**Response to reviews**

*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 larger and excellent suggestion of adding PI scores further below in a different reply.*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. My intended point in the manuscript was that a very, very low number of trials is used to estimate IRAP scores compared to other comparable tasks in the literature, in order to emphasise that the IRAP is likely to produce poor estimates at the individual level. In order to convey the variability of reaction times that the IRAP effect is estimated from – something often obscured to the modal reader, as you point out – I have included (a) a plot of the distribution of reaction times, (b) descriptive stats on the modal RT for each block and their SD in text.*  
  
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. Now, the analysis of the proportion of D scores that can be shown to differ from another is conducted only (a) within trial type and also (b) within-domain, so as to compare like-with-like.*

*In addition & on a related point: while revising the manuscript I came across Weir (2005, doi 10.1519/15184.1), who points out that two point estimates with error bars cannot be compared in the same way as one point estimate against a fixed value. E.g., whereas one effect size can be compared against a stable reference point (e.g., 0) using its 95% CIs, two estimates and their CIs can only be compared against one another when also considering the variability in the difference scores between them. Loosely speaking, there are now two moving targets, and this must be accounted for. Weir points out that the CI of one effect’s CIs must be multiplied by in order to use this interval for inference. I have therefore also updated my analysis to take this into account.*  
  
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 around the term “poorly estimated” for a non expert reader. The revised manuscript now clarifies that “poor” here refers to a relatively wide width of the confidence interval relative to the maximum observable range, and that this is a function of a combination of (a) the number of reaction time data points it is calculated from and (b) the properties of the task that produces those data points. As such, for large improvements to be observed in these estimations, changes are likely needed to the task itself and not merely the scoring algorithm used to process the data.*   
  
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.

*Given that this manuscript employs analyses that are not typical for IRAP manuscripts, my goal in it has been to try to balance the complexity with and clarity of message. In retrospect, I think you’re right that I leaned too far towards simplicity. The revised manuscript also includes PI scores to show that this doesn’t improve much, as well as other forms of complexity (e.g., multiple bootstrapping methods to show robustness).*   
  
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. The published IRAP literature uses a fairly homogenous set of statistical practices and small samples sizes, and the current manuscript seeks to move this forward two steps for this literature: a unprecedently large and open dataset, and a novel statistical method. Certainly, there will be more steps in the future, but this will be highly dependent on researchers rising to the task of producing other large datasets and embracing more innovative statistical approaches. I hope that the open dataset provided with this paper at least provides a first step that allows others to freely to explore new avenues for IRAP research like the one Reviewer 1 suggests there.*

*The issue here is not one of the complexity of running the analyses proposed – indeed this is only a few lines of code to respecify the confidence interval widths – but rather one of direction and space. Entire papers are written on the topic of justifying interval widths and cutoffs (eg the Justify Your Alpha paper), and this is beyond the score of the current paper.*  
  
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. Importantly, 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 unable to be used at the individual level; 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 (a different debate for a different day).*   
  
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 actually agree with my core argument: “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.” Other than this, the majority of their points are either unsubstantiated, incorrect, or differentially applied to the current manuscript. I am surprised by this adversarial tone given that fundamentally we agree that the IRAP as currently formed should not be used for individual purposes. Given that this point has not been made in the published literature, the current manuscript would seem to be a useful contribution.*

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.

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.

*Reviewer 2’s is incorrect here: no current or version of the manuscript analysed overall D scores. Individual trial types were always analysed, just not reported separately. This has now been updated in the manuscript, see reply to Reviewer 1’s comment above.*

*While this point does not apply, it’s perhaps worth noting however that some of Reviewer 2’s positions here were nonetheless out of line with the actual published IRAP literature: several published studies have indeed reported overall IRAP D scores in the absence of evidence that they load onto a single factor, including (but not limited to) the original IRAP publication (Barnes-Holmes et al., 2006). As such it can hardly be described as the IRAP’s raise d’etre.*   
  
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 again refers to fact that my analyses should make comparisons within-trial types and domains, which is now the case. Results remain substantively the same: confidence intervals are too wide for individual use. Again this is for very simple reasons: the small number of reaction times used to calculate D scores in the IRAP. No amount of what Reviewer 2 might accept to be extremely high quality data can overcome this. I encourage Reviewer 2 to assess this for themselves: my manuscript contains open source code that can be applied to any IRAP dataset. If Reviewer 2 maintains that results are a function of the quality of the data, I encourage them to use what they accept to be high quality data and demonstrate different results.*

*Separately, 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 58.8% 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.*

*Additionally, Reviewer 2 asserts that no attempt was made to contact other researchers for their data. I’m unclear as to what basis Reviewer 2 can make this claim. Multiple members of multiple productive IRAP labs were contacted when accumulating the dataset. I can no longer check as the listserv infrastructure changed and is no longer searchable (to my understanding), but I’m fairly confidence I even made a public post to the ACT/RFT listserves soliciting suitable data from anyone who had it. We were contacted by and received data from multiple individuals. The main reason for exclusion of some of the data we were provided with from these labs 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. Other prominent researchers who were contact either did not reply to our requests or refused to share their data. 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 members 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 vast 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). 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, including studies that were included in Vahey et al.’s (2015) meta analysis of the IRAP’s validity. 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.*

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. 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. Reviewer 2’s expanded quote simply does not support their argument. 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. Authors’ strongest claims must be supported, and not retreat to weaker forms of the argument when pressed. 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.*

*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. Indeed, two paragraphs from now Reviewer 2 acknowledges that they agree with my core point that this is not suitable, and this point has never been made in a published article.*

*Separately, I also must disagree with a different element of Reviewer 2’s characterisation. Even if it were true that “It is simply not yet an issue in the IRAP literature”, this is not a good reason to not discuss the issue ahead of time. 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.

