Journal of Contextual Behavioral Science

The Implicit Relational Assessment Procedure is not suitable for individual use

--Manuscript Draft--

Manuscript Number:	
Article Type:	Empirical Research
Keywords:	implicit relational assessment procedure; implicit attitudes; clinical assessment; meta-analysis
Corresponding Author:	Ian Hussey Ruhr University Bochum GERMANY
First Author:	Ian Hussey
Order of Authors:	Ian Hussey
Abstract:	Vahey et al.'s (2015) meta-analysis concluded 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. Using an alternative scoring algorithm, the Probabilistic Index (PI), did not consistently improve these metrics. Lastly, the IRAP was also shown to be substantially inferior to the most popular implicit measure, the Implicit Association Test, on each of these metrics. In its currently form, the IRAP is therefore unsuitable for individual level use or assessment in both research and applied contexts.
Suggested Reviewers:	Solomon Kurz, PhD Postdoc, The Chicago School of Professional Psychology a.solomon.kurz@gmail.com Solomon has experience with advanced statistical methods, R, and the IRAP
	Yoav Bar-Anan, PhD Professor, Ben-Gurion University of the Negev barananу@bgu.ac.il Yoav has experience with use and reuse of large scale implicit measures dataset such as those used in this manuscript.
Opposed Reviewers:	

RUHR UNIVERSITY BOCHUM | 44780 Bochum | Germany

FACULTY OF PSYCHOLOGY

Psychology of
Human Technology Interaction

Dr. Ian Hussey
Phone +32 (0)470 396842
Email Ian.hussey@rub.de

August 18, 2022

Manuscript Submission: "The Implicit Relational Assessment Procedure is not suitable for individual use"

Dear Prof Levin,

Please find attached my manuscript "The Implicit Relational Assessment Procedure is not suitable for individual use".

This is a majorly revised version of a manuscript originally submitted around two years ago, which was originally entitled "The Implicit Relational Assessment Procedure is not suitable for individual use due to very wide confidence intervals around *D* scores". This was previously assigned the manuscript ID JCBS-D-20-00033.

I apologies for the long delay in resubmitting this revised manuscript, and hope you're willing to consider it. In the intervening time I was between academic jobs and couldn't work on academic projects such as this.

I have very substantially revised the manuscript based on the reviewers' suggestions. This includes: (1) breaking down results by domain and trial type so that the heterogeneity of results between them is properly conveyed; (2) using meta-analyses to estimate effects across domains and trial-types; (3) including prediction intervals (sometimes called credibility intervals) on the meta-analytic effect sizes, which take into account the heterogeneity between domains and trial types; and (4) comparing the performance of the *D* score with an alternative scoring algorithm, the PI score.

Please see the (anonymized) response to reviewers letter for point-by-point responses to the reviewers. In it I refer to a reference that I had to remove for blinding: that reference is Hussey & Drake (under review at JCBS) https://psyarxiv.com/sp6jx/

Kind regards,

Ian Hussey

Response to reviewers

Thank you to both reviewers for their comments. I have very substantially revised the manuscript based on their suggestions and feel the manuscript is now much stronger thanks to this feedback.

In some places, the reviewers suggested things that the original analyses had actually done, but which the original manuscript was insufficiently clear about. For example, the reviewers correctly pointed out that (1) IRAP D scores and their confidence intervals should be calculated for each trial type, and (2) that comparisons between individuals should only be made within each domain and trial type, in order to compare like-with-like. Both of these were implemented in the original analyses, but I agree with the reviewers that this is simply not clear in the manuscript itself. This was an important flaw in the manuscript and I have corrected throughout the revised manuscript.

I have made several other large changes thanks to reviewers' feedback: (1) Results are broken down by domain and trial type so that the heterogeneity of results between them is properly conveyed; (2) I now use meta-analytic models to estimate effects across domains and trial-types; (3) the results of these meta-analyses include prediction intervals (sometimes called credibility intervals) that convey the heterogeneity between domains and trial types; and (4) I compare the performance of the D score with an alternative scoring algorithm, the PI score.

Please find my itemized response to all questions and comments in italics and coloured blue.

Associate Editor comments

Dear Authors,

Thank you so much for your submission. It reflects what I feel is a solid piece of work, and was quite compelling for the two expert reviewers who provided feedback on the manuscript (pasted below). As your note, they had a number of concerns with the current version of the manuscript that they would like to see addressed. I was pleased to see that the reviewers, while both expert, represented different kinds of JCBS readers, and thus, tended to complement one another in their reception of and recommendations for the manuscript. I hope you find the feedback interesting, and that you agree that responding to it might improve the paper's potential contribution to the CBS literature. I look forward to receiving your revised manuscript, complete with a response to reviewers that details all the changes made, point by point. Please do let me know if you have any questions or concerns about the feedback or the resubmission process.

Emily Sandoz, AE
JCBS

Reviewer #1 comments

While most research looking at implicit phenomena (IRAP, IAT, etc.) focuses on signals (i.e., indices of difference scores) relatively little attention gets paid to the noise inherent in these paradigms. The manuscript takes a sober view of IRAP D-scores looking at several existing data sets and looks at whether obtained D-scores are significantly different from zero. The introduction is rather spartan and more context would benefit the general readership of JCBS. This article is a good fit for this journal given the interest in including the IRAP and its variants in clinically relevant research. While the article addresses problems with interpreting D-scores it misses a large opportunity to identify whether alternate scores proposed as an improved option (PI-IRAP) perform better enough to justify interpreting individual scores. Doing so would substantially increase the value of the paper.

Over 100 IRAP papers have now been published, mostly in journals that are read by the CBS field. There are also now multiple articles that explain and discuss the procedure in great depth. Rather than repeat these again and either (a) having to block-quote large sections of text or (b) risk introducing imprecision by rewording descriptions of the task to avoid plagiarism, I have elected to point readers to the authoritative sources on descriptions of the IRAP if they are interested in more information on the task. I think this is common and representative of a measure that is widely used and known within a given field. For example, the methods section reads:

"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."

I return to the excellent suggestion of adding PI scores further below.

The following comments aim to support and strengthen the manuscript.

1) Page 2, second paragraph: While the total number of trials can be helpful in providing more reliable estimates of effects (barring fatigue) if the standard deviation for individual response times remains in the range of 300-500ms, then no number of trials will be insufficient to have confidence at the individual participant level for D-scores that are based on difference scores that are less than half of a standard deviation for the raw response times. It will serve the audience well for the paper to clarify exactly how variable the raw response data are. The clinically oriented readers will have too distant of a relationship with the raw data to get its full implications.

I thank Reviewer 1 for raising this point. I agree that conveying the variability of raw reaction times on the IRAP can be quite illuminating as to how challenging it is to quantify the effect. I attempted several different plots in order to convey this. However, I found that each of these served to open up a larger and separate discussion about the analysis of reaction time data in general. The revised manuscript is already quite lengthy at around 8600 words. On balance, I decided that this discussion of the principled reasons why it is difficult to quantify the IRAP effect at the individual level is better left for another day, perhaps as a paper in

itself.

2) Page 8, Percent of D-scores that differ from one another: please provide more context for this analysis. Were these individual participant's D-scores within a particular IRAP type (e.g., body shape)? If so, did all studies of a particular theme involve concepts that theoretically would be dimensional with one another where a discriminant analysis via D-scores would be meaningful? If all body shape targets were negatively valenced, then they may not differ much from one another. If random selection occurred across the entire pooled data set, then the rationale needs to be fleshed out further regarding why and when different scores would be predicted. If the general problem is that the procedure tends to produce a restricted set of values (in the context of individual reaction time variability), and thus is not in a position to produce values that would meaningfully distinguish one target from another, that should be made more clear.

Thank you for raising this point – you've highlighted an important oversight in the original manuscript that I've now fixed. The manuscript now conveys that all analyses were conducted within domain and trial type.

I should also point out that I have corrected the method by which I tested the proportion of scores that differ from one another. Following a white paper by the Cornell Statistical Consulting Unit (2008, Overlapping Confidence Intervals and Statistical Significance), I now use the confidence interval on the difference score between each pair of scores.

Discussion:
3) Is it the problem that D-scores are poorly estimated, or is it that their confidence intervals are so wide that D-scores are too imprecise to justify most inferences made in relation to them? Those who have been trained primarily in classical statistics, versus estimation statistics, will be at risk for reading "poorly estimated" non-technically and could interpret this statement as suggesting an alternate algorithm should be sought and there exists a "true" D-score to be estimated. It could be noted that dozens of algorithms have been explored for other implicit tasks such as the IAT and none have been able to escape the problems related to the variability inherent in response time data. Although the MAP for the IAT is smaller, it still covers a wide range of values considering the strength of inferences researchers would like to make in relation to IAT D-scores.

Thank you for this good point on clarity of language. In the revised manuscript I exclusively use the term "imprecisely estimated" rather than poorly estimated. This is on the basis that there is a commonly accepted correspondence between the width of a confidence interval and the precision of an estimate (e.g., Lakens, 2022 'Improving your statistical inferences'; Field & Gillett, 2010).

