**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*](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.*

*I am approached multiple times per year by clinicians who wish to use the IRAP at an individual level who cite Vahey et al. as their rationale for doing so. 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, which is correctly cited as a relevant motivation for the current article.*

*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), and the expanded quote that Reviewer 2 provides does not undermine this strong claim. 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*](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. It is worth noting that the (clinical) assessment literature is built upon such correlational studies, i.e., individual level clinical assessment utility is validated through correlational studies on groups of individuals.*

*Reviewer 2 is also factually mistaken when they state that Vahey et al. was a “meta-analysis that was solely concerned with group-level effects” – it was a meta-analysis of Pearson’s r correlations between individuals, not (for example) Cohen’s d values between groups’ means.*

*Reviewer 2 claims that Vahey et al.’s (2015) limitations section in their abstract clarifies this matter. It does not: there is no qualification on the claim to clinical assessment. I past the entirety of this section:*

*“Limitations: The present meta-effect is an estimate based upon an IRAP literature that is still evolving rapidly in the clinical domain, and so as per its accompanying credibility interval, all conclusions that follow are necessarily provisional even if bounded. Apart from the fact that the current meta-effect might be subject to inadvertent under- and/or over-estimations of the current literature, the present metaeffect might strengthen with further refinements of the IRAP.”*

*Simply put: Vahey et al. (2015) stated in plain English that “The present paper demonstrates the potential of the IRAP as a tool for clinical assessment” and does not at all clarify that they did not mean this simple declarative statement. It is important that we stick to what Vahey et al 2015 actually stated and what a reasonable reader would infer from their words. To do otherwise would be a motte-and-bailey fallacy. Authors’ strongest claims must be tested, rather than being retreated from to weaker forms of the argument once challenged.*

*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 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*](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.*