Manuscript Number: JCBS-D-22-00243     
  
The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis  
  
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 your manuscript to JCBS and for your patience with the peer-review process. Although I have not used the IRAP, I have done quite a bit of research using implicit measures and am familiar both with their potential to provide predictive information that cannot be obtained via self-report as well as the large amount of measurement variance that can often accompany such indirect assessments. I therefore see value in in work that carefully examines the quality of information provided by implicit measures, including work (like the current manuscript) that quantifies key methodological issues that might underly such forms of assessment. I have now obtained comments from three researchers and am happy to report that they all saw potential value in the work presented. However, the reviewers were split on their recommendations for the manuscript.  
  
Given these reviews, I am recommending a decision of revise and resubmit with major revisions. I feel that all three reviewers offered sets of reasonable suggestions and comments. I would strongly encourage you to make revisions addressing each of the comments/suggestions raised. I would also ask that you clearly detail the changes made to address each and every comment in your response letter, taking extra care to provide clear justification for any suggestions you ultimately do not incorporate. Given the depth of changes being requested, I cannot guarantee the ultimate acceptance of this manuscript. Having said that, you can increase the odds of acceptance by being more responsive to the comments and suggestions across all three reviewers and by clearly detailing those changes in your response letter.  
  
Thank you again for your patience. I look forward to seeing your revised manuscript.  
  
Sincerely,  
  
Ron Rogge  
Associate Editor, JCBS  
  
  
REVIEWER 1 COMMENTS  
Dear editor,  
  
Thank you for the opportunity to review the current manuscript which reports some interesting findings regarding the use of the IRAP, primarily it's use psychometrically. Given that I myself often use the IRAP in my own research (although not as a psychometric instrument), I found the topic to be quite interesting. I believe it is important that the issues raised within the manuscript are discussed and debated so that those interested in using the measure are aware of its potential benefits and pitfalls, and of nuances with variables/parameters (such as those that the current authors have targeted and indeed others) that may be critical to consider when deciding whether or not to employ the measure.  
  
However, while I believe that the current manuscript certainly has potential for publication in JCBS on that basis, I do not believe that it is suitable in its current form. That is, while the psychometric issues raised are important, it is my view that more well considered arguments and a greater deal of scholarship would be required to ensure the value of the manuscript in the literature. As I will outline in more detail below, I believe that more is needed both empirically and conceptually to make the manuscript as useful as possible for the journal's readership. In particular, at an empirical level I have serious concerns about the extent to which the current analyses reflect how the measure is actually used by researchers. At a broader conceptual level, it came as a surprise to me to find that the manuscript makes no attempt to make links to the wider conceptual literature on relational frame theory (RFT) and contextual behavioural science (CBS) relevant to the  
instrument. Indeed, it even seems that no IRAP literature (experimental or conceptual) is mentioned beyond 2018, 4 years ago. There is much relevant published work using the measure available in the literature since that time. This again seems to limit the utility of the manuscript in its current form and raise questions about its current relevance. Given not only that JCBS is a behavioural journal, but also given the conceptual foundations of the IRAP itself (RFT) and the relevance of those foundations within contextual behavioural science (CBS) more generally, I firmly believe the arguments presented within the current manuscript need to connect with this wider behavioural literature. In this vein, it also seems crucial that connection is made with the ways in which the measure is actually being used within CBS and behaviour analysis more generally (e.g., use of the IRAP as a measure of behavioural dynamics; analysis of individual trial-type effects rather than relying on  
the overall D-score). I believe that incorporating such changes and considerations will serve to produce a more well considered and indeed valuable contribution to the literature.  
  
EMPIRICAL CONSIDERATIONS AND CONCERNS  
  
The first concern that I have with regards the current empirical analysis is the extent to which it actually represents and is relevant to how IRAP data is typically analysed and reported. Specifically, it appears that the primary analyses are conducted with respect to the overall D-score produced by the IRAPs involved (an overall score collapsing data from all 4 trial-types). Although the overall D-score is sometimes (but not always) reported, it rarely if ever is the primary score used in analyses with the measure (I am not aware of any study that relies on this score as the primary datum for analysis). Rather, work with the IRAP typically involves reporting and analysing each individual trial-type or effects across two trial-types. As such, the utility of conducting the current analyses on, and drawing conclusions from, the overall D-score seems extremely limited - because it does not speak to how people actually analyse data using the measure. It seems, therefore, that  
it may be far more useful to conduct reliability and validity analyses across individual trial-types rather than solely on the overall D-score. Indeed, it seems that the authors have a golden opportunity here - reporting analyses that can inform researchers based on how the IRAP is actually used and how its data are routinely reported. For example, does one trial-type report better reliability than others? Is one trial-type relative to others more reliable within a given domain?  
  
Another related empirical issue that was not considered in the current analysis and that from the IRAP literature seems to highlight a potentially a critical variable is the performance criteria employed in the studies used included in the current analyses. Specifically, were participants required to meet IRAP accuracy and latency criteria at the block level or rather at the trial-type level? Research has suggested that requiring participants meet criteria at the trial-type rather than block level may be critical in acquiring desirable levels of stimulus control within the IRAP (e.g., Leech et al., 2018, 2020), in behaviour-analysis anyway. Consider, for example, a participant that maintains criteria on only 2 or 3 trial-types but still passes the IRAP based on accuracy and latency criteria being applied at the level of the overall block. From a behavioural point of view, this suggests a lack of adequate stimulus control over the behaviour being assessed. And indeed, if one does not have adequate stimulus control, the variables involved in controlling performance on the measure are also increasingly unknown. In such a case, by definition, reliability is going to be lost -- too many variables are free to vary.  
  
