Manuscript Number: JCBS-D-22-00266     
  
The Implicit Relational Assessment Procedure is not suitable for individual use  
  
Dear Dr. Hussey,  
  
Thank you for submitting your manuscript to the Journal of Contextual Behavioral Science. The AE, Dr. Rogge, received 3 reviews of your manuscript and provided his own feedback. Based on these reviews (found below) we have decided to ask you to revise and resubmit. We are asking that the revised manuscript be submitted by Mar 06, 2023.  
  
When you resubmit your manuscript please take care to note all comments along with how they were addressed or why they were not addressed in a separate Response to Reviewers file.  Also, please ensure that no identifying information is included in the Response to Reviewers, as this would unmask the reviewers and delay processing of your manuscript considerably. For example, do not sign the Response to Reviewers or provide it on letterhead.  
  
To submit your revised manuscript, please log in as an author at <https://www.editorialmanager.com/jcbs/>, and navigate to the "Submissions Needing Revision" folder.    
  
Thank you for the opportunity to consider your work. Please contact me, Michael Levin, if you have any concerns or questions about this decision, the revision process, or about JCBS in general.  
  
Regards,       
Michael Levin     
Editor-in-Chief    
Journal of Contextual Behavioral Science  
  
Associate Editor and Reviewer comments:  
  
Dear authors,  
  
Thank you for submitting this IRAP manuscript to JCBS. Given the measurement noise associated with implicit measures, I see tremendous value in determining the most effective scoring methods for measures like the IRAP. I have now obtained comments from three reviewers including Emily Sandoz, the previous Associate Editor who handled the first submission of this manuscript. All three reviewers see potential in this manuscript and feel that your changes addressed many of the concerns and suggestions raised in that first round of review 2 years ago. Having said that, the reviewers raised a number of additional issues that remain concerns. As a result, I am recommending a decision of revise and resubmit with major revisions at this point. Please be sure to address all of the comments from all three reviewers, including the extensive suggestions of Reviewer 3 as much as possible. Given the range of opinions across the three reviewers, I cannot guarantee the ultimate acceptance of  
a revised manuscript, as that will largely depend upon the depth and quality of your revisions to address the remaining comments from this round of review.  
  
Please be sure to provide a point-by-point detailed account of the revisions you made in response to each comment in your response letter. It will also be important to provide thorough explanations and justifications for any suggested changes you choose not to implement in your revision.  
  
Thank you again for such an interesting submission to JCBS. I look forward to reading your revised manuscript.  
Sincerely,  
Ron Rogge  
Associate Editor, JCBS  
  
REVIEWER 1 COMMENTS  
I appreciate the opportunity to review this manuscript. I think it would be useful to the authors for me to unmask myself and disclose that I (Emily Sandoz) served as Associate Editor on the initial submission and am now serving as a reviewer.  
  
First, I appreciate the author's careful responsiveness to the reviews. They were demanding, and written from complementary perspectives, which always makes it more difficult to be truly responsive. It seems to me that the author took these reviews as an opportunity to explore and improve upon weaknesses in the manuscript that might limit contribution to the literature. What results is, in my perspective, a greatly improved piece of work mostly ready for publication.  
  
There are two issues I would like to see addressed, however. First, there are some odd grammatical errors that should be taken care of in copy editing:  
  
For example:  
  
p. 9 "I employs"  
p. 24 "and there is assume"  
  
Second, the discussion kind of just ends. While I agree that the author is not responsible for fixing the IRAP or softening the blow of their data, I wonder if they could speak more directly to what they would want readers interested in individual-level clinical assessment (e.g., the clinicians that approach the authors) to conclude beyond "I can't use the IRAP for this purpose." Just due to its length and the complexity of the analyses and the reasoning supporting them (which the author , the article will not be \*fully\* accessible to so many of those that stand to have their behavior changed by the data it describes. I think the accessibility would be improved by the author communicating clearly what they understand about what clinicians are looking for, and how they might advise clinicians who continue to pursue this goal.  
  
I look forward to seeing this manuscript in print.  
  
Best,  
Em  
  
REVIEWER 2 COMMENTS  
The revised manuscript contains clarified and additional analyses that improve the strength of its contribution to the literature. The author has adequately addressed this reviewer's concerns.  
  
