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

Ian Hussey & Chad E. Drake

Evidence for the IRAP's 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 Here, we extend this evidence base through meta-analyses of all published and unpublished studies conducted in two labs. Individual participant data was used to estimate both internal consistency and test-retest reliability across a large number of domains ($k = 16$) and participants ($N = 1576$). Results suggest that internal consistency is poor ($\alpha = .51$, 95% CI [.46, .56]) and test-retest reliability is very poor (ICC[2,1] = .20, 95% CI [.05, .34]). 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 approaching fifteen years old and having been used in over 100 published articles, 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 (Flake & Fried, 2019). 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). Quantification of the IRAP's measurement properties is therefore important to interpreting the results of existing research, and its utility in future work.

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 $r_{SB} = .65$, 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). 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 1207 participants for the meta-

analysis of internal consistency and 124 participants assessing test-retest reliability. Using Greenwald & Lai's (2020) data and code, computationally reproduced estimates were calculated for internal consistency (Cronbach's α = .56, 95% CI [.46, .65], 95% CR [.03, .85]) and for test-retest reliability (r = .45, 95% CI [.33, .55]).

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 typically accepted cut-offs 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.

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., ICC[2,1]: 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

All code and data needed to reproduce our analyses is available on the Open Science Framework, along with all word and image stimuli, instructions, responding rules, and task parameters used in each of the IRAPs (osf.io/v3twe).

Data was pooled from all IRAP studies we have been involved in. Inclusion criteria were use of at least one IRAP and access to raw data and the task parameters used in the study. Exclusion criteria were embargos on data that are soon to be published, whose data could therefore not be made open for this meta-analysis. Three studies met exclusion criteria. Two of these were in domains that are already represented in the included data (i.e., friend-enemy and Lincoln-Hitler).

Data from 35 IRAPs across 16 different content domains (see Figure 2) included a total of 1576 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 2). Some of this data has been published for other purposes (Drake et al., 2015, 2016, 2018; Finn et al., 2016; Hussey, Daly, et al., 2015; see supplementary materials). However, the large majority of this data was not considered by either of the two published meta-analyses of the IRAP's reliability, with the exception of a subset of the friend-enemy IRAPs (Drake et al., 2016) which was used in Greenwald and Lai (2020).

Participants

All participants provided informed consent prior to participation, and studies were approved by the local institutional review boards. Where demographics data was available, a majority of participants were women (64.4% female, 35.4% male, 0.2% non-binary), young adults (M_{age} = 19.5, SD = 3.3), 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

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 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 three pairs of test blocks until they meet performance criteria (e.g., median reaction time $<$ 2000 ms and percentage accuracy $>$ 80%), 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, Thompson, et al., 2015). All word and image stimuli, instructions, responding rules, and task parameters for each IRAP can be found in the Supplementary Materials (osf.io/v3twe).

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, Thompson, 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 also excluded if their mean reaction times in the IRAP test blocks were ± 2 Median Absolute Deviations from the median, in order to exclude implausibly fast or inappropriately slow responding. A total of 112 participants (7.1%) were excluded on this basis leaving 1464 participants in the internal consistency sample and 354 in the test-retest sample.

Results

Meta-analytic strategy

All data processing and analyses were done in R (R Core Team, 2020). Intraclass Correlation Coefficients were calculated using the psych package (Revelle, 2016). Meta-analyses were conducted using the metafor package (Viechtbauer, 2010, version 2.4-0) and Restricted Maximum Likelihood (REML) estimation. Meta-analysis of internal consistency estimates involved Bartlett transformations prior to analysis and inverse Bartlett transformations of meta-estimates 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., $r_{SB} = \frac{2r}{1+r}$). 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: r_{SB} = .54, 95% CI [.49, .59], 95% CR [.49, .59], I^2 = 0.1%, H^2 = 1.0.

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: r_{SB} = .52, 95% CI [.47, .57], 95% CR [.47, .57], I^2 = 0.0%, H^2 = 1.0. 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 (e.g., 2000). In each permutation, the data is split into two randomly determined halves, D scores are calculated for each, Pearson's r correlations are calculated from these two sets of D scores, and then a Spearman-Brown correlation is applied. A distribution of estimates is therefore obtained across permutations. This distribution is then parameterized: the mean value is used as the estimate, and the quantile method is used to find 95% Confidence Intervals. Parsons et al. (2019) noted that this method approximates Cronbach's α, and remove assumptions associated with specific split strategies (e.g., regarding learning occurring with the task between the first vs. second half).

