**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, I have elected to point readers to authoritative sources if they are interested in more information. 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 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.

*XXX*  
  
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.

*XXX*  
  
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.

*XXX*  
  
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.

*XXX*  
  
  
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: a large and open dataset, and a novel (for this literature) 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.*   
  
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.

*This is an insightful comment. I too have followed this suggestion by Payne and others re IAT D scores. However, the evidence for this is so far relatively weak: one or two studies by Payne, but nothing that fully rebuts the very strong evidence provided by Nosek & Hansen 2008. I have added some discussion of this to the discussion XXXXX.*

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**

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.

*I agree with Reviewer 2, and for this reason the manuscript does indeed report D scores calculated at the trial-type level. However, Reviewer 2 highlights that this point was not made clear in the previous manuscript, and I thank them for pointing this out. I have clarified this as follows:*

*XXX*

*Additionally, I have added additional output from the analyses which show that no trial type shows superior performance to the others. (note that all IRAP data was processed so as to assign the theoretically dominant trial type for each IRAP to “trial type 1”).*   
  
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).

*Here I must disagree with reviewer 2. The analyses already present output at the domain level (see Figure 2), and the results do not demonstrate any significant variation in performance between domains. This article presents an unprecedently large dataset for the IRAP literature in terms of both number of domains and total sample sizes. In addition, many of the IRAPs included employed or replicated IRAPs from published IRAP studies, including those used in domains showing clinical relevance, and studies included in Vahey et al.’s (2015) meta-analysis of clinically relevant IRAP effects. If there is a fundamental issue with the IRAPs employed in this article, this is a fundament issue for the IRAP literature as a whole and therefore one that is important to discuss (via the current article), not a flaw specific to the current article. It’s also worth noting that, having systematically reviewed them recently myself, not a single published IRAP article of the over 100 published to date reports such pilot testing. It would be inappropriate to hold this one study to a different standard, especially given the already large sample sizes.*

*With regard to obtaining data from other labs, Reviewer 2 is mistaken in this regard: I sounded out multiple other researchers to contribute their data. In all cases, researchers were unwilling to share their data, the data was unavailable, or was in a format to make it unfeasible to process into the required format for analysis. It is of course true that it may have been possible to collect even more data again, but it is important to not miss the wood for the trees here: the median IRAP study has a sample size of just under 40 compared to the 889 reported here, and this sample size is more than adequate to substantiate the claims made. That said, for the sake of precision, I have made the following rewording in the manuscript:*

*“Using all published and unpublished file-drawer data already available to me”.*  
  
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 target group of researchers (clinical researchers), but it makes no mention of “group level” use as Reviewer 2 states. The 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. The reviewer’s expanded quote 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. I.e., authors strongest claims must be evidenced, and not retreat to weaker forms of the augment 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. Regardless of Vahey et al.’s possible intentions here when they said this but possibly meant something else, they make a simple declarative statement that has had a bearing on the behaviour of readers of their article. Indeed, I am approached multiple times a year by clinicians who wish to use the IRAP at the individual level, many of whom point to this specific claim and sentence by Vahey et al. The current article is anything but a straw man: it directly addresses a use case that the CBS community repeatedly asks about, based in part on a specific claim made in the conclusion of the Vahey et al meta-analysis.*

*Separately, I also must disagree with a different element of Reviewer 2’s characterisation. 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. 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.

*First, Reviewer 2 states “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.*” *It’s surprising to me that R2 considers agrees that this is an open question – i.e., can the IRAP be used at the individual level – and yet seems to resistant to exploring answers to this question using the data and analyses this article presents. My point here is that R2 seems to at least acknowledge that this is a question worth asking – I agree.*

*Second, R2 argues that “the standard errors…typically reported with published group-level IRAP effects are dramatically narrower than those reported being reported by the current authors for individual level effects.” I am aware of no evidence that this is the case, and it is presented without evidence. I am very happy to hear, for example, specific critique of the analytic techniques that I have employed here that may suggest my estimates are incorrect. Data come from a large number of studies conducted in multiple domains and at multiple labs. There is no evidence to suggest that these specific studies have produced fundamentally different data to other published IRAP studies. I would like to stress that I am very open to making changes based on critiques of methods or analyses. However, this reads as more of a rejection out of hand that these studies are valid, for unclear reasons.*

*Third, R2 raises questions about how to assess the validity of the IRAP stimulus sets at the individual level and whether this corresponds with their validity at the group level, and suggests that the current article does not do this. I entirely agree – this is a completely different research question and beyond the scope of the current article, which has a clearly defined research question (are confidence intervals around IRAP effects sufficiently narrow to allow for individual level use) that does not include this. It is important to be circumspect about what any one article in the (IRAP) literature can or has typically accomplished in the past, and ask whether we are applying the same standards here. Lastly, I must reiterate again that several of the IRAPs employed within the current article are replications/reuses of IRAPs already used in published research. So, any questions about their validity are questions for the field generally, not the current article.*

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. 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>

*“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.” – this assumption is at the core of the IRAP D score. I am not arguing for or against this assumption, I am merely reflecting the modal practices in the IRAP literature….*

*R2 suggests that the bootstrapping method I employed may inflate the width of confidence intervals. This is an interesting suggestion and I thank the reviewer for this point. In order to address it, I have also calculated confidence intervals using an alternative method, i.e., numerically from the means and standard errors of the means (i.e., M +/- SE\*1.96). Results support the same conclusions as using the bootstrapped percentile method, suggesting that conclusions are robust to the method used/are not an artifact of the method. I now note this in the text: XXXX.*

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.

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 primary analyses, this involved the percentile method: confidence intervals were defined as the 2.5th and 97.5th percentiles of this distribution of D scores estimates. For the standard error method, which was added as a sensitivity analysis, a the standard error of the mean (SEM) was calculated, and confidence intervals were defined as the Mean ± SEM\*1.96. For a boot 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).

XXX  
  
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):*

*“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 must disagree with R2’s characterisation of my use of the term confidence interval. I agree that confidence intervals are frequently misinterpreted in the literature, but having carefully re-read my manuscript I am confident that I have not made such mistakes here. I exclusively refer to confidence intervals being used for statistical inference following NHST principles. For example, the results discuss the proportion of cases in which D scores can be said to be different from one another based on the nonoverlap of their confidence intervals. At no point do I (incorrectly) claim that, for example, confidence intervals contain the true value with 95% probability, or other common misconceptions. If R2 can highlight some specific instance of an error I am happy to correct it.*

*Furthermore, it should be noted that credibility intervals (in the frequentist use of the phrase but not the Bayesian, noting that I employ a frequentist analysis here), which take into account heterogeneity among estimates, are always at least as wide as confidence intervals. So, were one to calculate credibility intervals (which I would not endorse here as it asks and answers a different question, but to entertain the thought for a moment), the results would be either the same or even worse news for the use of the IRAP at the individual level, but by definition not better news. That is to say, the use of “confidence intervals” is both coherent and correct within the current article to the best of my knowledge. I am happy to correct specific examples of misuse if they are suggested.*