Reviewer 2 indicated they were "not aware of published or unpublished IRAP research that either attempts or recommends using the IRAP for the clinical assessment or diagnosis of individuals." This is not a substantive critique of the present analyses. The ability of ACT or process-based therapies to produce changes in implicit beliefs is of broad interest to the ACBS community, particularly since many members of this community harbor rather conventional and/or cognitivist views regarding beliefs as antecedents to behavior. Even those with more behavior analytic views are interested in whether clinical interventions can change the function of relevant stimuli. To these audiences, the IRAP appears to be a promising tool to shed light on these issues. The present analyses suggest that more development work needs to be done on the IRAP and/or alternative approaches need to be taken.  
  
A few minor typo's were identified that should be addressed:  
Page 9, line 18: "employs" should be "employ"  
Page 24, line 2: awkward wording: "is assume"  
Page 24, line 5: the word "be" can be omitted  
  
  
REVIEWER 3 COMMENTS  
Thank you for the opportunity to review the updated manuscript titled "The Implicit Relational Assessment Procedure is not suitable for individual use". I would like to begin by stating that I think there is much value in the proposed manuscript, if only the points being made were tempered by a more liberal interpretation of the manuscript's data and the extant literature. With the exception of some aspects of the word 'therefore', I am utterly in agreement with the core conclusion of the manuscript's abstract: that "In its currently form, the IRAP is therefore unsuitable for individual level use or assessment in both research and applied contexts."

I also agree with the authors that it's worth warning clinicians and/or clinical/applied researchers not to use IRAP scores to diagnose/assess individuals. I too have witnessed some confusion about this when informally discussing the IRAP with people wishing to learn more about the task -- but this has been sporadic and of  
course, like you I would have discouraged such uses on such occasions. Nonetheless, I am open to the possibility that lots of clinical/applied researchers are mistakenly conducting IRAP assessments and/or research that involves interpreting individual trial-type D-IRAP scores as if they are diagnostic of individual people's behavior. I would support the authors' efforts to make such warnings if only they would do so in a more transparent, tentative and inclusive manner. Would the authors, for example, consider recasting their proposed manuscript as being motivated by concerns about being repeatedly approached by clinicians and/or clinical researchers wishing to use the IRAP for individual diagnoses/assessments? There is currently no mention of this in the proposed manuscript and it would be helpful to clarify for the readership of JCBS what motivated the proposed article.

*Author response: Thank you for this point – this omission was not by accident. In previous versions of the manuscript I thought this discussion of motivation and recommendations was too anecdotal for publication. I think it’s still important to be cautious, but buoyed by your encouragement here I have reincluded a few lines on this. Page XX:*

*XXXX*

Perhaps you could poll published IRAP researchers about how often this problem arises for them? This would not take very much time to do because as you say the community of IRAP researchers is not particularly large. It may even be possible for the JCBS and the Psychological Record to provide you with data concerning the number of IRAP manuscripts they have rejected for publication that involved interpreting IRAP trial-type D scores individually. This would strengthen your argument by providing some evidence that clinicians/researchers are actively using the IRAP to provide individual-level diagnosis/assessments outside of the literature (i.e. as distinct from those who have been warned off doing so on an ad hoc basis by you and I). I would support this as a premise for writing an IRAP paper that warned the JCBS readership not to interpret IRAP scores individually.

*Author response: I have elected not to take up this suggestion about polling IRAP researchers on the basis that the reliability of this data may be very low. As a preliminary exercise, I tried to recall myself the frequency of these suggestions and realised quickly that I couldn’t do any better than “sufficiently frequently that I started specifically avoiding discussing the IRAP with anyone who wasn’t an active or aspiring IRAP researcher at conferences.” So in recently years it hasn’t be frequent, but confounded by my avoiding these contexts. A second reason this would be more difficult than the reviewer suggests is that of the active IRAP researchers, many do not see eye to eye and would be unlikely to offer me their help with this question. On the grounds of reliability and feasibility, I opted not to pursue this line of action.*

