The IRAP is not suitable for individual use due to very wide confidence intervals around *D* scores

Ian Hussey

A meta-analysis suggested that the Implicit Relational Assessment Procedure (IRAP) has potential "as a tool for clinical assessment". Here I present evidence to the contrary. Using all published and unpublished file-drawer data available to me, I bootstrapping 95% Confidence Intervals for each IRAP *D* score. Results demonstrate that Confidence Intervals are extremely wide: regardless of the estimated *D* score, the data is equally compatible with a 'true' score lying anywhere in the range of very negative to very positive. The IRAP is therefore not currently suitable for individual level use or assessment in research or applied settings.

The Implicit Relational Assessment Procedure (IRAP) is a reaction-time based task that has seen significant use as both a measure of implicit attitudes (Dawson et al., 2009) stand within Contextual Behavioral Science (Perez et al., 2019). In their meta-analysis of clinically relevant IRAP studies, Vahey et al. (2015) argued that the IRAP has potential "as a tool for clinical assessment" (p.64). However, for the IRAP to have individual-level utility, for clinical use or otherwise, scores produced by the task would need to be well estimated and come with a low degree of uncertainty.

Unfortunately, there is good *a priori* reason to believe that the IRAP's scores – 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 pairs of reaction times. This is 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 mean that any given individual's IRAP effect is likely to be poorly estimated. However, specifically how well has not yet been quantified. This study therefore uses existing data from a large number of published and unpublished studies to estimate (1) the width of confidence intervals around IRAP *D* scores, (2) the proportion of *D* scores that can accurately be said to differ from zero (i.e., where evidence of an IRAP effect was obtained), and (3) the proportion of *D* scores that can be said to differ from one another (i.e., agnostic to the zero point).

Method

Data

This study used all published and unpublished (file-drawer) data from my own work and that conducted in Prof Chad Drake's research group. Inclusion criteria were 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. Data came from 13 different substantive domains: body shape, Christianity-Islam, suffering and development between countries, disgust, hunger, ideographic evaluations of friends and enemies, gender stereotypes, life and death, race, sexuality and arousal, valenced words, and shapes and colors. Some domains involved more than one IRAP, for example there were two variants of race IRAPs. Some of this data has been used for different purposes in published articles elsewhere (Drake et al., 2016, 2018; Finn et al., 2016; Hussey, Daly, et al., 2015).

Participants

Data was obtained for 889 participants. This sample size is therefore significantly larger than the total sample size studied in Vahey et al.'s (2015) meta-analysis of clinically relevant IRAP research, and is roughly 20 time larger than the modal published IRAP study. Where demographic data was available, the sample was 63.2% women, 36.6% male, and 0.2% identified using another label; $M_{age} = 30.0$, $SD = 5.7$. Sample sizes for each IRAP ranged from $N = 10$ to 100, $M = 35.6$, $SD = 24.2$. All participants provided informed consent and studies were approved by the local institutional review board.

Measures

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

Results

Bootstrapped confidence intervals on IRAP *D* scores

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

Participants are typically described as demonstrating a positive D score if its value is descriptively above zero, and a negative D score if it is descriptively below zero. However, it is worth noting that this breaks with the traditions of hypothesis testing that are well established within psychology: A D score should only be described as positive if we can demonstrate that it is significantly different from zero. Depending on the width of the confidence intervals, it may be the case that even descriptively "large" D scores are not in fact credibly different from zero.

In order to quantify the uncertainty around individual D scores, I therefore bootstrapped confidence intervals. To the best of my knowledge, no published IRAP research has calculated or reported confidence intervals on individual's D scores before now. This was accomplished using the R packages purrr (Henry et al., 2020) and rsample (Kuhn et al., 2020), using the percentile method and 2000 resamples. All R code to reproduce the analyses is available on the Open Science Framework (osf.io/mb4ph).

Percentage of significant *D* scores

Results showed that only 19.0% of D scores can be said to be significantly different from zero. As such, in the great majority of cases, individuals cannot be said to demonstrate an IRAP effect (or, by implication, a theoretical abstraction from this such as an implicit bias or a brief and immediate relational response, see Hughes et al., 2011). This result is illustrated in Figures 1 and 2. Although there was some variation between IRAPs, results demonstrated that significant IRAP effects are consistently found only in a minority of cases, range = 11.2% to 46.4%, $M = 20.4\%$, $SD = 7.8\%$.