Of course, considering such a point is one of the reasons why an article such as this one has the potential to be of considerable value - it can inform people as to what ways (e.g., employing what criteria) that the IRAP may be relatively more or less reliable. But such analyses are not conducted currently and it seems to me that the criteria may have been solely employed at the block level. Of course, it may be the case that this variable does not change the overall reliability of the measure relative to employing the block level criteria. But as the manuscript currently stands we do not know.

*Author response: Yes, IRAPs were delivered in a standard fashion and had to meet accuracy and latency criteria. Only participants who passed the practice blocks were included in the analyses. Additionally, participants were excluded if they failed to maintain performance in the test blocks.* XXX  
  
Relatedly, I have some further concerns regarding the inclusion/exclusion criteria employed by the authors. They state that "Participants were also excluded if their mean reaction times in the IRAP test blocks were ± 2 Median Absolute Deviations from the median, in order to exclude implausibly fast or inappropriately slow responding." This is routine and good to see. However, for most IRAP studies, participants data are removed if they fail to maintain these criteria throughout the test blocks (see above issue related to adequate stimulus control for why). Were participants that did not maintain criteria throughout the test blocks included in the current analyses or removed? If not then why not? I cannot see information pertaining to this point in the current manuscript. If it is the case that I have missed it then I believe that it needs to be made clearer in a revised version. If these exclusions have not been made, it further calls the utility of the manuscript in its  
current form into question as it does not speak to how most researchers actually employ and analyse data gathered from the instrument. I believe that it would be extremely useful going forward for such exclusions to be made and analyses reconducted accordingly on this basis, and with regards the trial-type versus overall D-score point made above.  
  
In a similar vein, it seems that some of the IRAP data analysed employed latency criteria at 2000ms while others employed 3000ms. Did this make a difference? And again, both when applied at the block versus individual trial-type level? Indeed, I am aware of one study already in the literature that assessed IRAP reliability when applying criteria from 3000 to 2000ms that suggests that this variable makes a difference in performance on the measure (Barnes-Holmes et al., 2010). How do the current findings link to that work? Variables such as these (e.g., latency criteria employed; whether performance was required at the block or individual trial-type level) already highlighted in the literature as relevant need to be systematically explored in order for the manuscript to provide useful advice beyond the "just don't use the IRAP" advice that the authors currently provide. I believe that such further analyses (comparing reliability and validity when criteria are applied at the  
block vs TT level; conducting analyses across individual trial-types as opposed to solely the overall D-score) are critical for the current manuscript going forward. If not, there seem to be too many potentially important variables left unknown rendering the analyses in their current form, and by extension their conclusions, somewhat loose. As such, they are at risk of falling short of the necessary level of scholarship required for an article to be make a valuable contribution to the literature.

*Author response: XXX*

Relatedly, it is not clear what all other varying criteria were or were not employed on the multiple IRAP studies that generated data included in the current analyses. The authors provide a link for readers to check the data themselves on an open sources pre-registration website which is of course useful for a reader who wishes to sift through the data in greater detail. However, the purpose of an article is to inform and assist your readership. As such, in addition to the link to the detailed material provided, I believe that the authors should provide a summary of some of the necessary information within the body of the article too. For example, what were the parameters of the task employed? How many required criteria at the block versus individual trial-type level? How many of the studies were already published versus unpublished? How many employed 2000 vs 3000ms response latency criteria? How many employed specific versus general instructions (published research suggests  
this makes a difference to IRAP performance; Finn et al., 2016). Is there any information available as to the amount of previous exposure to the IRAP participants had (published research has found that this variable impacts upon the magnitude of the effects produced on the measure; Finn et al., 2018). These are all important parameters that may vary and in turn help speak to a lack of reliability and validity in a meta-analysis such as this. Do the authors think such wide variability has implications for their findings and conclusions?

*Author response: The reviewer argues that many of the factors they mention influence “IRAP performance”, but this is very ambiguous. There is no evidence to date that any of these factors influence the IRAP’s reliability, only mean scores on the task. The reviewer also proposes that these factors should be investigated as moderators, rather than variation in them be a source of heterogeneity in how reserachers actually use the task. XXX*  
  
Returning again to the above point about scholarship, it is also not appropriate to hinge an argument off of a paper that has not gone through the peer review process and is therefore not published within the academic literature. Specifically, "Hussey (2020)" is mentioned at a number of points throughout the article as evidence that the IRAP should not be employed for individual participant use. Upon closer inspection of this article in the References section, however, it becomes clear that this manuscript is an unpublished pre-print on PsyArXiv. This is not an appropriate source for the current article and should be removed, along with the arguments it is used to support, unless such points can be sufficiently argued with other reliable sources in the literature.