*I do like the idea about examining reasons for journal rejection, but it would be difficult to attribute rejection to this specific issue without access to the full text of peer reviews, which is often difficult due to journal policy. Even if access was granted, this would easily represent a full mixed-methods research project in and of itself, quite beyond substantiating a single point in the current manuscript. Moreover, my impression and experience is that it is not so much that there are fully-written up IRAP research projects that adopt this approach but failed at the last hurdle of peer review, but that these projects die much earlier in the research process (e.g., in the planning, data collection, or analysis phase).*   
  
I recommend the above course of action because it is highly misleading to premise the proposed article on the following opening statement at the beginning of the manuscript's abstract: "Vahey et al.'s (2015) meta-analysis concluded that the Implicit Relational Assessment Procedure has potential 'as a tool for clinical assessment'. Here I present evidence to the contrary." I am perplexed as to why the authors' response to my previous feedback refuses to acknowledge that the following interpretation is even possible: "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." I for one am willing to admit that the relevant paragraph at the end of Vahey et al (2015) could be misunderstood and could in hindsight, with regard to this newly raised issue, be re-written more clearly (i.e. with regard to the fact that the current version of the IRAP does not produce scores that can be individually interpreted). I would support the authors' of the proposed article in making this point provided that they also acknowledged that this was not necessarily the intent of Vahey et al (2015). It would only be fair  
to acknowledge that none of the three authors of Vahey et al (2015) have since promoted the interpretation of individual IRAP scores for clinical/applied purposes in any literature that I am aware of. Are the authors' of the proposed manuscript aware of these practices being promoted privately by any of the authors of Vahey et al (2015)? If not, then in good faith, they must relent from casting aspersions about these authors' intentions based solely upon one quote taken out of the context with respect to the remainder of the sentence and paragraph within which it was embedded -- none of which recommended or mentioned recommending the interpretation of individual IRAP scores using the current version of the IRAP. In fact, none of the rest of Vahey et al (2015) stated anything about interpreting IRAP effects individually -- rather the whole paper was concerned with group-level analyses. The whole point of that paragraph was to promote the need for further development of the IRAP's experimental precision.

*Author response: XXX*  
  
As an aside, I would recommend that the authors would remove the word 'recent' from the following bullet point in the proposed manuscript's 'highlights' section: "Recent meta-analysis claimed the IRAP has potential for clinical assessment". The meta-analysis to which it refers, Vahey et al (2015), was first published almost eight years ago and many new methodological and theoretical perspectives have since emerged in the IRAP literature.

*Author response: I have rephrased the point in the highlights to “Researchers have shown interest in the IRAP’s potential for clinical assessment”*

The beginning of the proposed manuscript's introduction further stated: "There is increasing interest in using the IRAP at the individual level. For example, in their meta-analysis of clinically relevant IRAP studies, Vahey et al. (2015) concluded that the IRAP has potential "as a tool for clinical assessment" (p.64). Subsequently, a recent study has reported individual level analyses of IRAP data (Finn et al., 2019), suggesting interest in the individual level utility of the task in both research and applied settings." It is true that Finn et al. claimed the following of in their discussion section: "A feature of the current research, therefore, is that it appears to show some promise in using what many consider to be a measure that is best used only at the group level of analysis for individual participant research." However, at no point did Finn et al. recommend that individual IRAP scores could currently be used for the purposes of diagnosis/assessment on an individual basis. They merely claimed that their research findings tentatively raised the prospect that individual level research might be feasible, and as such did not support the aforementioned premise of the proposed article.

*Author response: You state that Finn et al. (2019) does not call for the IRAP could be used for diagnosis/assessment. You are correct, but your implication is that the quote from the current manuscript is at odds with this, when it is not. As per your quote, the line in the current manuscript is “there is increasing interest in using the IRAP at the individual level.” Your quote from Finn et al. (2019) supports the accuracy of this. To restate your quote: “it appears to show some promise in using what many consider to be a measure that is best used only at the group level of analysis for individual participant research." I have not mischaracterised Finn et al. (2019) as calling for assessment, I have cited it in a sentence that states that there is interest in using the IRAP at the individual level. Thank you for suggesting this quote from Finn et al. (2019), I have included it in the manuscript to further reinforce the point regarding interest in using the IRAP at the individual level, and to ensure that it is clear that the authors of Finn et al. (2019) have their words accurately represented here. Page XX:*

*XXX*

*NB Your comment here used the word “diagnosis”. Please note that at no point whatsoever do past or present versions of the current manuscript employ the term “diagnosis” or any version of this word stem.*