4) The final paragraph briefly mentions some possible routes forward for those interested in salvaging the IRAP for research interested in individual scores. In particular the PI-IRAP is alluded to (via reference) but not explored. The current critique would be more progressive if the present data set were also analyzed using the PI-IRAP. This is a substantial revision request. If the aim of the paper remains to simply demonstrate that D-scores have too wide of confidence intervals to be interpretable, the paper provides a sober critique but fails to

identify whether there are any viable paths forward for individual-level IRAP research. If however, the aim is to identify whether IRAP data at the individual level are interpretable, including PI-IRAP analysis data within this document will speak to whether this measure is well suited for individual-level analyses as De Schryver and colleagues (2018) suggest. The data set used in this manuscript is much larger and more varied (more topics) than the one used in De Schryver et al (2018) providing a more general assessment of the claims of PI superiority. In addition, confidence intervals calculated for the PI-IRAP values will support parallel comparisons with the D-scores for the same data set. This substantial revision request has the potential to dramatically increase the citation value of this manuscript.

Thank you for this suggestion – I agree. The revised manuscript directly compares the D and PI scores.

5) There are other confidence interval-based approaches to data analysis (e.g., equivalence testing) that leave open the door to intervals other than the traditional 95% CI. The selection of different upper and lower bounds requires theoretical justification (hopefully empirically anchored!). It is possible that IRAP researchers could do the work to identify exactly what size of an effect is of theoretical and practical interest and evaluate whether D-scores or PI-IRAP scores obtained reliably surpass that threshold. One limitation of the present analysis is that it is conventionally conservative in the parameters chosen. If researchers had a basis for justifying more liberal CIs, then a greater percentage of D-scores (or PI-IRAP scores) could be viewed as representing evidence of IRAP effects.

I don't at all disagree with Reviewer 1's comments here, but this would seem to be an issue for another day and another paper, as first this basis would need to be found and then applied within the current analytic strategy. The former task is very much an entire project in and of itself - entire papers are written on the topic of justifying interval widths and cutoffs (eg the Justify Your Alpha paper) - and far beyond the scope of the current manuscript. Luckily, the analysis scripts for the current manuscript are freely available, so if researchers accomplish this basis of justification in the future they can very easily reassess the IRAP's individual level utility.

6) Some IAT researchers have embraced the theory that D-scores reflect cultural/community norms rather than individual responses. The idea being that the variability in responding at the individual level is too high to make pinpoint inferences, but observing similar patterns of responding across a community suggests greater reliability at the community level. Should PI-IRAP analyses prove PI-scores to have confidence intervals that are poorly interpreted at the individual level, then it may be useful to note that like the IAT, the IRAP may point us toward interpreting data in terms of community norms and behavioral histories rather than individual ones.

Thank you for this very well informed comment. This debate about what this idea (i.e., Payne's Bias of the Crowds hypothesis) is ongoing. The behaviorist in me agrees with Payne's premise that, for example, racist environments make racist people. However, his statistical claims do not map onto his verbal claim: recent research has pointed out that Payne's original evidence for this claim is based on a statistical artifact, i.e., is merely the benefits of aggregation on decreased measurement error (Connor, P., & Evers, E. R. K. [2020]. The bias

of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. Perspectives on Psychological Science. https://doi.org/10.1177/1745691620931492). Additionally, the evidence for the implicit bias being an individual phenomenon is quite strong (Nosek & Hansen 2008 being a very convincing demonstration).

Regardless, while the bias of crowds could be one avenue for future IRAP research that some may opt to pursue, I think it's important that we don't be seen to move the goalposts here, or introduce a "rescue hypothesis", as Lakatos would label it. The current manuscript addresses the idea that the IRAP in its current form is likely to be unsuitable for individual level use; a specific question with a relatively precise answer in this manuscript. I think this debate in the current manuscript is a self-contained and coherent one, even if there are other literatures that want to argue that implicit measures generally can be used for other purposes. Nonetheless these are excellent points which have given me things to think about.

Minor comments:

Page 2, first full paragraph: "stand" is a typo

Page 2, second paragraph: "mean" should be "means"

Page 3, participants paragraph: Typo—"20 time" should be "20 times"

Page 8, first paragraph: Typo—"which perennially raised" should be "which is perennially raised"

All corrected

Reviewer #2

*I would like to preface my replies to Reviewer 2 by noting that despite their numerous objections, they seem to actually agree with my two core arguments: i.e., they state "As they rightly point out in their introduction it is inadvisable to analyse individual IRAP (trial-type) D-IRAP scores because they are each based upon a relatively small number of response time measurements" and later "*there is as no published IRAP literature that interprets D-IRAP scores at the level of individual participants". *My manuscript appears to make a point that R2 agrees with, which specifies how the task should and shouldn't be used, and this point is currently absent from the literature. As such, the manuscript surely represents a useful contribution to the literature.*

Thank you for the opportunity to peer review the proposed manuscript. It highlights some interesting considerations in relation to the using the IRAP as a diagnostic tool for individual level assessment of clinically relevant behaviour. The authors focused upon highlighting the need for any such measure to exhibit a sufficiently small standard error (at the individual level) to allow for meaningful clinical distinctions to be reliably made on an individual basis.

They chose to assess this in terms of bootstrapped 95% confidence intervals. At first glance this may seem like a reasonable approach to take. However, as I will explain below there are multiple fundamental problems with the authors' interpretations of the relevant confidence intervals. In my opinion, these problems if left unresolved, are in danger of misleading the readership of the Journal of Contextual Behavioral Science.

The first and most fundamental problem with the authors' analysis is the fact that it is composed entirely of D-IRAP scores that are averaged across IRAP trial-types. In other words, the authors entirely ignore any distinction among IRAP trial-types even though this was the raison d'etre for the IRAP. It only makes sense to interpret an 'overall' D-IRAP score having already established that the four trial-type DIRAP scores comprising it all load on the same common factor (strictly speaking to a comparable extent). Computing an overall D-IRAP score without considering the relative contribution of each of its constituent trial-type D-IRAP scores, or their relationship with each other, is a severe mischaracterization of what a given IRAP is capable of in principle. By mixing together all four IRAP trial-type effects into a single overall DIRAP score it is simply not possible to interpret the functional (aka experimental) meaning of those overall D-IRAP scores.

See response to R1 above. I apologise for the lack of clarity in the original manuscript. Both the original and the revised manuscript did do all calculation of scores at the trial type level and did all comparisons within domain and trial type in order to only compare like with like. No overall D scores were calculated. I have clarified this throughout the revised manuscript.

By design, not all IRAP trial-types are created equally. Typically, the first trial-type is the one chosen to overlap most strongly with whatever (clinically relevant) behavioral function is in question. The IRAP is constructed such that subsequent trial-types must then be composed in distinction from the sample and target stimuli comprising this first trial-type. As a result, these remaining three trial-types will necessarily overlap to a lesser degree with the criterion behavioral function in question. There is even a model that specifically explains why the latter three trial-types are bound to differ from each other with respect to their overlap with any given behavioral function (i.e. the differential arbitrarily applicable relational responding effects [DAARRE] model). While the DAARRE model does not describe all of the reasons why IRAP trial-type D-IRAP scores might differ from each other, it at least establishes the fact that they are bound to systematically differ from each other with respect to any given behavioural function. This means that for any given IRAP, some trial-types are bound to produce DIRAPs that are more functionally valid and therefore precise (i.e. implying lower standard error) than others. The authors fail to account for this in their analyses and thereby systematically underestimate and mischaracterize the psychometric potential of the IRAP. Ideally, the authors should have at least identified which trial-type in each IRAP was chosen as the anchor/basis for the remaining three trial-types. This would have allowed them to assess the statistical precision of any given IRAP on its own terms.

I agree that there is important variation between trial types and domains. The revised manuscript now reports (a) metrics split by domain and trial type so that the heterogeneity can be observed, and (b) meta-analyses across domains and trial types, including prediction intervals (sometimes called credibility intervals) in order to estimate the impact of this heterogeneity.

The second fundamental problem with the authors' analysis is its complete disregard for the quality of the IRAP data it included. Some IRAP stimulus sets are more likely to overlap with their target behavioral function to a greater degree than others. By definition, any given IRAP trial-type will only produce D-IRAP scores that are reliable and valid to the extent that they were composed of sample, target and response stimuli that were chosen to overlap with a given behavioral function that consistently arises in a given sample population (with some given common behavioural history). Unfortunately, a large proportion of published IRAP studies provide very little or no information their IRAP stimulus selection procedures; much less how they relate to sampling some given behavioral function in a given target population that characteristically exhibits some corresponding behavioural history. Ideally, some form of pilot-testing would be used to develop/tune the relevant IRAP stimulus set with respect to the target sample, and thus behavioural function in question. Reports of any such pilot testing are unfortunately lacking in much of the IRAP literature. It seems likely that these stimulus design issues are even worse among the unpublished IRAP literature (i.e. lower quality papers are more likely to remain unpublished than published). This is important because a large portion of the data analyzed by the current authors was from unpublished data from just two researchers; and without any regard for the relative quality of individual IRAP trial-types in that data (even in their published research) - much less with respect to the data available in the wider IRAP literature (i.e. it is puzzling that the authors did not attempt to obtain raw data from any other IRAP researchers in relation to their published IRAP research - this is contrary to the authors' claim in their abstract that they analysed all IRAP data available to them).

