IRAP should be discriminable from other participants scores (i.e., there should be detectable variation between participants).

This was assessed using the lame logic as calculating individual PI scores, and indeed an adaption of the same code, by exhaustively assessing whether each individual’s *D* score’s interval excluded each other participant’s *D* score. In order to only compare like with like, these comparisons were made within domain and trial type. Where bootstrapping failed due to very small sample size (i.e., both proportion and variance estimated to be 0), these estimates were removed for plotting and meta-analysis.

Comparisons between an interval and a static reference point such as zero can be done using 95% Confidence Intervals (i.e., following common practices for Null Hypothesis Significance Testing). However, comparisons between two *D* scores must also take into account the uncertainty in the estimation of both scores rather than just one. Weir (2005) demonstrates that the appropriate method for this is to widen the 95% Confidence Interval by multiplying its width by . For the purpose of the current analyses, these can be referred to as 95% Discrimination Intervals (DIs). These were calculated for each *D* score from the *D* score’s confidence intervals. Each interval’s discriminability from all other D scores within the same domain and trial type was then assessed via exhaustive pairwise comparisons. The proportion of *D* scores that could be discriminated from other *D* scores was then estimated for each trial type and domain via bootstrapping (i.e., its point estimate and variance) using the same method as for the IRAP *D* scores. Figure 3 illustrates the proportion of other *D* and PI scores that a given individual’s score can be discriminated from for each trial type and domain. The plot suggests significant heterogeneity may be present between the trial types and domains.

The estimates were then subjected to a similar analysis as the previous one, with identical transformations, weightings, and both fixed and random effect specifications. Only the dependent variable was changed to the proportion of discriminable scores. Results of model 2 demonstrated that, across domains and trial types, the meta-analytic proportion of *D* scores that were found to be discriminable from one another was =0.06, 95% CI [0.04, 0.08], 95% PI [0.01, 0.25] (see Figure 5, middle panel, *D* scores). To again put the prediction interval in simple terms: across a wide variety of domains, some assessed via multiple different stimulus sets, and even between different trial types, only 1–25% of individuals’ *D* scores were found to be discriminable from the other individuals’ *D* scores within the same domain and trial type, typically 6%. As there was little detectable variation between individuals’ *D* scores, this is an additional line of evidence that suggests the IRAP does not have utility at the individual level.

**Figure 4.** Proportion of the observed range covered by individual participants’ *D* scores’ 95% Confidence Intervals within each domain and trial type.



### Proportion of observed range covered by individual scores

It is also useful and important to understand an individual *D* score’s confidence interval width in the context of the maximum range of scores. Loosely speaking, if a scores on a given depression scale told you that an individual lay in the range of 2 to 8, it is important to know whether the maximum range of the scale is 1 to 100, in which case the individual is on the low end, or 1 to 10, in which little can be said about the individual and the scale may have little utility. If the IRAP has utility at the individual level, a given participant’s estimated score on the IRAP should be cover a relatively small proportion of the total observed range of all participants’ intervals (i.e., individuals’ scores should exclude them from large proportions of the observed continuum).

Although *D* scores have a maximum theoretical range of -2 to 2, such extreme values are not typically observable. The observed range of *D* score’s 95% Confidence Intervals (i.e., from the lowest lower CI to the highest upper CI) across all participants, domains, and trial types was *D* = -1.51 to 1.71. In order to estimate the proportion of the observed range covered by a given participant’s interval, the width of the interval was divided by the observed range of intervals within each trial type and domain, and its mean and variance were then calculated (see Figure 4). These proportions were then subjected to a similar analysis as the previous two, with identical transformations, and both fixed and random effect specifications. The dependent variable was changed to the proportion of the observed range covered by individual 95% Confidence Intervals (within each trial type and domain). Results of model 3 demonstrated that, across domains and trial types, the meta-analytic proportion of the observed range covered by individual intervals was =0.51, 95% CI [0.50, 0.53], 95% PI [0.42, 0.61] (see Figure 5, lower panel, *D* scores). To again put the prediction interval in simple terms: across a wide variety of domains, some assessed via multiple different stimulus sets, and even between different trial types, individual intervals were found to cover 42–61% of the observed range of *D* score intervals within the same domain and trial type, typically 51%. As individuals’ intervals covered a large proportion of the total observed range, this third line of evidence suggests the IRAP does not have utility at the individual level.