However, I would like to acknowledge that it was unfortunate that Finn et al (2019) referred to their analyses as being individual-level because they clearly weren't; and I would disagree that their data provided a demonstration of the useful interpretation of individual IRAP scores for research purposes. It is true that the final subsection of Finn et al's (2019) results section is titled: "Individual-level analysis". However, this title is unfortunately quite misleading and problematic because none of the analyses it contains involve interpreting IRAP scores individually per se (i.e. in absolute terms). To the contrary, all of the analyses contained within that section are group-based analyses and explicitly referred to as such within that sub-section despite its title. More specifically, this sub-section included a binomial probability test along with multiple group based descriptive statistics such as group averages and inter-individual frequencies. This sub-section did also contain a figure that illustrated the full range individual-level IRAP scores that were of interest -- but this figure does not constitute an individual level analysis because it characterizes the observed inter-individual distribution of observed IRAP scores collectively; a group-level analysis. More specifically, those individual-level IRAP scores in that figure were only interpreted with respect to each other on a rank ordinal basis at the group level, not individually and certainly not individually with respect to any independent benchmark/diagnosis/prognosis. The authors of the proposed manuscript could certainly use Finn et al.'s conclusions about individual-level IRAP research as an opening premise that would justify their concerns about the possibility of confusion emerging in the literature concerning whether IRAP scores can meaningfully be interpreted on an individual basis. However, it would be misleading to claim that Finn et al (2019) recommended interpreting IRAP scores individually except on an exploratory basis in basic research. They at no time recommended interpreting IRAP scores individually for applied diagnosis/assessment, or even within applied research.

*Author response: Good point about their actual analyses, I agree. I have reworded references to Finn et al. (2019) to refer to their statements about interest in the IRAP’s individual level utility rather than their actual analyses. I will also contact the authors of Finn et al. (2019) to point out this issue and solicit their interest in issuing a correction here. Page XX now reads:*

*XXX*   
  
I would like to thank the authors of the proposed manuscript for clarifying that their analyses were performed in terms of trial-type D scores rather than overall D scores; and that they have also stratified their analyses by domain to test whether this generated any heterogeneity. Likewise, I was glad to see that the authors have also incorporated credibility intervals into the latest version of their analyses. I appreciate that a lot of work went into updating their analyses. Nonetheless, some of my initial concerns still remain. I will do my best to lay them out as briefly and constructively as I can:  
  
-- When I read back the following comment that I made in my previous round of reviews I must admit that I cringed: "The second fundamental problem with the authors' analysis is its complete disregard for the quality of the IRAP data it included." I apologize for the blunt and quite possibly offensive way that I phrased this. In hindsight, I can see how this could be interpreted as questioning the ability of the authors' of the proposed manuscript to conduct high quality IRAP research. This was not my intention but was most likely due to haste. The point I was trying to make in the rest of the relevant paragraph in my feedback is that the proposed manuscript does not acknowledge the fact that not every piece of IRAP research ever conducted, even by Prof. D. Barnes-Holmes, should or indeed has been published (e.g. lots of IRAP research has been conducted/designed by undergraduate students with little research training and the quality of such research consequently varied  
widely).

*Author response: It may be the case that lots of data is collected by people with little research training, but does not apply to the data under consideration here. All data was designed and collected by IRAP experts with an established record of doing so successfully.*

I don't know of anyone who has a 100% success rate with regard to producing high quality empirical research. I also admit that the peer review process is biased towards non-null findings among other problems, and I am generally in agreement with the arguments put forward by O'Boyle et al (2017). However, it affirms the consequent, to conclude on this basis that the absence of any peer review is a superior option. It begs the question, what are we doing here engaging in the current process of peer review if it does not add some value; if it is not an important (even if not absolute) source of quality assurance? As such, I would recommend that the proposed manuscript provides at least some information about the proportion of unpublished research that it contains (i.e. it currently doesn't), and/or additional information about any quality control procedures that were performed (i.e. in addition to the selection criteria mentioned, which are relevant to quality controlbut weren't explicitly presented as such in the manuscript). As the proposed manuscript currently stands, it does not sufficiently deal with the issue of quality control to justify the claim that quality control can be taken for granted, without any need for peer review; and this is the point I was originally trying to make albeit rather clumsily.