The main point here refers to fact that my analyses should make comparisons within-trial types and domains, and that heterogeneity of effect should be quantified, which is now the case. Results remain substantively the same: confidence intervals are too wide for individual use.

I found it slightly odd that reviewer 2 sees it as a problem that I include data "from just two researchers" given that the IRAP literature is an extremely small field. A systematic review that I am currently working on shows that 59% of IRAP articles include its creator, Dermot Barnes-Holmes, as an author. This has never been advanced as a reason to discount or disregard over half the literature.

The manuscript now clarifies that multiple members of multiple productive IRAP labs were contacted when accumulating the dataset. We put out broad invitations to provide data and received data from multiple individuals. The main reasons for excluding data from other researchers, in line with the inclusion and exclusion criteria listed in the manuscript, were (1) that they used older versions of the IRAP that did not produce reaction time level data in an easy to process format, (2) that while they had employed the necessary version(s) of the IRAP that they did not retain these files and so could not send them to us. Additionally, while the manuscript refers to the data "available to two researchers", it should be noted that the datasets included close to a dozen collaborators from multiple labs, including past and current members of lab of the creator of the task, Dermot Barnes-Holmes.

Reviewer 2 also asserts that (a) the dataset being used here is mostly unpublished data, and (b) that unpublished (IRAP) data is generally of lower quality than published data. The first claim is factually incorrect: although the dataset contains some unpublished data, the majority is published (see references in manuscript that describe the data sources). The second claim that unpublished (IRAP) data is generally of lower quality was also not substantiated by Reviewer 2. It may be true that published data tends to suffer from publication bias towards supporting the hypothesis but there is no evidence that they are of lower quality (e.g., O'Boyle, Banks, Gonzalez-Mulé, [2017] The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. doi 10.1177/0149206314527133). Again, reviewer 2 somewhat overreaches here by assuming that the usual standards of IRAP data collection were not followed in some of these studies (e.g., pilot testing of stimulus sets). The manuscript now clarifies that both published and unpublished data were included in the meta-analyses, following Cochrane guidelines.

I'm not sure how else to defend this point other than by pointing out that many of the datasets included in my analysis are from published articles that include the task's creator, Dermot Barnes-Holmes, as a co-author. The conclusions hold in these samples, same as in all other samples that I analyse. Loosely speaking, if Reviewer 2 doesn't consider Dermot qualified to run IRAP studies to a sufficient quality, I don't know who is or what would appease Reviewer 2 here as is then an issue with the entirety of the IRAP literature rather than just the current paper.

The third major problem with the authors' analysis is that it is premised upon a straw man argument. The basic argument set forth by the authors is that a meta-analysis conducted by Vahey, Nicholson and Barnes-Holmes (2015) wrongly claimed that the IRAP can currently be used as a tool for clinical assessment. For example, the authors stated at the beginning of their introduction that "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."
The whole point of the authors' manuscript, as they present it, is to empirically substantiate the contention that existing IRAP researchers (including Vahey et al.) are wrong in considering the IRAP as being suitable for clinical assessment. This is an unfortunate mischaracterization of the IRAP literature and the Vahey et al. meta-analysis more specifically. I am not aware of any published or unpublished IRAP research that either attempts or recommends using the IRAP for the clinical assessment or diagnosis of individuals. It is simply not yet an issue in the IRAP literature.

The authors repeated quoted Vahey et al. (2015) as having "argued that the IRAP has potential 'as a tool for clinical assessment' (p.64)" as the basis for their above rationale. However, this short excerpt of Vahey et al. (2015) is quoted out of context. It has a very different meaning when viewed with respect to the rest of the sentence from which it was plucked in the discussion section. The full sentence in question is "The present paper demonstrates the potential of the IRAP as a tool for clinical assessment and it is hoped that the present meta-analysis will prove useful to clinical researchers who are considering using

the IRAP as a measure." This sentence explicitly refers to clinical researchers as opposed to clinicians, and in the context of a meta-analysis that was solely concerned with group-level effects, the 'potential for clinical assessment' mentioned in that sentence clearly refers to group-level rather than individual-level effects. A few sentences later in the relevant paragraph Vahey et al. go on to further clarify what they mean by this 'potential' - namely, the potential for continuing to improve the precision of clinically-relevant IRAPs via research that systematically refines the IRAP itself (i.e. much like the present authors suggest in the final paragraph of their proposed manuscript). Indeed, the 'Limitations' section of Vahey et al.'s abstract explicitly clarifies the matter (in addition to various other parts of the manuscript) without ever mentioning 'the potential of the IRAP for clinical assessment' at an individual level.

Unfortunately I must disagree with Reviewer 2 here.

Vahey et al. (2015) makes a simple declarative statement, that "The present paper demonstrates the potential of the IRAP as a tool for clinical assessment" (p 64). The expanded quote that Reviewer 2 provides does indeed reference a group of researchers (clinical researchers), but it makes no mention of "group level" use as Reviewer 2 states, nor does it say that the assertion only refers to this group. The APA dictionary of psychology defines clinical assessment as "the systematic evaluation and measurement of psychological, biological, and social factors in ***a person*** *presenting with a possible psychological disorder" (https://dictionary.apa.org/clinical-assessment) - I.e., a person-level inference. This is the literal and common understanding of the claim made in Vahey et al.*

Reviewer 2's implication seems to be that clinical researchers are exclusively interested in group level research, which is simply not the case, and goes beyond the text and readers' likely inferences from it. It is important that we stick to what Vahey et al 2015 actually state and the questions those claims raise, or else we fall into a motte-and-bailey fallacy. That is, authors' strongest claims must be supported, and not merely retreat to weaker forms of the argument when pressed.

Even if we put aside disagreements about this quote, Vahey et al.'s statement represents an important sentiment. I am approached multiple times a year by clinicians who wish to use the IRAP at the individual level, many of whom point to this specific claim and sentence by Vahey et al. The current article is anything but a straw man: it directly addresses a use-case that the CBS community repeatedly asks about, based in part on a specific claim made in the conclusion of the Vahey et al meta-analysis.

Separately, I also must disagree with a different element of Reviewer 2's characterisation that "It is simply not yet an issue in the IRAP literature". Many years of effort could be saved by defining now based on providing this this evidence ahead of time that the IRAP cannot be used effectively in this way. Clinicians, researchers and patients could be saved much effort by defining contexts in which the task is unlikely to be useful ahead of time, including through the current article. Furthermore, my own experience suggests that there have already been such efforts wasted. Some of the clinicians who ask me about the IRAP's utility at the individual level report having already attempted to use it in this way. Just because no published work has used it this way so far does not mean this is not happening – there are

many reasons why it might not (yet) appear in the academic record, including the fact that clinicians often don't care about publishing as much as full time researchers, or indeed the fact that they likely found null results or no utility, and such results are harder to publish.

The fourth major problem with the authors' analysis is their interpretation of the confidence intervals that they computed for each individual D-IRAP score. As they rightly point out in their introduction it is inadvisable to analyse individual IRAP (trial-type) D-IRAP scores because they are each based upon a relatively small number of response time measurements. This is particularly problematic for trial-type D-IRAP scores but still a considerable issue for the overall D-IRAP scores that the authors chose to analyse. This is presumably the main reason why, as I have already stated above, that there is as no published IRAP literature that interprets D-IRAP scores at the level of individual participants.

Indeed, in the last paragraph of their discussion section the authors acknowledge that the number of response trials comprising a given D-IRAP score is a fundamental limiting factor in its statistical precision. For example, the standard errors (i.e. equivalent to ~half the distance spanned on either side of the man by a 95% confidence interval) typically reported with published group-level IRAP effects are dramatically narrower than those reported being reported by the current authors for individual level effects. This means that the former IRAP effects can be used to make clinically meaningful distinctions (at the group level), but the former cannot. Without increasing the number of pairs of response time measurements included in individual-level IRAP effects to a comparable level as group-level effects, it is simply not possible to assess the validity of the corresponding IRAP stimulus set per se. This is already well-known in the IRAP literature. The question remains as to how and whether the standard error/CI associated with individual-level IRAP effects could be narrowed by increasing the number of IRAP trials at the individual level - but the authors are theoretically and empirically silent about this except with parting allusion in their last paragraph.

"Without increasing the number of pairs of response time measurements included in individual-level IRAP effects to a comparable level as group-level effects, it is simply not possible to assess the validity of the corresponding IRAP stimulus set per se. This is already well-known in the IRAP literature."