*Author response: It is not unscholarly to reference preprints, following both APA standards and Cochrane’s recommendation that the grey literature be included (REF). The reference is accurately labelled as a preprint, the reader is in no way deceived about its status.*  
  
Finally, the authors come to the conclusion that the IRAP performs "unacceptably low" on a number of psychometric criteria. To maximise the utility of this conclusion for those who wish to employ a reaction-time-based/implicit measure, it seems important to provide other potential alternatives and thus, comparisons to alternatives. While there are a couple of brief passing comparisons made to the IAT, I do not think this is well considered enough. How well does the IRAP perform in comparison to the IAT and other prominent implicit measures such as evaluative priming, the AMP, etc.? In other words, should researchers use an IAT instead, for example? What other implicit/reaction-time-based measures actually do meet these criteria? In addition, consideration in conducting these analyses would have to be given to the differing task structures of these measures. Given the different structure of, for example, the IRAP and the IAT (the IRAP has 4 separate trial-types in any give block, the IAT only has 2), calculating, for example, test-retest reliability in the same way for both measures does not seem to be an appropriate comparison. Comparisons taking these differences into account and/or to other implicit measures that produce 4 separate combinations as the IRAP does seems like they would be far more useful (i.e., comparing analogous measures).  
  
Such analyses and comparisons would not only provide a more balanced argument overall, but also suggest a more constructive and useful way forward. The critiques raised currently with regards internal validity and test-retest reliability would of course still stand but how do other measures compare? What could be used instead and in what contexts? What are the advantages and disadvantages? It is my view that incorporating such considerations would considerably increase the utility of the manuscript as it could then direct interested researchers where to go if the IRAP is not suitable for their needs, or if it is how best to employ it.  
  
Overall, with respect to empirical considerations going forward, I would like to see more information and analyses pertaining to the potential role of different parameters on test-retest reliability and internal consistency within the IRAP. Currently, the studies involved in the analyses seem to differ in terms of a number of variables. For example, variation seems to be evident in some of the criteria employed, no information is provided about whether criteria were maintained at the block or individual trial-type level, nor whether criteria was maintained throughout test blocks. In addition, no information is provided as to whether such criteria make a difference to the analyses presented, particularly when considered in the context of stimulus control. These analyses seem particularly important given the authors broad unfavourable conclusions for use of the instrument in whole cloth. I would also like to see an analysis of how the measure compares to other similar measures  
in the literature. Such additions will surely increase the utility of the current manuscript to the readership and provide some direction for researchers going forward.  
  
CONCPEPTUAL  
  
Another concern regarding how the arguments in the manuscript are currently presented pertain to the distinct lack of conceptual considerations made. The development of the IRAP was conceptually rooted firmly within RFT as a way to assess natural verbal relations (Barnes-Holmes et al., 2008). It is not totally clear to me if and how the current critiques apply to the IRAP when used in this way (as opposed to as a psychometric or more mainstream oriented instrument; a recent article by Barnes-Holmes and Harte, 2022, comes to mind that may be useful for the authors when considering this point going forward). It thus seems important to link the current arguments to its theoretical roots and consider whether the psychometric-based arguments and analyses presented currently have any implications for its use in this way. This point has a number of facets that I will attempt to articulate below.  
  
First, it came as a surprise to me that the current article did not make any mention of, or attempt to link to, conceptual developments that have emerged over the last 4 or 5 years within IRAP research, and indeed within RFT more generally. Specifically, the differential arbitrarily applicable relational responding effects (DAARRE; Finn et al., 2018) model of IRAP effects and the hyper-dimensional, multi-level (HDML; Barnes-Holmes et al., 2020) framework for considering dynamic patterns of relational responding seem particularly relevant. How do the current arguments connect with the DAARRE model (as a means of interpreting patterns of effects obtained on the IRAP) and with empirical work that has provided support for this model and the variables it highlights as important (e.g., Finn et al., 2019; Kavanagh et al., 2019; Pinto et al., 2020; Schmidt et al., 2021; the authors might also find it useful to consider the doctoral thesis of Martin Finn in refining their arguments  
going forward)?  
  
Furthermore, connecting to broader conceptual developments within RFT also seem important. Indeed, the first article describing what later became the HDML framework for analysing the dynamics of relational responding (which the IRAP conceptually purports to assess) was published in this very journal (Barnes-Holmes et al., 2017). It is thus surprising that the current article does not attempt to make contact to these relevant developments in the literature and it is my view that such contact would be needed (and indeed, extremely valuable) going forward. Can considering IRAP performances along the levels and dimensions of the HDML help make conceptual sense of the issues raised in terms of validity and reliability? Do the current critiques apply to work that uses the DAARRE model to guide use of the IRAP as a means of analysing the dynamics of responding involved on the IRAP? Of course, this is by no means to say that the authors have to come out in favour of this  
perspective. But rather that on the grounds of decent scholarship, connecting with a far wider array of directly relevant conceptual and empirical analyses seems necessary and valuable.  
  