**Figure 5.** Results of meta-analyses 1 to 3 assessing the individual level utility of IRAP *D* scores and comparing them with IRAP PI scores.



## Comparing individual level utility of IRAP *D* scores with the IRAP PI score

Each of the three previous meta-analytic models also included the data scored using the PI as well as the *D* score, and assessed differences between the two. Results are reported below for each. No direct comparison between the MAP of the width of the confidence intervals of the *D* versus PI was possible because they have different maximum possible ranges (i.e., are on different scales).

### Proportion of non-zero scores

In model 1, the meta-analytic proportion of PI scores that were found to differ from the zero point was =0.08, 95% CI [0.05, 0.12], 95% PI [0.01, 0.46] (see Figure 5, upper panel, PI scores). Scoring the IRAP with the PI score instead of the *D* score did not improve this proportion, = 0.00, *p* = .56.

### Proportion of scores that differ from one another

In model 2, the meta-analytic proportion of PI scores that were found to be discriminable from one another was =0.05, 95% CI [0.03, 0.06], 95% PI [0.01, 0.21] (see Figure 5, middle panel, PI scores). Scoring the IRAP with the PI score instead of the *D* score resulted in a worse proportion of discriminable scores, although the magnitude of change was small, = -0.01, *p* = .009.

### Proportion of observed range covered by individual scores

In model 3, the proportion of the observed range of PI scores covered by individuals 95% Confidence Intervals was =0.49, 95% CI [0.47, 0.51], 95% PI [0.40, 0.58] (see Figure 5, lower panel, PI scores). Scoring the IRAP with the PI score instead of the *D* score resulted in an improved, smaller proportion of coverage, although the magnitude of change was small, = -0.02, *p* < .001.

## Comparing individual level utility of IRAP *D* scores with IAT *D* scores

XXXX. note that both use *D* scores. Note changes to the random effect structure.

[discuss why estimates of IRAP proportions are different to previous models, and therefore why delta scores don’t necessarily correspond.] Analogous to Figure 1, IAT *D* scores and their confidence intervals for every domain can be found in Figure 2S in the supplementary materials. Analogous to Figure 1S, the MAP IAT *D* score 95% CI widths can be found in Figure 3S in the supplementary materials

### Proportion of non-zero scores

The proportion with of IAT *D* scores that exclude the zero point by domain can be found in Figure 4S in the supplementary materials. Model 4 was similar to model 1. The fixed effect comparison in model 4 was task rather than scoring method (i.e., IRAP *D* score vs. IAT *D* score rather than IRAP *D* score vs. IRAP PI score). Because the IAT has only one trial type, the random intercept only specified domain rather than domain and trial type. Because there was substantial differences in proportions between the tasks, this was allowed to vary in the random structure too by specifying a random slope for task. The meta-analytic proportion of IAT *D* scores that were found to differ from the zero point was =0.57, 95% CI [0.52, 0.62], 95% PI [0.06, 0.97] (see Figure 6, upper panel, IAT *D* scores). This was significantly better than for the IRAP *D* scores and the magnitude of the difference in proportions was large, = 0.51, *p* < .001.

### Proportion of scores that differ from one another

The proportion with of IAT *D* scores that were discriminable from other *D* scores in each domain can be found in Figure 5S in the supplementary materials. Model 5 was similar to model 2. The same modifications to the fixed and random effects were made as model 4. The meta-analytic proportion of IAT *D* scores that were found to be discriminable from one another was =0.44, 95% CI [0.39, 0.48], 95% PI [0.08, 0.87] (see Figure 6, middle panel, IAT *D* scores). This was significantly better than for the IRAP *D* scores and the magnitude of the difference in proportions was large, = 0.40, *p* < .001.

### Proportion of observed range covered by individual scores

The proportion with of the observed interval width covered by individual IAT *D* scores 95% Confidence Intervals in each domain can be found in Figure 6S in the supplementary materials. Model 6 was similar to model 3. The same modifications to the fixed and random effects were made as model 4. The proportion of the observed range of IAT *D* scores covered by individuals 95% Confidence Intervals was =0.26, 95% CI [0.26, 0.27], 95% PI [0.16, 0.41] (see Figure 6, lower panel, IAT *D* scores). This was significantly better than for the IRAP *D* scores and the magnitude of the difference in proportions was large, = 0.24, *p* < .001.