*Reviewer response: I have no idea what you’re talking about at this point. We are peer reviewing the claims made on the basis of this data right now. You are suggesting \*\*as part of the process of peer reviewing this work\*\* that I am somehow asserting that no peer review is needed for this work. Absolutely no one is asserting this.*

*I am very happy to take feedback and quality assurance. The data is openly available, which is more than almost all published IRAP research to date can say. You are free to inspect this data and tell me if it is in some way lower quality than other work – as part of the peer review process that I have solicited here. It is entirely circular to argue that this work can’t be published because it hasn’t already been published.*

*I am not affirming the consequent by suggesting that unpublished data be included, I am pointing to best practices defined by the world authority organisation on evidence synthesis, the Cochrane foundation, which suggests that (as yet) unpublished data be included.*

*In order to attempt to satisfy the reviewer here, I have included additional details of the data collection procedures associated with each study included in the dataset, and how their design and data collection procedures were identical between already published and as yet unpublished work. See the supplementary materials (p. XX).*

As regards, the fact that the relevant data was derived only from two researcher's labs it would similarly be helpful if the authors acknowledged even the possibility in their discussion section that this data might not be entirely representative of the wider IRAP literature -- not least because it seems that the authors improvised their own, unmentioned methods of quality control (i.e. except indirectly perhaps via their stated inclusion criteria, but this was not explained as such).

*Author response: For better or worse, the fact that this data was collected in two labs actually makes it very representative of the IRAP literature as a whole. This recent preprint (REF) presents analyses of the authorship of the IRAP literature from 2006 to 2022 and suggests that the task’s creator and his students authored 75% of the IRAP literature (REF). My point here is that the concentration of IRAP research within a small number of labs is unfortunate but is a problem that is much larger than the current manuscript. I welcome the acknowledgement of limitations, and have therefore taken the reviewers comments on board by adding a limitations section that acknowledging the source of the data (while comparing this to the literature as a whole with reference to the previously cited preprint). Page XX:*

*XXX*

I see that the proposed manuscript now contains the following new clarification at the beginning of the methods section: "This dataset was constructed by contacting individuals at multiple labs conducting IRAP research and asking them to contribute their well-organized trial level IRAP data, as well as all stimuli and task parameters to generate that data, to an openly available dataset." Does this mean that the updated analysis now contains data from outside of the two labs referred to in the original manuscript? If so, then it would be helpful if the authors clarified in tabular form (even if in an appendix table) how many (and maybe even which) datasets, along with their sample sizes, were respectively produced by each such lab. Indeed, it only seems fair to credit those other labs by naming them, even if the specific names are temporarily redacted during peer review.

*Author note: No, the data comes from two labs. The change you refer to was made based on your question in your previous review about why other labs were not contacted to contribute data. The text now clarifies that other labs were contacted, and some offered data, but data in the necessary format was only provided and entered into the dataset by these two labs. Note that “lab” may have caused confusion here. It has been used in reference to research projects in which the contributing authors were involved and where they had permission to contribute data. Not all data was collected in the universities at which they were affiliated, some was collected by their coauthors at other institutions. In this sense, the data comes from two investigators collaborations, but multiple sites. The supplementary materials have been expanded as you request.*   
  
-- I liked the authors' explanation of prediction/credibility intervals on page 18 of the proposed manuscript (and their distinction from the Bayesian equivalent), however I do not agree with the way in which they have characterized confidence intervals: "...confidence intervals represent a long run probability of the true (i.e., data generating) value,...". As illustrated in the following weblink that I provided in my previous feedback, <https://rpsychologist.com/d3/ci/>, an individual confidence interval is like the roll of a dice rather than an estimate of some true value. Yes, any given 95% confidence interval has a 95% chance of including the supposed true effect in question, however as illustrated by the above interactive weblink each individual confidence interval oscillates wildly around that unknown true effect. It might be useful for me to distinguish here between this meaning of 'statistical precision' in the long run, and 'experimental precision' which is more concerned with the uncertainty/stability of point estimates per se (unlike confidence intervals).

Nonetheless, the proposed articles still persists in using confidence intervals to illustrate in the figures various points about the observed bootstrapped IRAP effect sizes.