At a more general level, it seems to me that a more adequate and complete argument is needed for why test-retest reliability and internal consistency as measures are deemed themselves critically important. That is, assuming that these measures of reliability and validity are crucial in assessing the utility of the IRAP or any measure is predicated on a psychometric assumption -- that the instrument they are assessing is reflecting a constant construct. Indeed, why else would one expect that performance on a measure would remain consistent across time? If that assumption is not made when employing the IRAP but rather the assumption that the IRAP captures a "snapshot" of behaviour at a certain point in time, the issues to do with its reliability and (construct) validity seem to fall away to a degree. In other words, if the measure is not assumed to be a proxy of a constant construct, are issues with poor internal consistency and test-retest reliability less of a problem? This  
seems particularly important to consider and defend within an article to be published in a journal of contextual behavioural science and given that much work using the IRAP employs and interprets the measure in this way (a measure of dynamic relational responding as defined within RFT). More informally, yes, it very well may be a bad measure of a constant construct, but why is that relevant to work that uses it as a measure of relational responding at a certain point (or points) in time? Within the latter conceptualisation, if you ask a participant to do something over and over again (e.g., respond on one or more IRAPs, or indeed across blocks of trials within a single IRAP), the behaviour observed at time 1 is fundamentally different from the behaviour observed at time 2 - they are not the same behaviour by definition (e.g., because it is a different point in time and with a history of responding on the IRAP). Thus, don't low levels of reliability conceptually make sense?  
  
Of course, I am not necessarily saying that test-retest reliability and internal consistency should not be considered important. But given that the claims made seem to be largely founded upon a construct-based assumption, it is my view that this needs to be sufficiently justified early on and argued within the different uses of the IRAP (as a measure of an implicit construct or of the dynamics of derived relational responding as defined within RFT). This seems particularly necessary within the context of a behaviour science journal and a measure that in many cases is used to assess the behavioral dynamics involved in natural verbal relations. On balance, if the current arguments are being made primarily for the use of the IRAP as an implicit mainstream measure alone, then the author's criticisms do certainly seem to carry more weight. But if that is the case they are making then it would need to be made far clearer early on that the current critiques pertain to its use in  
that way. As such, much of the conclusions made as applying to use of the measure in whole cloth would need to be substantially revised and reframed. Revisions would also need to be made to the article title on this basis.  
  
Overall, therefore, a more sophisticated discussion of how the current findings connect to the wider conceptual literature on use of the IRAP within behaviour science seem needed. Clarification also seems needed for how these critiques apply to different uses of the IRAP (implicit measure vs dynamics of relational responding), as well as a stronger argument for the importance of internal consistency and test-retest reliability in these contexts. These considerations seem particularly important when considering the current journal and what would be useful for its readership. Otherwise the utility of this piece is, in my view, rather limited.  
  
TWO FINAL MINOR POINTS  
  
Page 7 - "Participants typically complete between one and three pairs of test blocks until they meet performance criteria (e.g., median reaction time < 2000 ms and percentage accuracy > 80%), followed by three pairs of test blocks from which scores are calculated." It seems to be that the first mention of "test blocks" in this sentence should actually read "practice blocks."  
  
Page 10 - "As illustrated in Figure 1, results indicated bimodality in both estimates of effect size and heterogeneity that was driven by data from a single domain: sexuality (Sexuality IRAP 1: α = .84, 95% CI [.65, .93], Sexuality IRAP 2: α = .94, 95% CI [.84, .98]), suggesting that it represented an outlier that biased the results." This is an interesting finding that is never considered. Why may it have been the case that this domain in particular performed differently than others?  
  
CONCLUSION  
  
To conclude, I think a revised version of the current manuscript must engage with a number of empirical and conceptual issues in the service of increasing the utility of the piece and the arguments it puts forward. Namely:  
  
\* Additional analyses should be conducted reflecting more closely how the measure is actually used. For example, analysing reliability across individual trial-types as well as using (but not relying solely upon) the overall-D-score. Trial-type analyses are the primary analyses conducted in published IRAP work. Relatedly, information should be provided as to whether participants were excluded if they did not maintain adequate accuracy and latency criteria across the test blocks. If participants were not excluded on this basis further analyses should be conducted with these exclusions made.  
  
\* Additional analyses should be conducted to assess how internal reliability differs for participants who maintain performance criteria at the trial-type level versus the block level, at 2000ms and 3000ms response latency. Additional comparisons should be made with other reaction time based measures to provide useful recommendations as to which measure may be most useful along the criteria the authors have deemed important (internal consistency and test-retest reliability). It is of course possible that such additional analyses suggested above will reveal similarly poor outcomes. But based on the analyses presented currently we cannot know.  
  
\* Regardless of whether such analyses make a difference to levels of reliability and internal consistency, connection and consideration should be made to the wider relevant conceptual and empirical literature in the area (e.g., the DAARRE model). How do these issues fair when considering the instrument through an RFT lens (as a measure of dynamic and changing behaviour)? Does considering responding on the measure in this way help explain the poor reliability and internal consistency?  
  
\* Clarification of the construct argument seems needed as it remains ill defined. If what is being measured is performance on the IRAP itself then it seems like it may be important to look at the impact of other variables when employing the measure (e.g., dimensions highlighted within the HDML) to guide research and explaining the effects obtained. On the other hand, if the IRAP is being used as a proxy for something else, the issues with reliability and validity should definitely be heeded. There is no consideration for these nuances currently.  
  
\* The argument against single participant use should be removed as it hinges on an unpublished pre-print. Unless other more reputable sources can be provided to argue this point. If the latter is the case, how do these arguments connect with the conceptual developments mentioned above and with work that has employed the IRAP in a single participant context (e.g., the Finn, 2020, doctoral work).  
  