R2's states: "This means that the [group-level] IRAP effects can be used to make clinically meaningful distinctions (at the group level), but [individual level IRAP effects] cannot." This is not the case. R2's comment here involves a common statistical error: they have confused the dependent and independent variables. Fried & Kievit (2016, 'The volumes of subcortical regions in depressed and healthy individuals are strikingly similar: a reinterpretation of the results by Schmaal et al.') illustrates this error elsewhere.

To illustrate using R2's example: known group membership (IV) is often shown to predict mean group-level IRAP scores (DV) in an ANOVA (which necessarily takes a continuous DV and categorical IV). This does not mean that that the IRAP can make clinically meaningful distinctions, quite the opposite: it means that already known clinically distinct groups cause different mean scores on the IRAP. For R2's claim to be supported, IRAP scores (IV) must

predict clinical group (DV) in, for example, a logistic regression. This type of analysis is very rare in the IRAP literature. Kosnes et al. is one exception.

R2 states "Without increasing the number of pairs of response time measurements included in individual-level IRAP effects to a comparable level as group-level effects, it is simply not possible to assess the validity of the corresponding IRAP stimulus set per se. This is already well-known in the IRAP literature." R2 does not reference an article that make this point "well-known" in the IRAP literature. To the best of my knowledge, no paper has made this point. As such, the current manuscript seems to represent a novel contribution.

Reviewer 2 seems to be disagreeing with my arguments as being simultaneously (a) incorrect and (b) already established facts, and it cannot be both.

The authors could reconceptualize their manuscript as an attempt to quantify the variability of IRAP data at an individual level. They could use this to highlight just how much additional work is needed in the (incremental) design of the IRAP to achieve useful individual-level of analyses of IRAP effects. I for one would very much welcome a constructive analysis examining (the need for and) how to further develop the IRAP for greater precision and accuracy.

I agree. This exactly line of analysis is pursued in a different manuscript, which I unfortunately cannot cite due to peer review blinding, but I have provided a reference to the editor.

However, the authors should be aware that bootstrapped confidence intervals are not a panacea for the statistical instability of small, and (typically) positively skewed response latency samples. When building a sample population distribution using bootstrapping (to calculate 95% confidence intervals using the percentile method) each instance of the sample statistic comprising this distribution is derived from the same sample with replacement (e.g. the same extreme outlier could be selected more than once for a given re-sample even though it was only in the original sample once). If the original sample is small and positively skewed as in the authors' analyses, then the resulting re-sample D-IRAP estimate is bound to vary more from re-sample to re-sample than if it had originated from a larger corresponding sample of response latencies. It is well-established that the bootstrapped percentile method of calculating confidence intervals is systematically biased (toward inflation) with small and positively skewed sample sizes. More fundamentally, bootstrapped confidence intervals assume that the relevant sample points are 'independent and identically distributed' - this is obviously not a tenable assumption with respect to the individual IRAP response latencies comprising a given D-IRAP score. Those response latencies are bound to be related to each other across time (i.e. repeated measures), between consistent and inconsistent blocks, and also in complex confounded ways among trial-types. As such, the bootstrapped DIRAP re-sample estimates that the authors computed were in principle bound to exhibit a greater degree of variability than their non-bootstrapped counterparts. Therefore, the resulting bootstrapped confidence intervals computed from across these repeated bootstrapped estimates were systematically inflated. See the following weblinks for further information on the above points in summary: https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

https://stats.stackexchange.com/questions/355781/is-it-true-that-the-percentile-bootstrap-should-never-be-used
https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/1365-2656.12382

To begin with a constructive point, I have updated the analyses to use the Bias Corrected and Accelerated (BCA) method to bootstrap confidence intervals on individuals' IRAP scores, as R2 recommended paper by Puth et al (2015).

Reviewer 2 argues that "bootstrapped confidence intervals assume that the relevant sample points are 'independent and identically distributed' - this is obviously not a tenable assumption with respect to the individual IRAP response latencies comprising a given D-IRAP score.". Here, Reviewer 2 has misattributed this assumption to being specific to my manuscript: these are not merely assumptions of my analyses, but assumptions of the IRAP's D scoring procedure that the current manuscript consciously inherits. If Reviewer 2 believes these assumptions to be untenable, this is a fatal flaw with the all published IRAP papers, well beyond the current manuscript. However, reviewer 2 seems to hold a niche non-mainstream position on the appropriate analysis of reaction time data (e.g., in contrast with papers such as Ratcliff, 1993 and Whelan, 2008).

*Reviewer 2 states that "As such, the bootstrapped DIRAP re-sample estimates that the authors computed were in principle bound to exhibit a greater degree of variability than their non-bootstrapped counterparts." This claim is incorrect. First, the estimates themselves were not bootstrapped, they are calculated as normal in any IRAP paper. Second, if Reviewer 2 is actually revering to their confidence intervals, its critical to note that to my knowledge there is *no such thing as a non-bootstrapped counterpart to a 95% confidence interval on an IRAP D score*. The variance of the D score effect size has not been defined mathematically, so no SEM*1.96-style confidence intervals can be calculated mathematically. Here, reviewer 2 is referring to methods that don't exist as if they are well known. I went to some lengths when writing this paper trying to find or create such a mathematical strategy and came up empty handed. This is because the D score method of pooling SD is an odd choice from a math perspective, as it means that the numerator and the denominator are highly correlated and the range of the D score is limited to -2 to +2. De Schryver et al. 2015 expand on the odd mathematical properties of the D score, which emerged out of convenience of calculation in Microsoft Excel/SPSS rather than mathematical soundness. If Reviewer 2 can derive how to calculate the variance of the IRAP D score (note: not Cohens' d but Greenwald D, which have different maximum ranges and distributions) I am more than happy to add such scores.*

As an aside, it would be better if the authors explained for the reader, at least in summary, how they calculated the bootstrapped DIRAP sampling distribution used to compute each DIRAP confidence interval. Without knowledge of the R programming language, and the time to sift through the code you refer them to, readers of the JCBS would be unable to determine your methods or therefore reproduce them.

Good idea - I have added a brief explainer on bootstrapping to the results section (p10-11):

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 minimize 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.

It is also worth noting that DIRAP scores are specifically designed (and empirically developed) to minimise the positive skew typically exhibited by raw response latencies (e.g. see Nosek et al, 2003; https://faculty.washington.edu/agg/pdf/GB&N.JPSP.2003.pdf). This begs the question as to why one wouldn't examine the precision of a given trial-type DIRAP score in terms of its bootstrapped sampling distribution among a given sample of individuals for whom that trial-type has a similar behavioral function(s).

I agree; as stated previously this is the analytic strategy employed within the current and previous version of the manuscript.

Incidentally, it is puzzling that the authors claim in their introduction that D-IRAP scores are typically based upon 36 pairs of response times, when 18 is the classic number for individual trial-type scores in the IRAP literature (i.e. six target stimuli and one sample stimulus per trial-type), and 72 is the number of pairs typically comprising the overall scores they ultimately use for their analyses.

I thank the reviewer for catching this. The manuscript now reads: "In a typical IRAP, a D score for a given trial-type is calculated from only 18 pairs of reaction times."

It is also problematic that the authors frequently presented their confidence intervals as if they were credibility intervals. Unlike credibility intervals, confidence intervals do not measure the precision of a given estimate because they are prone to oscillating in both location and width from sample to sample (for a graphical illustration see here: https://rpsychologist.com/d3/ci/; see also http://rynesherman.com/blog/misinterpreting-confidence-intervals/ & http://datacolada.org/28 & http://www.timvanderzee.com/not-interpret-confidence-intervals/).

I agree that confidence intervals are frequently misinterpreted in the literature, and the forms these misinterpretations take (see Greenland et al., 2016). Having carefully triple checked my manuscript I am confident that I have not made such mistakes here. If R2 can find any I am very happy to correct them.

However Reviewer 2 does seem to be mistaken themselves. They state: "confidence intervals do not measure the precision of a given estimate". This is not the case. For example, in his authoritative open source course, Laken's states in the opening paragraph "Confidence intervals provide a way to quantify the precision of an estimate". https://lakens.github.io/statistical_inferences/confint.html. Furthermore, one of the more commonly cited discussions of the difference between confidence and credibility intervals is Field & Gillett (2010), in which they too state "confidence intervals measure the precision of an estimate, whereas credibility intervals reflect whether validity can be generalized" (p.675). Here too, Reviewer 2 seems to have a niche position on what is commonly and correctly understood by this statistical term, which should not be used to discount the current article from publication.

Highlights

- Recent meta-analysis claimed the IRAP has potential for clinical assessment
- Bootstrapped confidence intervals around IRAP *D* scores are very wide
- The majority of *D* scores are not significantly different from zero or from other *D* scores
- Using an alternative scoring algorithm, the PI, did not improve performance
- The IRAP is not suitable for individual use or assessment in research or applied settings

The Implicit Relational Assessment Procedure is not suitable for individual use

Ian Hussey

Author note: Ian Hussey, Faculty of Psychology, Ruhr University Bochum, Germany.

ian.hussey@rub.de. IH was supported by Ghent University grant 01P05517 and the META-REP

Priority Program of the German Research Foundation (#464488178). All data and code is available at osf.io/mb4ph

Abstract

Vahey et al.'s (2015) meta-analysis concluded 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. Using an alternative scoring algorithm, the Probabilistic Index (PI), did not consistently improve these metrics. Lastly, the IRAP was also shown to be substantially inferior to the most popular implicit measure, the Implicit Association Test, on each of these metrics. In its currently form, the IRAP is therefore unsuitable for individual level use or assessment in both research and applied contexts.