*Author response:*

*The reviewer contradicts themselves here. You state that “an individual confidence interval is like the roll of a dice rather than an estimate of some true value.” However, even Kristoffer Magnusson’s site, which you link to (*[*https://rpsychologist.com/d3/ci/*](https://rpsychologist.com/d3/ci/)*) defines “95 % confidence is a confidence that in the long-run 95 % of the CIs will include the population mean.” Statistical precision is the width of the confidence interval. Note: not its location (accuracy), but its width. ….*

*This distinction you make in the use of statistical precision vs experimental precision is not one that is used or accepted elsewhere. You don’t provide references to this. The precision of a 95% Confidence Interval has an accepted meaning, which I employ here. I now include citations to affirm that this common use is my intended meaning and not others. Your comment above about Vahey et al.’s (2015) intentions behind the use of their words must be extended to me too: I am being clear about what I mean by the word precision, it is not necessary for me to adopt your meaning, nor is your meaning widely accepted or even discussed elsewhere.*

Furthermore, it is clearly indicated in Figure 1's caption that it displays confidence intervals, but Figures 2 & 3 leave the reader guessing about this issue. As I will explain in my next point, this issue is currently moot to a more serious issue, but it should be born in mind should the authors of the proposed manuscript choose to take my constructive advice as follows.

*Author response: the manuscript now states on page XX that all error bars in the figures represent 95% Confidence Intervals.*

-- The foundation of the current analysis, the bootstrapped method by which its D-scores are calculated is still fundamentally problematic for the reasons I outlined in the previous round of peer review. I understand that the authors used the same basic approach to calculating their trial-type D scores, except with fewer data points per D score. I would however strongly caution against this practice of using fewer data points than usual. There are already a minimal number of 18 pairs of data points gathered within each IRAP trial-type on any given administration of the IRAP, and reducing this any further seems particularly inadvisable given that small sample sizes inevitably yield very unstable descriptive statistics, including effect sizes. If anything, I would say that the IRAP needs to be improved so that it can gather more data-points per trial-type without fatiguing participants or reducing its experimental precision. Otherwise, current versions of the IRAP will continue to produces scores with statistical and experimental precision that is insufficient for individual interpretation/assessment/diagnosis.

*Author response: We cannot have our cake and eat it too. The flaw that you highlight is a flaw with the IRAP, not the analyses. You wanted trial type level analyses, so I provided them. The reduction in the number of trials per score was an entirely foreseeable consequence of this. What you present as a problem with my analysis is in fact a problem for the IRAP.*

Indeed, as things currently stand each of the IRAP's trial-type D scores already contain substantially fewer trials than comparable latency-based implicit/task-based measures. As such, it would be irresponsible to attempt to interpret and honestly quite misleading to attempt to interpret individual IRAP scores in the proposed manuscript that were generated from even fewer data-points than usual.

*Author response: You are agreeing with the conclusion of the current manuscript (that the IRAP is not suitable for individual use) but presenting it as a reason not to publish the work. There is nothing misleading or irresponsible about the manuscript and I resent the implication of dishonesty. I think it is misleading and irresponsible to attribute flaws with the task to the authors who are trying to highlight these issues. It is the definition of shooting the messenger.*