**Figure 6.** Results of meta-analyses 4 to 6 comparing the individual level utility of IRAP *D* scores with IAT *D* scores.

****

# Discussion

Results from four different metrics suggest that the IRAP does not have utility at the individual level: (1) A given IRAP *D* score’s confidence intervals are likely to be very wide (i.e., ±0.66); (2) with the result that only a small minority of *D* scores actually represent evidence of IRAP effects (19%) (3) or are significantly different from other *D* scores (30%), and (4) XXX.

Except in the case of extreme scores, an individual *D* score is in general so poorly estimated as to allow for almost no inferences about the individual. This point can be illustrated with a simple example: if a participant completed an IRAP and demonstrated a *D* score = 0.30, we might traditionally describe this as a positive IRAP effect. However, when the confidence intervals around *D* scores are considered, we would more accurately say that the participant’s true score lies somewhere in the range of very negative (*D* = -0.41) to very large (*D* = 0.91) – bearing in mind that 95% of all observed D scores fell within the range of *D* = -0.66 to 0.93. As such, individual *D* scores are very poorly estimated, and are consistent with such a wide range of conclusions that few inferences can be made from an individual’s data.

In the great majority of cases, individuals therefore cannot be said to demonstrate an IRAP effect or, by implication, any theoretical abstraction from such an effect regardless of ones theoretical framework, such as an implicit bias or attitude, or a brief and immediate relational response (Hughes et al., 2011). These biases can be more reliably detected at the group level due to the benefit of aggregation (e.g., increased statistical power, see Connor & Evers, 2020).

Results also suggest that using an alternative scoring method, the PI, does not improve the IRAP’s individual utility.

In order to place the IRAP’s individual level performance in the context of a comparable task, it was compared to the most popular implicit measure, the Implicit Association Test. Across all metrics, the IAT significantly and substantially outperformed the IRAP.

The IRAP was recently argued to be better suited to investigating behavior analytic questions than questions around implicit attitudes (REF). It is important to note that the analysis of individual level IRAP data can serve neither purpose if responses on the IRAP are very poorly estimated, regardless of whether you couch this in psychometric language (e.g., poor individual level estimation) or behavior analytic language (e.g., poor stimulus control within the task). For example, Finn et al. (2019) adopted an explicitly behavior analytic approach, consistent with the recent recommendations of Barnes-Holmes (REF) for a behavior analytic program of IRAP research. However, their analyses of individual level IRAP data (including presentation of similar caterpillar plots as used in this article) are equally stymied by the IRAP’s poor estimation of individual level effects as would, for example, explicitly social-cognitive research using the IRAP as a measure of implicit attitudes.

Notionally, the estimation of individual scores could be improved. This could take many forms, including greatly lengthening the procedure, for example by a factor of four or so. Given that reliability is determined in part by task length, this would also help raise the IRAP’s internal consistency and test-retest reliability, which are currently between poor and unacceptably low (Hussey & Drake, 2020a). However, task lengthening is often less effective in practice due to increased participant fatigue. Additionally, this may make the task unreasonably long for each participant (e.g., >45 minutes). As such, other forms of improvement to the task itself in conjunction with lengthening the task itself are likely necessary. From a behavioral perspective, it appears that there is a need to enhance stimulus control over behaviour within the task in order to improve signal-to-noise ratio (in the psychometric sense) of the data it produces.

# References

Albright, J. (2019). *Boot Package in R: Understanding Bootstrap Confidence Interval Output*. https://blog.methodsconsultants.com/posts/understanding-bootstrap-confidence-interval-output-from-the-r-boot-package/

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., & 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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Canty, A. J. (2002). Resampling Methods in R: The boot Package. *R News*, *2/3*, 2–7.

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

Drake, C. E., Primeaux, S., & Thomas, J. (2018). Comparing Implicit Gender Stereotypes Between Women and Men with the Implicit Relational Assessment Procedure. *Gender Issues*, *35*(1), 3–20. https://doi.org/10.1007/s12147-017-9189-6

Drake, C. E., Seymour, K. H., & Habib, R. (2016). Testing the IRAP: Exploring the Reliability and Fakability of an Idiographic Approach to Interpersonal Attitudes. *The Psychological Record*, *66*(1), 153–163. https://doi.org/10.1007/s40732-015-0160-1