The Implicit Relational Assessment Procedure is not suitable for individual use

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), suggesting interest in the individual level utility of the task in both research and applied settings.

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 precisely 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 imprecisely estimated at the individual level. In a typical IRAP, *D* scores are calculated from only 18 pairs of 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 imprecisely estimated.

Surprisingly, no research to date has quantified how well IRAP effects are estimated at the individual level. This study therefore calculated confidence intervals around individual participants' IRAP effects, using confidence intervals around individual scores. This was done using a large open dataset containing many different domains. These intervals were 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 could 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 was 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 was then compared with the a more robust method: the Probabilistic Index (PI: De Schryver et al., 2018).

Finally, recent debate in the broader implicit measures literature has suggested that these measures are particularly noisy and are reliant on group-level aggregation to produce reliable and replicable effects (Connor & Evers, 2020). It is therefore worth considering whether the IRAP's individual level performance is particular to it, or representative of other implicit measures. That is, what is the IRAP's individual level performance relative to a closely related task? In order to assess this question, the IRAP's utility for individual assessment was compared to the most common popular implicit measure, the Implicit Association Test (IAT: Greenwald et al., 1998), again using a large open IAT dataset of assessing many different domains.

Method

Data

IRAP data

Data were taken from a large open dataset of all IRAP research. This dataset was constructed by contacting individuals at multiple labs conducing IRAP research and asking them to contribute their well-organized trial level IRAP data, as well as all stimuli and task parameters to generate that data, to an openly available dataset. The resulting dataset contained all data from two labs [reference blinded for peer review]. Inclusion criteria used for the curation of that dataset were listed by [reference blinded for peer review] 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 [reference blinded for peer review]). 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

Data was taken the Attitudes, Identity and Individual Differences (AIID) dataset: a large, multivariate, planned missing data study on implicit and explicit attitudes within many different domains designed for reuse. See the AIID dataset documentation for the design of the study, all data, code, and materials (Hussey et al., 2019). The AIID dataset was suitable for the current research question because, like the IRAP dataset, it included data from a large number of different attitude domains (95 in total, including race, religion, social groups, celebrities, politicians, political parties, and brand preferences). Inclusion criteria were self-reported fluent English, complete trial-level IAT data, and use of an evaluative IAT. Exclusion criteria were performance-based accuracy and latency exclusions often employed in IAT studies (specifically those used in Nosek et al., 2007). These exclusion criteria are relatively strict, and recent research has suggested they may be conservative (Greenwald et al., 2022). However, the full AIID dataset is over 230,000 participants. This represented an overabundance of data for the current research question, especially given that bootstrapping confidence intervals on individual participants' *D* scores is relatively computationally intensive. Therefore a subset of 100 participants per attitude domain after exclusions were randomly sampled from the full public dataset, to make 9500 participants total. This allowed us to employ the previously mentioned conservative exclusion criteria in order to ensure high data quality. Code to reproduce this random sampling is available in the supplementary materials.

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 was 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 was 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 66.4% women and 33.6% male; M_{age} = 31.6, SD = 12.5. 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 of the general procedure 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 these 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. 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. Typically, but not exclusively, one score is calculated for each of the four trial types on the task. These are formed by the relating of two classes of sample stimuli (e.g., Black people vs. White people) and two classes of target stimuli (e.g., positive vs. negative) to make four trial types (e.g., Black people – positive, Black people – negative, White people – positive and White people – negative).

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. In contrast to the IRAP, IAT scores are almost exclusively quantified as a single score representing an overall bias (e.g., towards responding faster to "Black people – positive / White people – negative" relative to "Black people – negative / White people – positive").

Scoring methods

IRAP and IAT studies typically use the *D* scoring method to convert each participant's reaction times into analyzable scores for each individual. 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 that are above vs. below zero (Finn et al., 2019).

De Schryver et al. (2018) discussed 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 compatibility with the existing IRAP literature, I 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 zero point (i.e., $D = 0$, PI = 0.50), and a negative effect if it is descriptively below it. 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 state that 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.

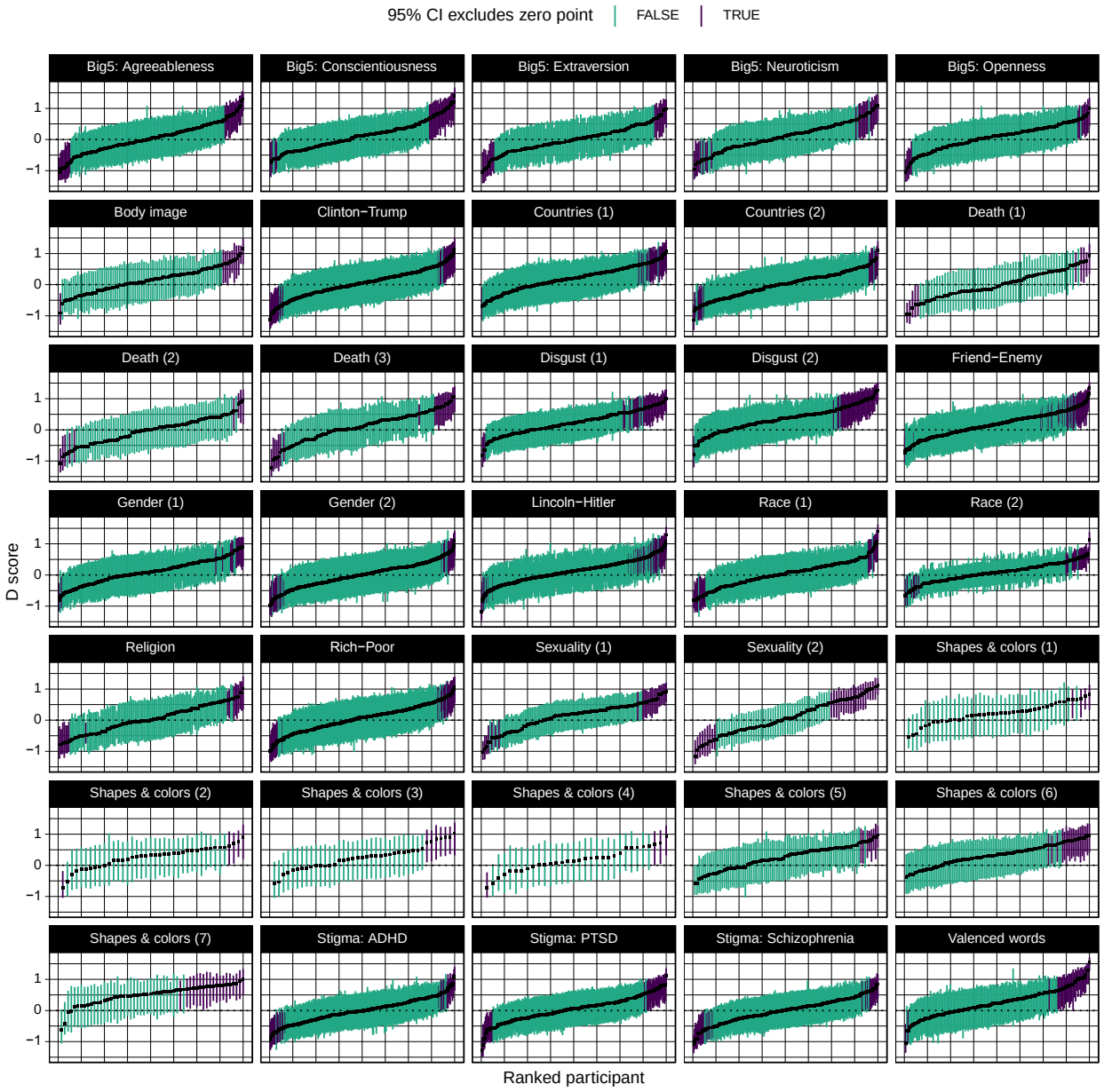
A common method for calculating confidence intervals the arithmetic method (e.g., CI = mean ± SEM*z, where z for 95% interval = 1.96; Swinscow & Campbell, 1997). However, this requires the standard error of the mean of the effect size, or a derivative of it such as its variance, to be specified. To the 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 minimize 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 (https://osf.io/mb4ph/?view_only=1daffd44aeb147b8bdaf7dcdc8862f69).