Using the permutation method, the meta-analytic estimate of internal consistency was found to be poor, α = .54, 95% CI [.48, .59], 95% CR [.42, .62]. A small degree of heterogeneity was found between estimates, $Q(df = 34) = 44.82$, $p = .101$, $\tau^2 = 0.01$, $I^2 = 7.9\%$, $H^2 = 1.1$. A Graphical analysis of Study Heterogeneity plot (GOSH: Olkin et al., 2012) was used to attempt to understand this heterogeneity by assessing whether the meta-estimate was unduly influenced by outliers. This analysis uses resampling to calculate the distribution of estimates of effect size and heterogeneity (i.e., I^2) using a large number of subsets (10000) of possible combinations of the effect sizes. As illustrated in Figure 1, results indicated bimodality in both estimates of effect size and heterogeneity that was driven by data from a single domain: sexuality (Sexuality IRAP 1: α = .84, 95% CI [.65, .93], Sexuality IRAP 2: α = .93, 95% CI [.82, .97]), suggesting that it represented an outlier that biased the results. When this effect size was excluded as an outlier, the meta-analytic estimate of internal consistency was found to be poor, α = .51, 95% CI [.46, .56], 95% CR [.46, .56], with no heterogeneity, $Q(df = 32) = 21.59$, $p = .918$, $\tau^2 = 0.00$, $I^2 = 0.0\%$, $H^2 = 1.0$. See Figure 2 (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 conclusions are therefore based on this estimate.

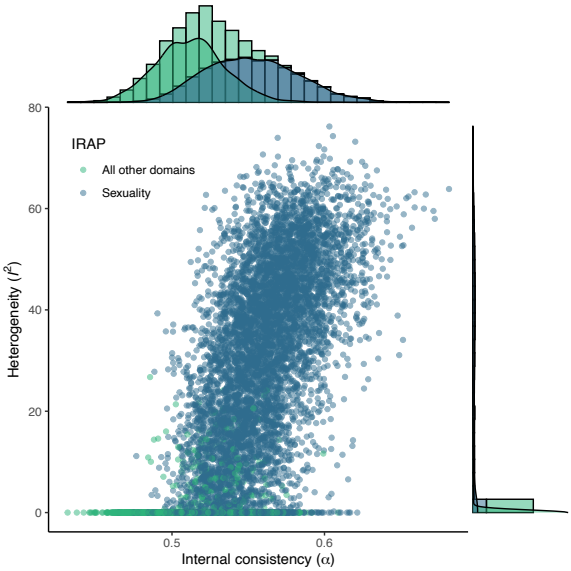

Figure 1. GOSH plot for internal consistency.

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, $r = .14$, 95% CI [-.07, .35], 95% CR [-.39, .60], $I^2 = 72.8\%$, $H^2 = 3.7$. Test-retest correlations were negative for three IRAPs (i.e., gender, body image, and race). This result may be most useful