*I will briefly interject on Reviewer 2’s longer comment here to point out that no article in the published IRAP literature has actually made this point – except for the manuscript we are currently discussing. It seems that, fundamentally, Reviewer 2 and I agree here on this point that is currently absent in the literature. If so, some form of this argument is surely worthy of publication given its novelty. Readers currently have no citable source for this rule about how the IRAP should and should not be used.*

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.

*Quick interjection: reviewer 2 is incorrect here, no current or version of the manuscript analysed overall D scores. Individual trial types were always analysed, just not reported separately.*

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.

*Here, Reviewer 2 is objectively out of line with the demonstrable facts of the published IRAP literature. They argue that this point “is already well-known in the IRAP literature” without reference to a single paper. To the best of my knowledge, having recently systematically reviewed the RIAP literature, no such paper exists. If Reviewer 2 can point to such a paper that was published prior to the writing of this manuscript, I will hear this point and revise my position. It is extremely unclear to me whether Reviewer 2 disagrees with my arguments as being incorrect or disagrees with them as being already known to be true, 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.

*To quickly interject, this exactly line of analysis is pursued in a different manuscript: see Hussey & Drake (under review) https://psyarxiv.com/sp6jx/*

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>

*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 the assumptions to my manuscript: these are not merely assumptions of my analysis, but assumptions of the IRAP’s D scoring procedure that the current manuscript consciously inherits. I.e., these assumptions are common to the IRAP D score. If Reviewer 2 genuinely believes these assumptions to be untenable, this is a fatal flaw with the vast majority of published IRAP papers, well beyond the current manuscript. It’s also worth noting that this opinion runs contrary to the majority of thinking not only in the IRAP literature but the analysis of reaction times generally (e.g., classics such as Ratcliff 1993, or Dermot Barnes-Holmes’ former PhD student Whelan 2008). Reviewer 2 appears to have a very niche position on the appropriate analysis of RT data. If it were not for the blinding of peer review, I would be interested to see an example of how Reviewer 2 analyses their own reaction time data in a way that does not make these same assumptions.*

*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, and would strongly preregister that the same pattern holds across them to.*

*Finally and more constructively, Reviewer 2 claims that the percentile method of bootstrapping is known to be biased. In order to demonstrate that the conclusion is robust, results in the manuscript are now reported using all four commonly accepted methods of bootstrapping. Identical conclusions are observed across them. The key point to be appreciated here is not only that the results are robust to different bootstrapping strategies, but that (mathematical) non-bootstrapping strategies for calculating 95% CIs on IRAP D scores simply do not exist to date. So, the current manuscript represents a best effort to quantify the fact that IRAP D scores should not be used at the individual level.*

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:

*“Confidence intervals were calculated for each trial type, within each person, within each domain.”… “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. It is therefore particularly suitable for estimating confidence intervals around D scores, given that (a) the D score does not have a widely accepted method to calculate its standard error and (b) the distribution of reaction times used to calculate D score typically violates parametric assumptions. In this case, bootstrapping merely involved calculating D scores using random samples from the data, with replacement, a large number of times. The resulting distribution of bootstrapped D scores was then parameterized. For the percentile method, which is most intuitively accessible, confidence intervals were defined as the 2.5th and 97.5th percentiles of this distribution of D scores estimates. For a book length introduction to bootstrapping see for example the classical text by Mooney et al. (1993).”*

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, and 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, but also note that 72 reaction time pairs is for an overall D score. Taking R2’s suggestion above that IRAP data is typically and should be analyse at the trial-type level, I have therefore corrected this number in the introduction to refer to 18 trial type pairs for a typical trial-type IRAP D score (72/4). Unfortunately, this corrected number actually lowers rather than raises the figure:*

*“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 (see Greenland et al., 2016), but having carefully re-read my manuscript I am confident that I have not made such mistakes here, and Reviewer 2 has not cited any specific examples of errors. A search of the manuscript for the word string “precis” returns only the following results, none of which support Reviewer 2’s point here:*

* *“The specifics of the D score have been discussed in precise detail in other publications (e.g., Barnes-Holmes et al., 2010; Hussey, Thompson, et al., 2015), and therefore will only be summarized here.”*
* *“The estimation precision of the IRAP D score was also assessed by examining what proportion of randomly selected D scores were significantly different from other randomly selected D scores.” [NB: this does not equate CI width with precision]*
* *“It is also worth noting that similar analyses of data from another implicit measure, the Implicit Association Test, suggests that the IRAP’s estimation precision is substantially worse than the IAT’s”*

*Furthermore, Reviewer 2 seems to be mistaken themselves: the relative width of a confidence intervals are commonly and correctly interpreted as informing us about the relative precision of an estimate (e.g., see Laken’s authoritative course where he states “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)*). The term “Credibility Interval” is used in two senses, either referring a Bayesian form of confidence interval, or the prediction interval around a meta analytic effect size that takes the width of the standard deviation of the random effect components into account. Neither of these forms of credibility interval represents a metric that is not “prone to oscillating in both location and width from sample to sample”. Finally, a google scholar search for "credibility interval" "precision of estimate" returns only two results, neither of which support Reviewer 2’s position. Here again, 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.*