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

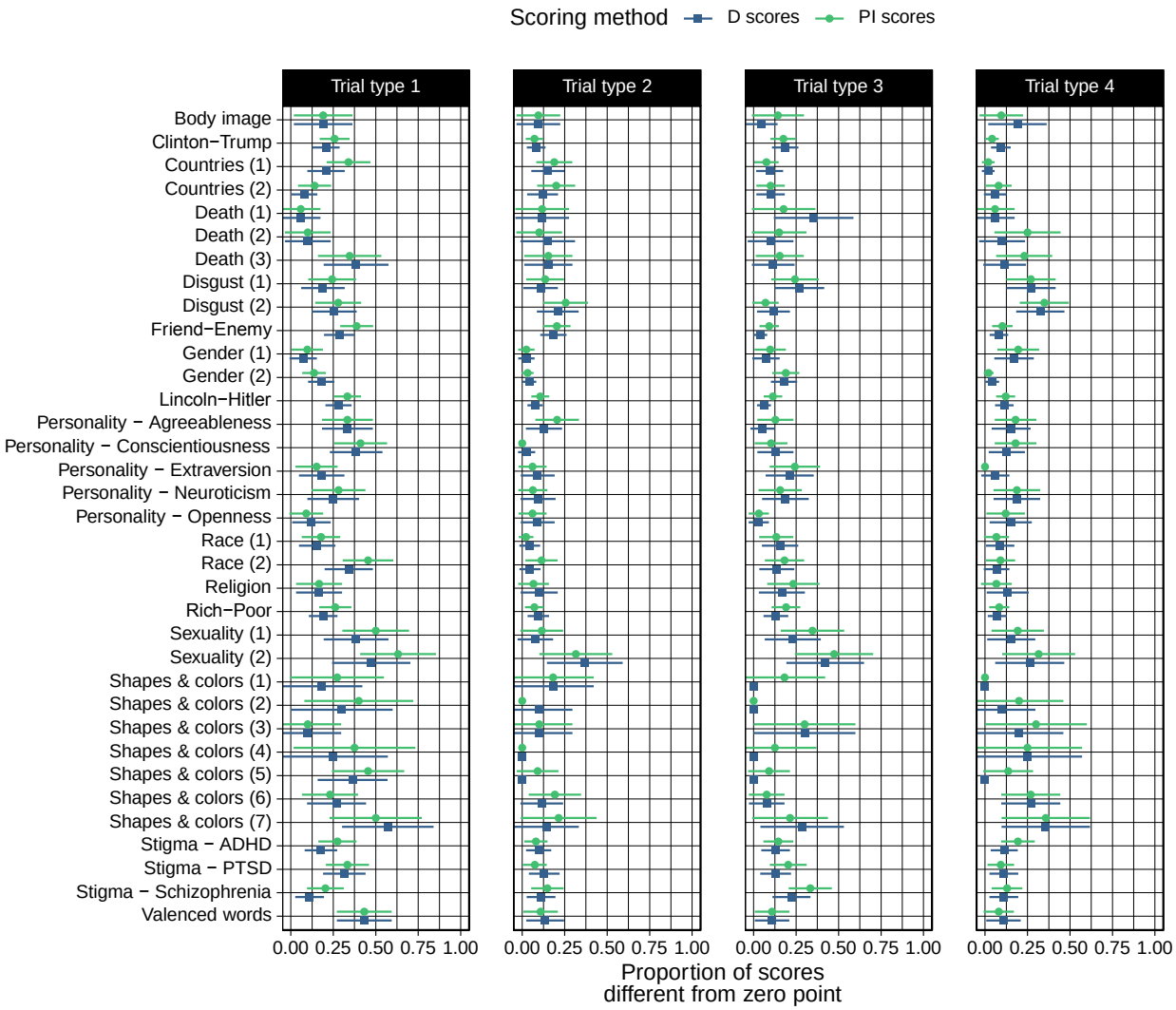
Participants' IRAP *D* scores and 95% Confidence Intervals are presented in Figure 1. These were clustered by domain, and arranged by ranking the participants by their *D* score. This type of plot is sometimes referred to as a caterpillar plot in the meta-analysis literature (e.g., Fernández-Castilla et al., 2020), and a similar form of plot – albeit without 95% CIs – was 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 were colored based on whether the interval excluded the zero point or not ($D = 0$). It was 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 $D = 1.31$ (i.e., D score ±0.66). 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 were 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., from the first publication to the most recent ones: Barnes-Holmes et al., 2006; Kavanagh et al., 2022). 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 were not significantly different from zero, and as such only a small minority of participants could 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.

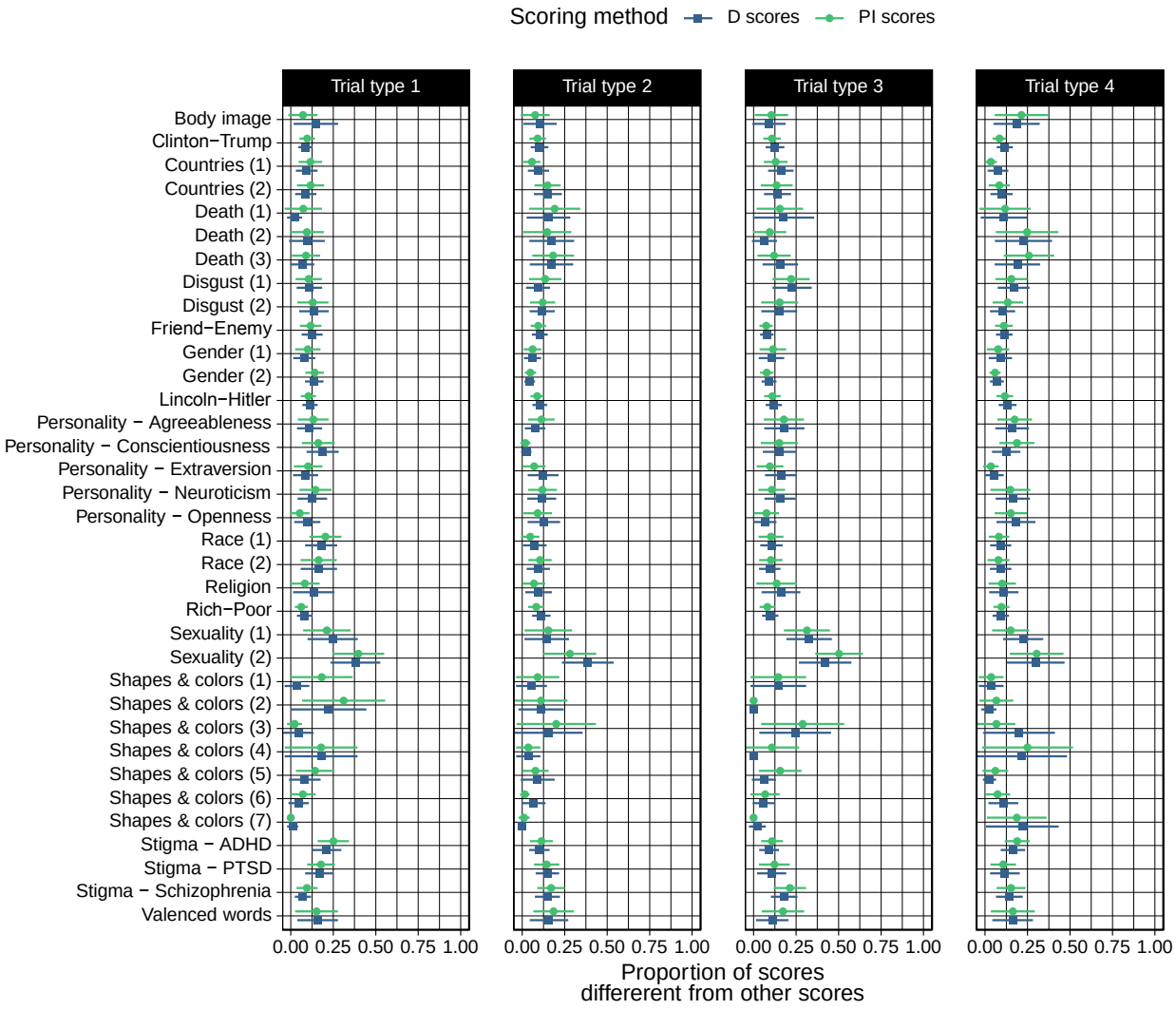
In order to quantify the proportion of individual participants with detectable biases, results were meta-analyzed across trial types, participants, and domains. This and all subsequent analyses were implemented as (meta-analytic) linear mixed effects models using the R packages lme4 (Bates et al., 2015) 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. Any scores of exactly 0 or 1 were offset by a small amount (e.g., 0.0001) to allow the model to run. 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. Any variances of zero were offset by a very small amount (e.g., 0.0001) to allow the model to run. 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, 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. They are sometimes referred to as Credibility Intervals, but this term can be confused with the Bayesian analogue to Confidence Intervals so I avoid it here. 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 $\theta = 0.08$, 95% CI [0.05, 0.12], 95% PI [0.01, 0.47]. These results are depicted in Figure 5 (upper panel, *D* scores) along with the next five meta-analytic models' results. 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 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. 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 sample and target 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).