I would really like to see the authors engage with and pursue the suggestions and comments I have outlined in this review. The article is interesting and presents considerable potential for a meaningful contribution to the literature if sufficiently revised. However, in its current form it does not, in my view, reflect or connect with how the IRAP is commonly used nor how its data are routinely presented. The richness of the current article could be easily developed if it considers the above comments. I should be clear, however, that in suggesting the authors address my concerns I believe it is important that they do so by fully engaging with these materials as opposed to simply addressing each with a line or two the Discussion. Furthermore, and as mentioned above, considering such points does not mean the authors must agree with these perspectives. However, on the grounds of scholarship and providing an article of value to the readership of JCBS and researchers interested  
in the IRAP, connecting with a far wider array of directly relevant conceptual and empirical analyses seems necessary (for the authors convenience, I have also included the references that I mentioned throughout at the end of this review). I believe that a revised version of the manuscript that does so could be an extremely valuable addition to the literature.  
  
References  
Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., & McEnteggart, C. (2017). From IRAP and REC model to a multi-dimensional multi-level framework for analysing the dynamics of arbitrarily applicable relational responding. Journal of Contextual Behavioral Science, 6(4), 473-483. [https://doi.org/10.1016/j.jcbs.2017.08.001  
Barnes-Holmes](https://doi.org/10.1016/j.jcbs.2017.08.001Barnes-Holmes), D., Barnes-Holmes, Y., & McEnteggart, C. (2020). Updating RFT (more field than frame) and its implications for process-based therapy. The Psychological Record, 70, 605-624. [https://doi.org/10.1007/s40732-019-00372-3  
Barnes-Holmes](https://doi.org/10.1007/s40732-019-00372-3Barnes-Holmes), D. & Harte, C. (2022b). The IRAP as a measure of implicit cognition: A case of Frankenstein's monster. Perspectives on Behaviour Science, 45, 559-578. [https://doi.org/10.1007/s40614-022-00352-z  
Barnes-Holmes](https://doi.org/10.1007/s40614-022-00352-zBarnes-Holmes), D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. The Psychological Record, 58, 497-516. [https://doi.org/10.1007/BF03395634  
Barnes-Holmes](https://doi.org/10.1007/BF03395634Barnes-Holmes), D., Murphy, A., Barnes-Holmes, Y., & Stewart, I. (2010). The implicit relational assessment procedure: Exploring the impact of private versus public contexts and the response latency criterion on pro-white and anti-black stereotyping among white Irish individuals. The Psychological Record, 60, 57-79. <https://doi.org/10.1007/BF03395694>  
Finn, M. (2020). Exploring the dynamics of arbitrarily applicable relational responding with the implicit relational assessment procedure [Unpublished doctoral dissertation, Ghent University].  
Finn, M., Barnes-Holmes, D., Hussey, I., & Graddy, J. (2016). Exploring the behavioral dynamics of the implicit relational assessment procedure: The impact of three types of introductory rules. The Psychological Record, 66(2), 309-321. <https://doi.org/10.1007/s40732-016-0173-4>  
Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect on the IRAP: Developing a differential arbitrarily applicable relational responding effects (DAARRE) model. The Psychological Record, 68(1), 11-25. <https://doi.org/10.1007/s40732-017-0262-z>  
Finn, M., Barnes-Holmes, D., McEnteggart, C., & Kavanagh, D. (2019). Predicting and influencing the single trial-type dominance effect. The Psychological Record, 69(3), 425-435. <https://doi.org/10.1007/s40732-019-00347-4>  
Kavanagh, D., Matthyssen, N., Barnes-Holmes, Y., Barnes-Holmes, D., McEnteggart, C., Vastano, R. (2019). Exploring the use of pictures of self and other in the IRAP: Reflecting upon the emergence of differential trial type effects. International Journal of Psychology and Psychological Therapy, 19(3), 323-336.  
Leech, A. & Barnes-Holmes, D. (2020). Training and testing for a transformation of fear and avoidance functions via combinatorial entailment using the implicit relational assessment procedure (IRAP): Further exploratory analyses. Behavioral Processes, 172, 104027. <https://doi.org/10.1016/j.beproc.2019.104027>  
Leech, A., Bouyrden, J., Bruijsten, N., Barnes-Holmes, D., & McEnteggart, C. (2018). Training and testing for a transformation of fear and avoidance functions via combinatorial entailment using the implicit relational assessment procedure (IRAP): The first study. Behavioral Processes, 157, 24-35. <http://doi.org/10.1016/j.beproc.2018.08.012>  
Pinto, J. A. R., de Almeida, R. V., & Bortoloti, R. (2020). The stimulus' orienting function may play an important role in IRAP performance: Supportive evidence from an eye-tracking study of brands. The Psychological Record, 70, 257-266. <http://doi.org/10.1007/s40732-020-00378-2>  
Schmidt, M., de Rose, J.C, & Bortoloti, R. (2021). Relating, orienting and evoking functions in an IRAP study involving emotional pictographs (emojis) used in electronic messages. Journal of Contextual Behavioral Science, 21, 80-87. <https://doi.org/10.1016/j.jcbs.2021.06.005>  
  
REVIEWER 2 COMMENTS  
Thank you for the opportunity to review this paper. The submission provides a precise, surgical review of the internal consistency of the Implicit Relational Assessment Procedure (IRAP), drawing upon published and unpublished data to demonstrate that the IRAP - certainly in its commonly-used format - is an unreliable, inconsistent measure.  
  