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, N., Onghena, P., & Noortgate, W. V. den. (2020). Visual Representations of Meta-Analyses of Multiple Outcomes: Extensions to Forest Plots, Funnel Plots, and Caterpillar Plots. *Methodology*, *16*(4), 299–315. https://doi.org/10.5964/meth.4013

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*, *66*(2), 309–321. https://doi.org/10.1007/s40732-016-0173-4

Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2017). Exploring the Single-Trial-Type-Dominance-Effect in the IRAP: Developing a Differential Arbitrarily Applicable Relational Responding Effects (DAARRE) Model. *The Psychological Record*, 1–15. https://doi.org/10.1007/s40732-017-0262-z

Finn, M., Barnes-Holmes, D., McEnteggart, C., & Kavanagh, D. (2019). Predicting and Influencing the Single-Trial-Type-Dominance-Effect: The First Study. *The Psychological Record*, *69*(3), 425–435. https://doi.org/10.1007/s40732-019-00347-4

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

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

Higgins, J., & Thomas, J. (2022). *Cochrane Handbook for Systematic Reviews of Interventions* (Version 6.3). The Cochrane Collaboration. Available from www.handbook.cochrane.org

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, *61*(3), 465–498.

Hussey, I. (2020). *Bootstrapped Confidence Intervals around IAT D scores*. https://osf.io/t6c74

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., Daly, T., & Barnes-Holmes, D. (2015). Life is Good, But Death Ain’t Bad Either: Counter-Intuitive Implicit Biases to Death in a Normative Population. *The Psychological Record*, *65*(4), 731–742. https://doi.org/10.1007/s40732-015-0142-3

Hussey, I., & Drake, C. E. (2020a). The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis. *Preprint*. https://doi.org/10.31234/osf.io/ge3k7

Hussey, I., & Drake, C. E. (2020b). *The Implicit Relational Assessment Procedure is not very sensitive to the attitudes and learning histories it is used to assess*. PsyArXiv. https://doi.org/10.31234/osf.io/sp6jx

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

Kavanagh, D., Hussey, I., McEnteggart, C., Barnes-Holmes, Y., & Barnes-Holmes, D. (2016). Using the IRAP to explore natural language statements. *Journal of Contextual Behavioral Science*, *5*(4), 247–251. https://doi.org/10.1016/j.jcbs.2016.10.001

Klein, C. (2020). *Confidence Intervals on Implicit Association Test Scores Are Really Rather Large* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/5djkh

Lenth, R. V., Buerkner, P., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.8.0). https://CRAN.R-project.org/package=emmeans

Levin, M. E., Hayes, S. C., & Waltz, T. (2010). Creating an implicit measure of cognition more suited to applied research: A test of the Mixed Trial—Implicit Relational Assessment Procedure (MT-IRAP). *International Journal of Behavioral Consultation and Therapy*, *6*(3), 245–262. psyh. https://doi.org/10.1037/h0100911

Liefooghe, B., Hughes, S., Schmidt, J., & De Houwer, J. (2019). Stroop-like effects of derived stimulus-stimulus relations. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(2), 327–349. https://doi.org/10.1037/xlm0000724

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Mooney, C. Z., Mooney, C. F., Mooney, C. L., Duval, R. D., & Duvall, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. sage.

Murphy, C., Lyons, K., Kelly, M., Barnes-Holmes, Y., & Barnes-Holmes, D. (2019). Using the Teacher IRAP (T-IRAP) interactive computerized programme to teach complex flexible relational responding with children with diagnosed autism spectrum disorder. *Behavior Analysis in Practice*, *12*(1), 52–65. https://doi.org/10.1007/s40617-018-00302-9

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality & Social Psychology Bulletin*, *31*(2), 166–180. https://doi.org/10.1177/0146167204271418

O’Shea, B., Watson, D. G., & Brown, G. D. A. (2016). Measuring implicit attitudes: A positive framing bias flaw in the Implicit Relational Assessment Procedure (IRAP). *Psychological Assessment*, *28*(2), 158–170. https://doi.org/10.1037/pas0000172

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(4), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

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

Swinscow, T. D. V., & Campbell, M. J. (1997). *Statistics at Square One* (9th Ed.). BMJ Publishing Group. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one

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

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

Viechtbauer, W. (2022). *A Comparison of the rma() and the lm(), lme(), and lmer() Functions*. https://www.metafor-project.org/doku.php/tips:rma\_vs\_lm\_lme\_lmer

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231–240. https://doi.org/10.1519/15184.1

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482.