The difference between two estimates is often assessed through the non-overlap of their confidence intervals. However, this can be shown to be an overly conservative approach on the basis that significant differences can exist when there is a slight overlap in intervals. As such, the more appropriate and liberal test is to assess the confidence interval on the difference score between the two estimates. This point was made particularly clearly in a report by the Cornell Statistical Consulting Unit (2008), whose formula (p.2) was used to assess pairwise differences between scores: the null hypothesis was rejected when

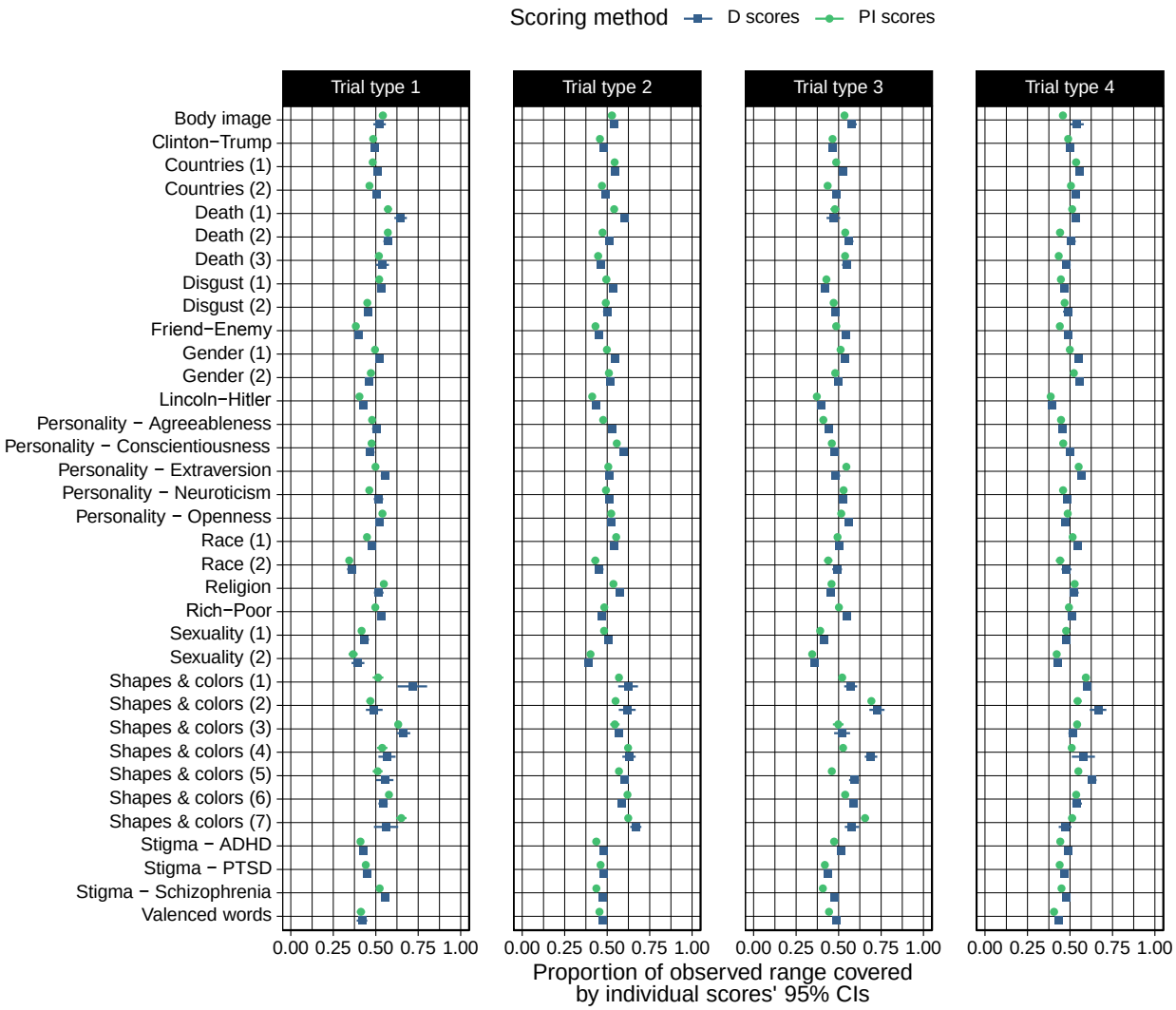
$$|x_1 - x_2| > 1.96 \times \sqrt{SE_1^2 + SE_2^2}$$

The proportion of scores that that could be discriminated from other scores (i.e., the proportion of significant differences) was then calculated and its 95% Confidence Intervals estimated via bootstrapping. In order to only compare like with like, these were calculated within domain and trial type. 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 $\theta = 0.08$, 95% CI [0.06, 0.11], 95% PI [0.02, 0.35] (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 2–35% of individuals' *D* scores were found to be discriminable from the other individuals' *D* scores within the same domain and trial type, typically 8%. 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' IRAP *D* scores' 95% Confidence Intervals within each domain and trial type.



Proportion of observed range covered by individual scores

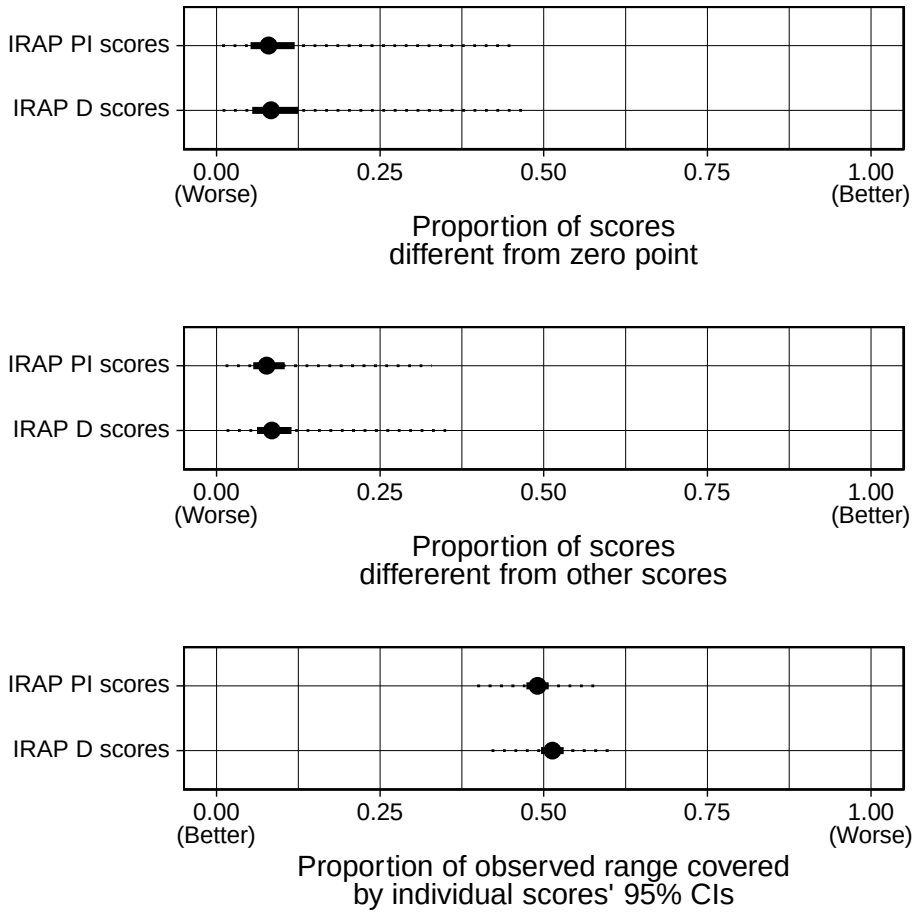
It is also useful and important to understand an individual *D* score's confidence interval width in the context of the maximum range of scores. Loosely speaking, if scores on a given depression scale told us that a given individual lay in the range of 2 to 8, it would be important to know whether the maximum range of the scale is 1 to 100 (in which case the individual could meaningfully be said to be on the lower end of the scale's continuum) or 1 to 10 (in which case

little could be said about where the individual lies on the continuum). If most participants lie within a large portion of all observed scores, and there is assume that there should be genuine variation between individuals, then 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 θ = 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 confidence 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' confidence intervals

covered a large proportion of the total observed range of intervals, this line of evidence also 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 in proportions 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 θ = 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, $\Delta\theta$ = 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 θ = 0.08, 95% CI [0.06, 0.10], 95% PI [0.01, 0.33] (see Figure 5, middle panel, PI scores). Scoring the IRAP with the PI score instead of the *D* score did not improve this proportion, $\Delta\theta$ = 0.00, p = .15.

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 θ = 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 very small, $\Delta\theta$ = -0.02, p < .001.