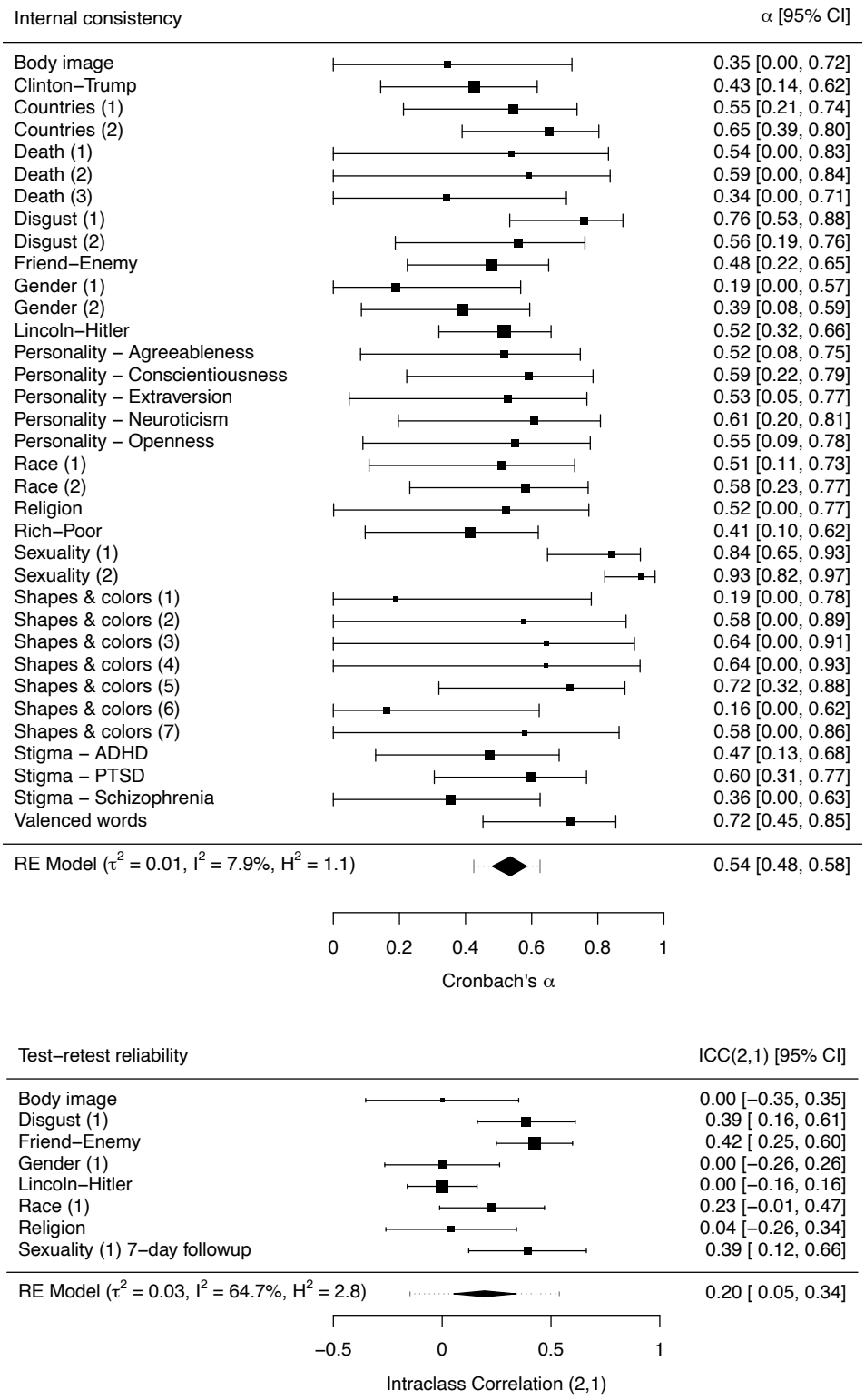

Figure 2. Forest plots.

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.

Test-retest via ICC(2,1). When using ICCs, results also suggested that test-retest reliability was very poor, ICC(2,1) = .20, 95% CI [.05, .34], 95% CR [-.15, .49]. A substantial degree of heterogeneity was found between the two studies, $Q(df = 7) = 21.4$, $p = .003$, $\tau^2 = 0.03$, $I^2 = 64.8\%$, $H^2 = 2.8$. Test-retest was near zero for half of the IRAPs (i.e., gender, body image, race, and Lincoln-Hitler).

A GOSH plot revealed no evidence of multimodality and therefore no evidence of outliers (see Figure 3). 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. Results can be found in Figure 2 (lower panel). 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. 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).

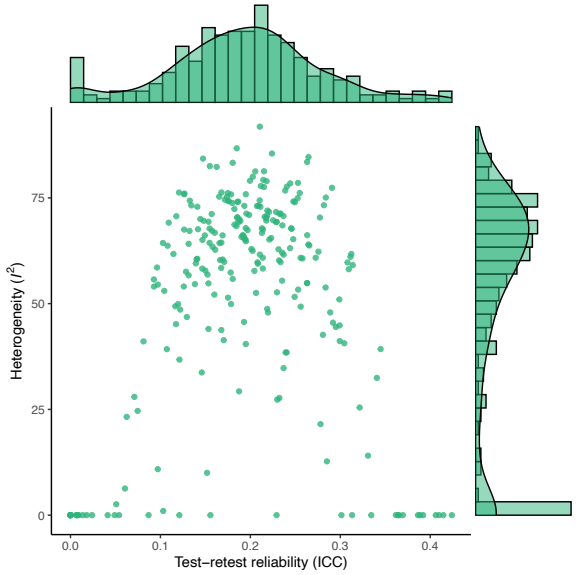

Figure 3. GOSH plot for test-retest reliability.

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 maximum observed estimate of the true correlation among two measures x and y (i.e., r_{xy}) is a function of the true correlation (ρ_{xy}) and also the reliability of both measures (i.e. their self-correlation ρ_{xx} and ρ_{yy}):

$$r_{xy} = \frac{\rho_{xy}}{\sqrt{\rho_{xx}\rho_{yy}}}$$

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 with a reliability of 1.0), we can use our meta-analyzed estimates of the IRAP's reliability to estimate the maximum correlations that could be observed between the two. 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 (ICC = .20) and internal consistency (α = .51). Maximum correlations with the IRAP (i.e., where true correlation $\rho_{xy} = 1$) was estimated to be r = .45 and .72, respectively. 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, a medium true correlation (ρ_{xy} = .5) would imply maximum observable correlations of r = .22 and .36, 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 'medium' 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.