The author(s) demonstrate a thoroughly transparent and diligent approach to the work: the data used for the analyses is openly available to interested parties; various 'traditional' and newer statistical approaches to the assessment of reliability are used to triangulate and scrutinise the findings; and various reformulations of IRAP trial data are explored in an attempt to obtain more favourable benefit-of-the-doubt reliability estimates.  
This work is crucial and should make a significant positive impact on the area: tools that are not fit for purpose should become obscure, irrespective of their pedigree; and theoretical and practical progress cannot be made if the tools we use produce unreliable or meaningless data. There is also a sound economic argument: researchers, clinicians and funders should not be expending significant resources on something that has now been shown (in the current work and two independent meta-analyses) to be unfit for purpose.  
  
There will, however, likely be some pushback from other researchers regarding this work, and it's important to demonstrate that - as a reviewer - I am aware of some of these arguments. The prime argument is likely to be that, as behaviourists, we are not really interested in psychometrics, classical test theory, etc. I have some sympathy with this position; however, it is often the case that, despite psychometrics not having a natural home in behaviourism, they are often reported in behavioural research - and of most relevance here - in most published IRAP studies. We can't have it both ways; if we report such data, then data of a similar kind can be collated and used to scrutinise the work we do. Secondly, as the author(s) note: 'even the animal-behaviorist working with rats in Skinner boxes must be concerned with whether the lever functions well to capture the animal's lever-pressing behavior'. Third, an argument could be made that the IRAP may have functional use,  
irrespective of its psychometric properties; while that may be the case in some instances (for example, the IRAP has shown some relationships with observable, external behaviours), given the reliability issues outlined here and elsewhere, and potential publication bias, it is difficult to speculate what those functional uses may be if the fundamentals of the tool are questionable at best. Finally, even if arguments that encompass the above (and others) can legitimately be made - the authors have been very explicit about their data, methods of analysis, and have made those all publicly available: so any counter argument or counter claim against this work should be made in the literature, in good open science tradition, using equally open and transparent methods.  
  
In summary, I think the work is important, relevant for CBS, and should be accepted for publication in JCBS. I have made some additional minor comments below that the authors could consider prior to publication:  
  
Page 2: The IRAP claims to capture (rather than does capture) strength of associations between concepts.  
Page 3: 'Just one study was found that reported test-retest reliability...' - it would be useful to provide a reference here.  
Page 4: Full stop missing at the end of the page.  
Page 5: It might be useful here to provide mini-biographies of the author(s) to demonstrate their expertise as IRAP researchers and to provide weight to the analyses presented.  
Page 7: 'Participants typically complete between one and three pairs of test blocks until…' - are these not practice blocks?  
Page 19: Although not core to the foci of this paper, it would have been interesting if the author(s) had provided data regarding the differential impact of allowing switching response options on subsequent completion/drop-out rate.  
Page 21: Typo? control over responding within responding.\  
  
REVIEWER 3 COMMENTS  
Thank you for the opportunity to review the proposed manuscript titled "The Implicit Relational Assessment Procedure demonstrates poor internal consistency and test-retest reliability: A meta-analysis". This manuscript provides new data regarding the internal and test-retest reliability of the IRAP, and illustrates a useful method of calculating the maximum correlation observable with any given IRAP effect size given the reliability of that effect size. Crucially, this information has the potential to encourage IRAP researchers to focus upon systematically and incrementally improving the reliability of the IRAP -- something which has been ostensibly lacking in the IRAP literature to date. Similarly, the proposed manuscript also very clearly highlights how impractical it would be to improve the IRAP's reliability purely by increasing the number of trials it includes. Indeed, during the discussion section of the proposed research, the authors tentatively explore some other  
potential ways in which IRAP researchers might research how to improve the reliability of the IRAP. However, none of these points are made, or even implied, in the proposed manuscript's abstract; rather they are presented in the proposed manuscripts as afterthoughts and given relatively little emphasis. This in my opinion is a missed opportunity as regards providing pragmatic and proactive recommendations to the readership of the Journal of Contextual Behavioral Science.

*Author response:* Thank you for this suggestion. This omission was mostly due to space limits. I’ve reworked the abstract to point to these issues being discussed in the paper. It now reads:

XXX  
  
The main headline emphasis of the proposed manuscript was not upon how to improve future implementations of the IRAP, but rather upon emphasizing how unusually poor the IRAP's reliability is. The proposed manuscript's abstract for example concludes rather fatalistically: "...that researchers should be very cautious about choosing to employ the IRAP or when interpreting its results". I would seriously question how useful this emphasis is to the readership of the Journal of Contextual Behavioral Science. Indeed, I would urge the authors to interpret the statistics that they report less emphatically than they have. While many of the concerns that the authors raise about the need to systematically improve the reliability of the IRAP are well founded, they frequently mischaracterize the implications of their two labs' findings for the wider IRAP literature.