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

The individual level performance of the IRAP was then compared with that of the IAT in order to provide a direct comparison with a closely related measure. The main purpose of these analyses is to estimate the direction and magnitude of differences between the IRAP and IAT. These analyses compare *D* scored data on both measures, as the most commonly employed scoring method. Note that, because of necessary changes to the random effect model specifications (see next section for details), the estimation of the IRAP's performance can differ

from models 1 to 3 reported earlier. Where any differences exist, models 1 to 3 represent the more appropriate and authoritative results for the purposes of evaluative the IRAP itself (as they fully model differences between IRAP trial types), whereas models 4 to 6 are useful for comparing the IRAP to the IAT for comparison purposes. 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

Model 4 was similar to model 1. The proportion with of IAT *D* scores that exclude the zero point by domain can be found in Figure 4S in the supplementary materials. 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 $\theta = 0.57$, 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, $\Delta\theta = 0.51$, $p < .001$.

Proportion of scores that differ from one another

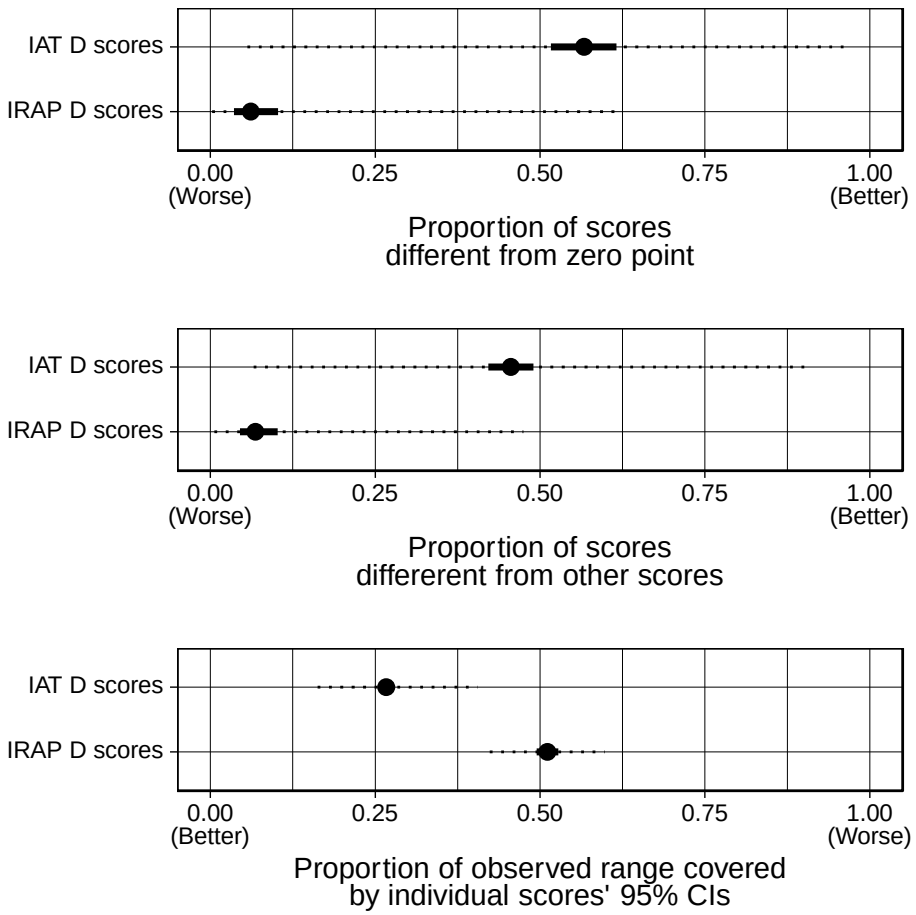
Model 5 was similar to model 2. The proportion with of IAT *D* scores that were discriminable from other *D* scores in each domain can be found in Figure 5S in the supplementary materials. The same modifications to the fixed and random effects were made as model 4. The meta-analytic proportion of IAT *D* scores that were found to be discriminable from

one another was $\theta = 0.46$, 95% CI [0.42, 0.49], 95% PI [0.07, 0.91] (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, $\Delta\theta = 0.39$, $p < .001$.

Proportion of observed range covered by individual scores

Model 6 was similar to model 3. 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. 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 $\theta = 0.26$, 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, $\Delta\theta = 0.24$, $p < .001$.

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



Discussion

In contrast to the conclusions of Vahey et al.'s (2015) meta-analysis, the current results suggest that the IRAP, in its current form, does not currently have potential "as a tool for clinical assessment". Analyses of four different metrics consistently suggested that the IRAP does not have utility at the individual level. A given IRAP *D* score's confidence intervals are likely to be very wide (i.e., a given participant's *D* score ±0.66), with the result that, in the great majority of cases, no IRAP effect is detectible at the individual level (92%), and individuals' *D* scores cannot be differentiated from other participants' scores in the same domain and trial type (92%). A given individuals' *D* scores' 95% Confidence Interval typically spans a very large proportion (51%) of

the observed range of all scores within a given domain and trial type, suggesting that little can be said with confidence about where the individual lies on the continuum that is being assessed.

Except for a minority of extreme scores, an individual *D* score is in general so imprecisely estimated as to allow for almost no inferences about the individual, including where on the scale they lie, that they demonstrated an IRAP effect, whether their score is different from other participants' scores. This point can be illustrated with a simple example: if a participant completed an IRAP and demonstrated a *D* score of 0.30 (i.e., greater than zero and putting them in the 66th percentile of all *D* scores in the dataset), we might traditionally describe this as a positive IRAP effect. If this was on the Black-positive trial type of a race IRAP, this would typically be interpreted as a pro-Black implicit bias. However, when the confidence intervals around *D* scores are considered (e.g., using the most probable confidence interval width: 95% CI [-0.36, 0.96]), we would more accurately say that the participant's data is (a) no detectable IRAP effect was demonstrated (i.e., 95% Confidence Intervals did not exclude the zero point), and (b) their data is equally compatible with them demonstrating a moderately anti-Black ($D = -0.36$) to strongly pro-Black effect ($D = 0.96$).

Results of a second set of meta-analytic models demonstrated that using an alternative scoring method, the PI, does not consistently or substantially improve the IRAP's individual level utility. This suggests that rescoring IRAP data does not represent a simple fix for the issue.

Finally, results from a third set of meta-analytic models assessed whether the IRAP is (a) as Connor & Ever (2020) recently argued, like many implicit measures in that it is a noisy measure at the individual level, or (b) whether the IRAP seems to have particularly bad performance relative to its peers. Across all metrics of individual level performance, the IRAP

performed significantly and substantially worse than the most popular implicit measure, the Implicit Association Test.

Implications for behavior analytic research using the IRAP

The IRAP was recently argued to be better suited to investigating behavior analytic questions than questions around implicit attitudes (Barnes-Holmes & Harte, 2022). It is important to note that the analysis of individual level IRAP data can serve neither theoretical purpose if its effects are very imprecisely estimated, regardless of whether you couch this in psychometric (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 and Harte's (2022) call for a behavior analytic program of IRAP research. However, Finn et al.'s (2019) individual level analyses IRAP data are equally as stymied by the IRAP's imprecise estimation of individual level effects as any explicitly social-cognitive research using the IRAP as a measure of implicit attitudes.

Possible ways to improve individual level utility

Notionally, the estimation of individual scores could be improved. This could take many forms, including substantially lengthening the procedure. 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). Simulation studies would be useful in estimating the relationship between lengthening the task by different degrees may improve individual level estimation. Ultimately, novel empirical studies would be needed to quantify this, given that task lengthening is often less effective in practice due to increased participant fatigue. Other forms of improvement to the task itself are likely

necessary. From a behavioral perspective, it appears that there is a need to enhance stimulus control over behavior within the task in order to improve signal-to-noise ratio 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., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, *32*(7), 169–177.

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.

Connor, P., & Evers, E. R. K. (2020). The Bias of Individuals (in Crowds): Why Implicit Bias Is Probably a Noisily Measured Individual-Level Construct. *Perspectives on Psychological Science*, *15*(6), 1329–1345. https://doi.org/10.1177/1745691620931492

Cornell Statistical Consulting Unit. (2008). *Overlapping Confidence Intervals and Statistical Significance* (Statnews #73). Cornell University.

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

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., & 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., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., Hahn, A., Hehman, E., Hofmann, W., Hughes, S., Hussey, I., Jordan, C., Kirby, T. A., Lai, C. K., Lang, J. W. B., Lindgren, K. P., Maison, D., Ostafin, B. D., Rae, J. R., … Wiers, R. W. (2022). Best research practices for using the Implicit Association Test. *Behavior Research Methods*, *54*(3), 1161–1180. https://doi.org/10.3758/s13428-021-01624-3

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

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., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J., & Nosek, B. A. (2019). *The Attitudes, Identities, and Individual Differences (AIID) Study and Dataset*. https://doi.org/10.17605/OSF.IO/PCJWF

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., Barnes-Holmes, Y., & Barnes-Holmes, D. (2022). Attempting to Analyze Perspective-Taking with a False Belief Vignette Using the Implicit Relational Assessment Procedure. *The Psychological Record*. https://doi.org/10.1007/s40732-021-00500-y

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

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

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88. https://doi.org/10.1080/10463280701489053

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

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

Conflict of interest statement

I report that I have no conflicts of interest with regard to the publication of this manuscript.