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.

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 ρ^* refers to the goal reliability, ρ refers to the current reliability, and n refers to the multiple of current test length:

$$n = \frac{\rho^*(1-\rho)}{\rho(1-\rho^*)}$$

Using the meta-analytic estimate of the IRAP's internal consistency (α = .51), in order to increase internal consistency to α = .70, the task would need to contain 2.2 times the number of trials it currently does. Using the meta-analytic estimate of test-retest reliability (ICC = .20), in order to increase internal consistency to ICC = .70, the task would need to contain 9.3 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 22 minutes and 2.5 hours to complete, depending on the type and level 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. 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: $\alpha = .53$, 95% CI [.46, .58], A scores: $\alpha = .55$, 95% CI [.48, .61], $Q_M(df = 1) = 0.50$, $p = .478$.

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 were observed in internal consistency between the block orders; consistent block first: $\alpha = .46$, 95% [.33, .56], inconsistent block first: $\alpha = .48$, 95% CI [.29, .62], Q_M $(df = 1) = 0.06$, $p = .810$.

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. Results demonstrated that internal consistency was found to be higher when response option locations were static, static: $\alpha = .61$, 95% [.52, .69], moving: $\alpha = .48$, 95% CI [.33, .60], Q_M $(df = 1) = 5.37$, $p = .021$.

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 near-zero. This work has several benefits compared to previous meta-analyses: (a) it is the largest analysis to date, (b) it is resistant to publication bias, as it is based on our complete file drawer data, (c) it used more optimal analytic methods, and (d) it is computationally reproducible due to sharing both data and code.

Our estimate of internal consistency ($\alpha = .51$, 95% CI [.46, .56]) was smaller than that those reported in both previously published meta-analyses ($r_{SB} = .65$, 95% CI [.54, .74]: Golijani-Moghaddam et al., 2013; $\alpha = .56$, 95% CI [.46, .65]: Greenwald & Lai, 2020). Our estimate of test-retest reliability (ICC = .20, 95% CI [.05, .34]) was much lower than those reported by either previously published meta-analysis ($r = .49$, 95% CI [.10, .75]: Golijani-Moghaddam et al., 2013; $r = .45$, 95% CI [.33, .55]: Greenwald & Lai, 2020). Differences in results may be due to one or more features of our work relative to previous research, such as our larger sample size and variety of domains, the resilience of whole-lab file-drawer meta-analyses to publication bias,

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). While our results differ from previous meta-analyses to some degree, the conclusions of all agree that the IRAP's internal consistency and test-retest reliability is poor at best.

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 two and a half hours long for it to provide high test-retest reliability. This is likely to be at odds with the goals and pragmatics of many forms of research. We also used moderator meta-analyses to explore whether three factors might increase internal consistency. First, 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. We also assessed whether two commonly manipulated procedural parameters might increase internal consistency: the order in which participants completed the blocks, and whether response option mappings were static or moving. No differences were found between block orders, but improvements were found between moving ($\alpha = .48$) and static ($\alpha = .61$) response option locations. However, even when response option locations were static, internal consistency remained to be lower than the most popular implicit measure, the IAT ($\alpha = .80$: Greenwald & Lai, 2020), as well as being lower than the typically recommended minimum cut-off values for psychological measures (e.g., $\alpha > .7$, .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 animal-behaviorist working with rats in Skinner boxes must be concerned with whether the lever functions well as a measure of the animal's lever-pressing behavior: if it is too heavy or 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.

Vahey et al.'s (2015) meta-analysis of criterion validity concluded that the IRAP 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). 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.

References

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

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. *Preprint*. https://doi.org/10.1101/2020.04.26.048306

Drake, C. E., Kramer, S., Sain, T., Swiatek, R., Kohn, K., & Murphy, M. (2015). Exploring the reliability and convergent validity of implicit racial evaluations. *Behavior and Social Issues*, *24*. https://doi.org/10.5210/bsi.v24i0.5496

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

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*, 1–13.

Flake, J. K., & Fried, E. I. (2019). *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Preprint.* https://doi.org/10.31234/osf.io/hs7wm

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

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

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., & 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., Mhaoileoin, D. N., 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

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.

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH - a graphical display of study heterogeneity. *Research Synthesis Methods*, *3*(3), 214–223. https://doi.org/10.1002/jrsm.1053

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

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

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0) [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. (2016). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. http://CRAN.R-project.org/package=psych

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

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