*Author response: Regarding the tone of the conclusions, I encourage the reviewer to read the recommendations in the spirit they were written. The manuscript does not say the IRAP cannot be improved or should not be used etc, it urges caution in the choice to use the task and interpret the results of existing work. This is not nearly as fatalistic as the reviewer implies; science is a cautious enterprise! To urge less than caution in light of these poor results would be to misrepresent the necessary implications of the results.*

*The reviewer raises an excellent point regarding generalizability, which I have expanded on. Supplementary materials discuss the labs data was collected in. References to the preprint showing that this breadth of data collection is typical for the majority of the rest of the IRAP literature.*  
  
In the proposed manuscript's introduction much effort is made to demonstrate that the IRAP performs unusually badly relative to comparable measures; and thus, should be avoided whenever possible. As I will unpack below, I do not believe that such fatalistic conclusions are justified by the facts presented in the proposed manuscript. The proposed manuscript's headline conclusions mischaracterize the IRAP literature in ways that are likely to create confusion and controversy rather than foster productive collaboration. I would therefore urge that authors consider changing the emphasis of their proposed manuscript to focus more upon the points they have made in relation the need for IRAP researchers to focus upon conducting research aimed at systematically and incrementally improving the reliability of the IRAP.

*Author response: This risks imposing a singular viewpoint on strategy and goals that readers may not share. Not all research must conclude “further research is necessary”. Philosophy of science has spent much time talking about the need for mechanisms of discarding poor ideas and creations. Just because the IRAP exists does not require that we improve it – it’s a big world out there, and researchers could do other things. I have spent a decade trying to improve the IRAP and generally found no way forward, and am choosing to divert my attention elsewhere. It is scientifically honest to report this honestly. It is not anti-scientific to be pessimistic in light of data, nor is it appropriate to require all authors to be optimistic in all cases.*   
  
-- In the second paragraph of the introduction the authors almost premised their paper on the following erroneous claims. Firstly, they claimed that in their meta-analysis "Vahey et al (2015) argued that the IRAP has potential as a tool for clinical assessment", before then juxtaposing this with the apparently contrary claim that two other meta-analyses had concluded that the IRAP has "low reliability" (Golijani-Moghaddam et al, [2013] and Greenwald & Lai, 2020), and that a third by Hussey et al (2020) concluded that it had "poor individual-level estimation" (the latter point being repeated at the end of the discussion section). The proposed manuscript directly goes on to make the classic argument that a measures cannot be valid if they unreliably; thus implying that there is a clear consensus that Vahey et al's findings are untenable. This line of argument has multiple embedded problems. Firstly, it is a highly misleading straw man argument to claim that Vahey et al (2015)  
recommended the IRAP for use in clinical assessment. Rather Vahey et al instead concluded in their discussion that: "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.

The second major flaw in the argument set out at the beginning of this paragraph is that neither Golijani-Moghaddam et al (2013) and Greenwald & Lai (2020) concluded that the IRAP had low reliability. These two meta-analyses reported internal reliability statistics for the IRAP that were comparable with other measures of implicit cognition: 0.60 in Greenwald & Lai (2020), and 0.65 in Golijani-Moghaddam et al (2013).

*Author response:*

Thirdly, the proposed manuscript cites Hussey et al.'s (2020) meta-analysis as having equivalent evidential weight as the other aforementioned meta-analyses even though it hasn't been published in a peer reviewed journal. I have read the relevant preprint and find it problematic in many of the same ways the proposed manuscript is problematic. In particular, it attempts to make the specious argument that IRAP researchers are currently trying to use the IRAP to perform individual-level analyses that amount to clinical assessments/diagnoses. To the contrary, the focus of the IRAP literature to date has been upon refining group-level analyses in the hope that it might one day be possible to improve the measure sufficiently to perform individual level analyses. Granted, the IRAP does seem to have somewhat stalled in its development, with little methodological innovation over the past few years; but it is highly misleading to claim that the IRAP is being used to diagnose individuals or to quantify the magnitude of their individual differences. Granted the IRAP does not meet the customary internal reliability of .80 required of self-report measures designed to quantify the magnitude of individual differences. However, it is hardly true that an internal reliability of 0.6-0.65 is poor in some absolute and irretrievable sense.

*Author response [to editor]:*   
  
-- There is very little IRAP data testing its test-retest reliability, and none of it was gathered in research that was specifically designed to maximize test re-test reliability (e.g. by carefully matching the baseline and follow-up testing contexts to at least in principle ensure a stable criterion behaviour that the IRAP was supposed to be capturing).

*Author response: This is not the case. The test-retest data included in the studies included in the current manuscript set out to examine this exact property. The reviewer is asserting things they could not know about the studies, and in doing so demonstrate their bias against this work.*

Even so, according to Greenwald & Lai's (2020) results the IRAP has exhibited very similar test retest reliability to other so called implicit measures. On page 3 of the proposed manuscript the authors chose not to report Greenwald and Lai's findings but instead used their own analytic techniques to recalculate slightly lower reliability statistics for the IRAP while presenting them as if they were Greenwald & Lai's findings.

*Author response: Greenwald and Lai’s findings are based on 74 individuals in three studies, and have a confidence interval that range from XX to XX.*

*Include joint analysis of both my and Greenwald & Lai’s data?*

Putting aside the misleading manner in which this was done, at best in doing so the proposed manuscript raises doubts about Greenwald & Lai's findings. It certainly doesn't establish that the IRAP has 'poor' reliability in some absolute sense, given that the proposed manuscripts substituted figures are still within the range of comparable implicit measures.