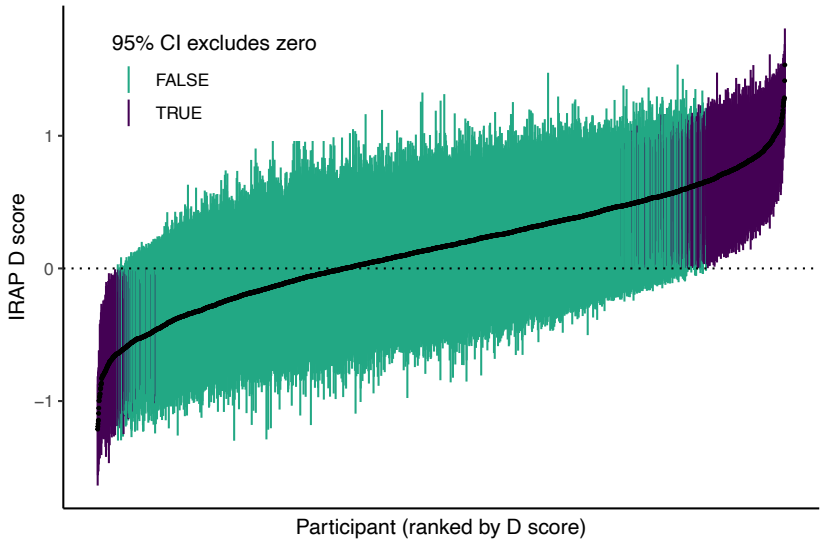

Figure 1. Bootstrapped confidence intervals around *D* scores. Participants are arranged by D score magnitude. Black points represent *D* scores, vertical lines represent 95% CIs. CIs that exclude zero are purple and those that do not are green.

Percent of *D* scores that differ from one another

The estimation precision of the IRAP *D* score was also assessed by examining what proportion of randomly selected *D* scores were significantly different from other randomly selected *D* scores. This comparison is useful as, unlike the above analysis, it is agnostic to whether the *D* score's zero point is meaningful (a point originally raised by Blanton & Jaccard, 2006, which perennially raised in the discussion of how IRAP effects should be interpreted). That is to say, perhaps zero is an arbitrary point to compare each *D* score against, and if so perhaps it is more useful to assess what proportion of *D* scores are different from one-another rather than different from zero. All *D* scores and their confidence intervals were compared against all others via pairwise comparisons. Confidence intervals on the proportion of cases that were different were again calculated via bootstrapping. Results demonstrated that any two randomly selected *D* scores were significantly different from one another in 30.3%, 95% CI [29.1, 31.4] of cases.

Confidence interval width

This result was driven by the fact that Confidence Intervals around any given D score were very wide. Given strong skew in their distribution, it was not appropriate to describe the distribution of confidence intervals around *D* scores using means or even medians. Instead, I employ the Maximum A Posteriori 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). Results showed that the most probable value for the width of a *D* score's confidence interval was MAP = 1.33. Results demonstrated that confidence intervals were of comparable width between IRAPs, range MAP = 1.00 to 1.35, $M = 1.29$, $SD = 0.08\%$.

It is also important to understand this confidence interval width in the context of the possible range. Although *D* scores have a maximum mathematical range of -2 to 2, such extreme values are not possible with the constraints of the IRAP itself. It is therefore more useful to observe that the 95% trimmed range of all observed *D* scores in the sample (i.e., the range of non-outlier *D* scores) was $D = -0.66$ to 0.93, or a total range of 1.59. That is, among the sample of 889 participants, 95% of *D* scores fell within this range. As such, the most probable confidence interval width across the sample (MAP = 1.33) spans 83.8% of the observed range of *D* scores.