At the very least, the proposed manuscript should have clarified how many of the 9 pairs of trials per trial-type that it resampled for each round of bootstrapping with replacement. The smaller each of these of bootstrapped subsets, the more unstable and thus broad the proposed manuscript's confidence intervals would have been. As per the references that I previously provided in the previous round of peer review, the problem is only compounded when the relevant data is positively skewed because there is a greater potential for any individual datapoint to skew the relevant sub-sample estimate (i.e. precisely because the data is 'skewed' rather than clustered around a mean/mode, thus violating the central limit theorem). With larger samples of response latency pairs (i.e. 18 or more), the underlying pairs of positively skewed response latency distributions will stabilize pretty quickly and with rapidly diminishing returns on sample size (as per most distributions). Once those pairs of response latency distributions are somewhat stable due to increase sample size we can set about calculating D scores with them, and it is important to note that the resulting D score distribution is not in turn skewed by instead approximately normally distributed even though it is bounded as +/-2 due to statistical standardization (i.e. akin to the scaling performed in producing Z scores which are also bounded for small sample sizes; see  
<https://www.tandfonline.com/doi/abs/10.1080/00031305.1988.10475530#:~:text=The%20magnitude%20of%20Z(n,exist%20for%20small%20data%20sets.)>. I would also point out that the articles cited by the authors of the proposed manuscript, Ratcliff (1993) and Whelan (2008), both refer to the analysis of raw response latencies rather than to standardized differences between pairs of response latencies that are experimentally matched with each other. One might argue that because D scores are usually calculated based upon adjacent pairs of response latencies that are from opposing trials within a given trial-type (i.e. often called consistent versus inconsistent), that any transient experimental confounds are largely (though admittedly not entirely) cancelled out by this subtraction process; certainly these standardized difference scores will in principle much more fully fulfill the 'independent and identically distributed' requirement than shuffled subsamples of raw latencies that haven't been temporally paired. I'm sure that the authors would agree that in some sense everything is connected to everything else in the World in some sense -- what I am highlight here is a matter of degree, and the extant literature is clear that small samples of response latency data produce notoriously unstable descriptive statistics (and this includes difference scores like the D scores).

The authors of the proposed manuscript claim to have resolved any such issues by adding a bias correction and accelerated (BCA) technique to their analysis that was derived from one of the article by Puth et al (2015). However, it should be noted that Puth et al (2015) were very clear in the final paragraph of their discussion section that these sorts of bias correction techniques are only supposed to be used on small samples when those samples are not skewed. As such, it is not an appropriate solution in the present case.  
  
-- The authors of the proposed manuscript also appealed that this was the best available approach and that there were no other ways of calculating bootstrapped estimates of credibility/confidence intervals around individual D scores. Given that the current version of the proposed manuscript focuses upon using group-level statistics to interpret a large range of individual bootstrapped IRAP scores, I would suggest an alternative approach that is in keeping with both the authors aims and with the group-level statistical methods they used once their bootstrapped D scores had been calculated. If the authors' goal is to use multiple individual D scores to make a general point about the statistical and/or experimental precision/stability of trial-type D score IN GENERAL, than why not simply take their existing dataset of non-bootstrapped trial-type D scores and bootstrap those to estimate confidence and prediction/credibility intervals for those stratified per known-group? Each of  
those known-groups is by definition designed to be similar in relevant benchmark/functional/criterion ways to each other, and so bootstrapping within each known-group could in principle provide a preliminary estimate of the statistical and experimental precision of the authors' trial-type IRAP data. In other words, to the extent that the relevant known-groups were carefully sampled according to independent benchmark criteria, then the resulting trial-type D-scores within each known group should be proportionately similar/stable to each other. This would be a commentary not just on the precision of trial-type D scores alone per se, but also upon the precision of any associated validity criteria such as the known-groups procedures used within the IRAP literature. Much like what the proposed manuscript is attempting to do, this would allow one to tentatively quantify the statistical and experimental precision of individual IRAP trial-type scores IN GENERAL. If one could do this  
then one could legitimately comment upon whether or not current version of the trial-type IRAP scores are GENERALLY statistically and/or experimentally precise enough to interpret on an individual basis.  
  
-- As an aside, it is my understanding that the distinction between IV and DV that the authors have made does not necessarily apply to the IRAP literature because if one is coming from a functional contextualistic perspective any given trial-type D score is supposed to be capturing the same behavior/verbal function as the criterion/benchmark variable in each case (i.e. rather than predicting some other independent phenomenon as in the IV versus DV distinction). From the perspective of functional contextualistic IRAP researchers the trial-type D score might differ topographically from some corresponding criterion variable but the aim is to design the IRAP so that it is functionally the same.

*Author response: This point is offered without reference any supporting literature (e.g., to work on functional contextual perspectives on statistics or psychometrics, of which I’m not aware of any), and without reference to how it relates to the current manuscript. As such, I’m not sure action the reviewer would like me to take. The analyses presented in the current manuscript adopt the same treatment of IVs and DVs that are employed nearly ubiquitously elsewhere. E.g. the IRAP effects estimate the impact of IRAP block (IV) on reaction time (DV) and not the other way around. It is not plausible that reaction times estimate which IRAP block one is being presented with.*