*Author response: Yes it does. Reliability is not to be compared only with other similar measures, it is a measurement property that defines the proportion of variance that is attributable to error, it has implications for statistical power, and implications for individual level precision of estimation, among other things. These are explicitly discussed within the manuscript. Here, the reviewer mischaracterises the current articles broad claims. It does seek to compare the IRAP only to other implicit measures, but assess the magnitude of its reliability in terms of those implications for the aforementioned related statistical properties. Implicit measures have been widely discussed and acknowledged as let down by their their poor test-retest (e.g., by Greenwald and Lai, 2020). Furthermore, the IRAP’s test-retest is explicitly not “within the range of comparable implicit measures”: confidence intervals around the IRAP’s test-retest are statistically significantly lower than the IAT. This is discussed on page XX.*

"Test Retest Reliabilities of the magnitudes shown in Table 1 are satisfactory for using most of Table 1's measures in correlational studies. They are, however, problematic for individual diagnostic uses—i.e., uses that seek either to describe traits of individuals or to characterize differences between individuals. There has not yet been any development of high-precision implicit measures." (Greenwald & Lai, 2020, p. 245).

*Author response: The reviewer presents a single citation – by the creator and proponent of the IAT – as being evidence of the acceptability of that measure. We should not be very surprised when authors describe their own measures as acceptable regardless of the reliability estimate obtained. Indeed, this finding was described by REF who found that test-retest estimates are ubiquitously described as “acceptiable” or “good” regardless of what numerical value is obtained (REF). A less biased source of advice for reliability might be a psychometrician who isn’t advocating for their own measure. For example, the mostly widely cited source for reliability cut-offs is REF, who gives a thorough treatment of when and why those values are valuable and necessary, and recommends .70, .80, or .90 depending on the use case.*   
  
-- The proposed manuscript attempts to make a virtue out of the fact that its data was largely, if not mostly (this is worryingly unclear), derived from the authors' file drawers. They claim that this will give a more faithful assessment of the IRAP's reliability than the existing literature. However, the proposed manuscript completely disregards the possibility that some of that research may not have been of sufficient quality to be safely published or therefore interpreted. Certainly, there is no mention in the proposed manuscript of what experimental designs gave rise to the authors' file drawer data -- or more to the point, how those experimental designs related to reliability at least in principle. To illustrate one aspect of my point, consider for a moment that in any given IRAP some trial-types describe phrases that are more in keeping with a given criterion behavior than others (i.e. in the sense of what we already know empirically and functionally about that  
behavior from previous non-IRAP research). As such, it would be ill-advised and specious to treat all four trial types as being equally suitable candidates to test either the reliability or the validity of the IRAP writ large. And yet, it appears from the proposed manuscript that the proposed manuscript's meta-analysis was based upon overall D-scores that collapsed each IRAP's four trial-types' data into one score.  
  
-- The proposed manuscript's introduction concluded by objecting at length about the methods used within the IRAP literature for calculating internal reliability. In particular, it argued from Parsons et al (2019) that the IRAP has not assessed the reliability of the size of the individual differences that it measured but instead has only measured the reliability of the rankings of those individual differences. This argument is besides the point given what I have already explained about the current evolution of the IRAP being formulated for group-based analyses rather than for quantifying individual differences per se. Even if one were to ignore this point, when the authors meta-analysis revealed that these new methods of computing the reliability of the IRAP D scores didn't actually produce results in this case that were practically different from more traditional methods of calculating reliability such as split-half methods.

*Author response: the reviewer may have misunderstood or mischaracterised these points. First, the method are not new, they are at least 30 years old. Second, the measurement property of reliability has implications for both the individual level and the group level. There is no way to divorce implications from one type of analysis from the other; to argue otherwise is to demonstrate that one does not fully comprehend the concept of reliability or the mathematical relationships that define it. For example, the Standard Error of Measurement (SEm), which can provide confidence intervals around an individual’s scores on a measure, is defined in part by the group-level test-retest reliability of the measure. Group level reliability and individual level precision are closely related concepts.*

-- The proposed manuscript repeated reports 'significant differences' where no such tests were performed (e.g. the bottom line on page 14), …

*Author response: Thank you for catching this slip: I forgot to include the confidence intervals on Greenwald & Lai’s estimate that allow for this inference test (although note that they were provided earlier in the introduction). P. XX now reads:*

*“Due to the non-overlap of the confidence intervals, the IRAP’s test-retest reliability (r = .14, 95% CI [-.08, .35]) was therefore found to be both statistically significantly and substantively lower in magnitude than the IAT’s (r = 0.50, 95% CI [0.45, 0.55]: Greenwald & Lai, 2020). Note that this comparison is done via the IRAP’s test-retest r estimate rather than ICC in order to compare like with like, given Greenwald and Lai (2020) estimated r.*

* *Update all estimates from Greenwald and lai to use their values not my incorrect ones!*

…and mischaracterizes a correlation of 0.5 as being medium-sized in order to exaggerate the point being made (page 16).

*Author note: Cohen’s (1988) guidelines for the interpretation of correlations are almost ubiquitous in many areas of science (e.g., the book containing these interpretation guidelines having been cited more than 240,000 times on google scholar). It is uncontroversial to refer to r = .50 as “medium” in size. A reference to Cohen (1988) has been included on page XX.*