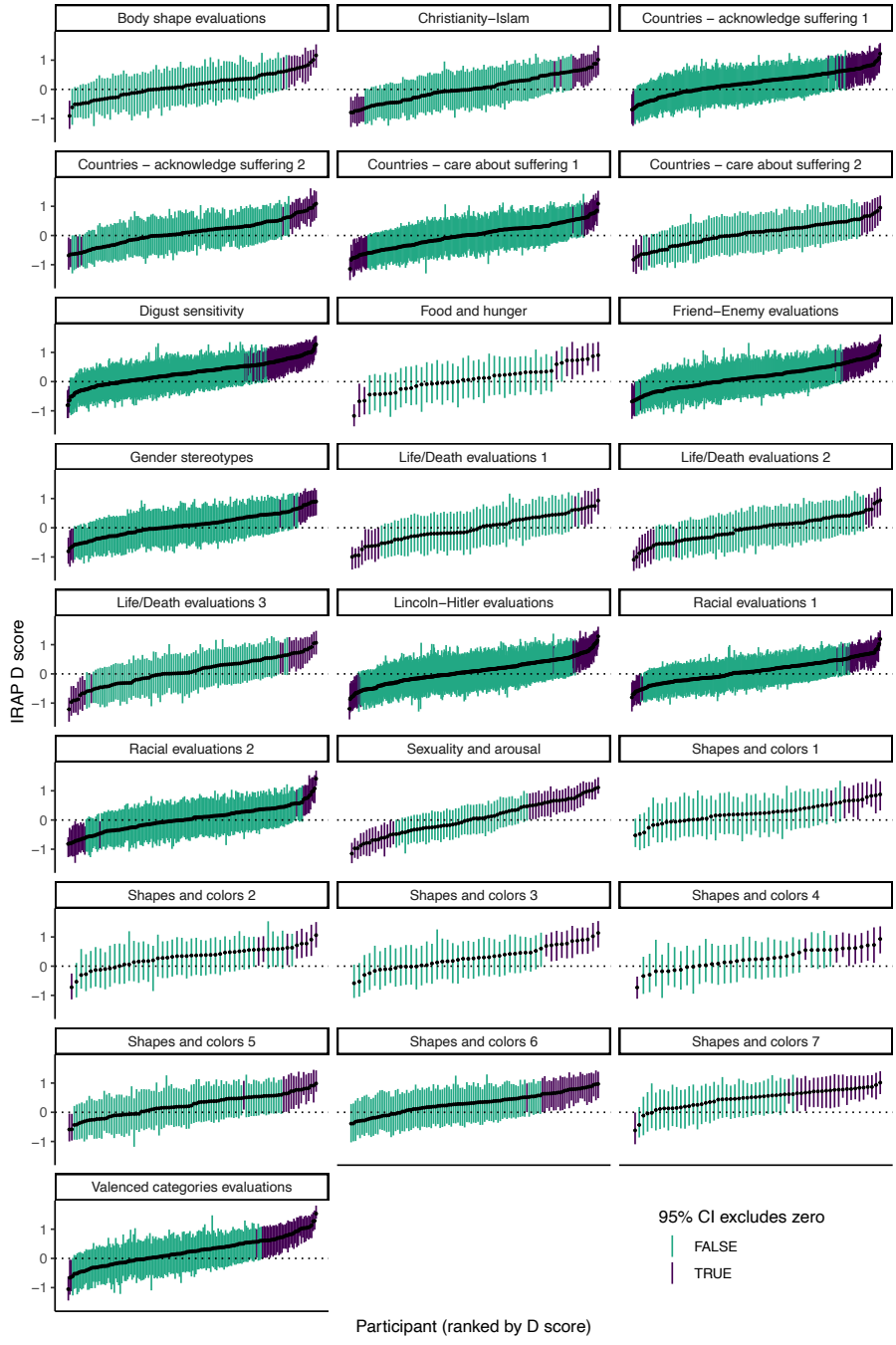

Figure 2. Bootstrapped confidence intervals around D scores, split by IRAP.

Discussion

Results provide convergent evidence under a range of different assumptions that IRAP D scores are very poorly estimated. A given D score's confidence intervals are likely to be very wide (i.e., ± 0.66); with the result that only a small minority of D scores actually represent evidence of IRAP effects (19%) or are significantly different from other D scores (30%). 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 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. As such, the IRAP, as in its current form, does not have individual (clinical) utility in research or applied settings (cf. Vahey et al., 2015).

It is also worth noting that similar analyses of data from another implicit measure, the Implicit Association Test, suggests that the IRAP's estimation precision is substantially worse than the IAT's (IRAP CI width MAP = 1.32, IAT CI width MAP = 0.75: see Hussey, 2020; Klein, 2020).

Notionally, the estimation of individual scores could be improved. This could take many forms, including use of more robust scoring algorithms (De Schryver et al., 2018), improvements to the task format itself to improve the signal to noise ratio (in a psychometric sense) and enhance stimulus control over behaviour within the task (in a behavioural sense). Estimation could also be improved by greatly lengthening the procedure by a factor of four or so, however, this may make the task unreasonably long for each participant (e.g., >45 minutes).

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.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*(1), 27–41. https://doi.org/10.1037/0003-066X.61.1.27

Dawson, D. L., Barnes-Holmes, D., Gresswell, D. M., Hart, A. J., & Gore, N. J. (2009). Assessing the Implicit Beliefs of Sexual Offenders Using the Implicit Relational Assessment Procedure A First Study. *Sexual Abuse: Journal of Research and Treatment*, *21*(1), 57–75. https://doi.org/10.1177/1079063208326928

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

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

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

Henry, L., Wickham, H., & RStudio. (2020). *purrr: Functional Programming Tools* (Version 0.3.4) [Computer software]. https://CRAN.R-project.org/package=purrr

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., 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., 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

Kuhn, M., Chow, F., Wickham, H., & RStudio. (2020). *rsample: General Resampling Infrastructure*

(Version 0.0.7) [Computer software]. https://CRAN.R-project.org/package=rsample

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

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

Perez, W. F., de Almeida, J. H., de Rose, J. C., Dorigon, A. H., de Vasconcellos, E. L., da Silva, M. A., Lima, N. D. P., de Almeida, R. B. M., Montan, R. N. M., & Barnes-Holmes, D. (2019). Implicit and Explicit Measures of Transformation of Function from Facial Expressions of Fear and of Happiness via Equivalence Relations. *The Psychological Record*, *69*(1), 13–24. https://doi.org/10.1007/s40732-018-0304-1

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

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

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