Manuscript Number: JCBS-D-23-00021    
  
A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using  
the Implicit Relational Assessment Procedure    
  
Dear Dr. Hussey,      
  
Thank you for submitting your manuscript to the Journal of Contextual Behavioral Science. The AE, Dr. Rogge, received 2 reviews of your manuscript and provided his own feedback. Based on these reviews (found below) we have decided to ask you to make minor revisions and resubmit. We are asking that the revised manuscript be submitted by Aug 11, 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 such an interesting review of the power and sample sizes within the IRAP literature. I feel that your manuscript is an excellent extension of the work of Fraley and colleagues (2022). More importantly, the comparison of the growth of sample sizes in the social literature and the IRAP literature in the last decade (i.e., since the discovery of the replication crisis) is a bit alarming and provides critical insights toward understanding limitations of that work and how future research on the IRAP needs to be improved. I have received two reviews of your manuscript and the reviewers agree with me on the importance of this work and its potential impact on the field. Both reviewers have provided comments and suggestions to help strengthen your work and I have a suggestion for improvement as well. Thankfully, most of these revisions involve fairly minor additions and revisions to your narrative. I am therefore happy to recommend a revise and resubmit with minor revisions at this point.  
  
Despite my enthusiasm for your manuscript, I want to note that this recommendation does not guarantee the ultimate acceptance of your manuscript. As with all papers in peer-review, the level of review following this revision and the ultimate decision on this manuscript depend heavily on the depth and responsiveness of your revisions to all of the comments raised as well as the level of detail provided in describing and explaining the revisions made within your response letter.  
  
I look forward to reading your revised manuscript and I hope the comments raised in this peer-review are helpful in revising it. Please let me know if you have any questions.  
  
Sincerely,  
  
Ron Rogge  
Associate Editor, JCBS  
  
ASSOCIATE EDITOR COMMENT:  
The sentence after you first introduce Figure 2, you explain what is in the graphs: “In this and all subsequent figures, the straight line represents the fitted Ordinary Least Squares linear regression line (discussed later) and the shaded region around it represents its 95% Confidence Interval.” The second half of that sentence does not make sense as none of your line graphs have shaded regions. Please delete that portion of the sentence.  
  
  
REVIEWER 1 COMMENTS:  
The purpose of this article was to review the empirical literature using the Implicit Relational Assessment Procedure, with a focus on the typical sample sizes used. The authors argue that, in the wake of the replication crisis in psychology, there has been a greater interest in improving research methods. One such call, which actually goes back several decades, is for greater statistical power. Power is useful in empirical research because, on the one hand, underpowered work cannot detect real effects that exist. But, it is also important because, in literatures where the typical power of studies is low, the false positive rate will be higher and published effect size estimates will be inflated.  
  
The author finds that the typical sample size in work using the IRAP is close to 64 and that the power of a typical study to detect a typical effect size of r = .20 is only 34%. Moreover, although there is some evidence that sample sizes may be increasing, they are increasing at a rate that, frankly, is unacceptable in light of these debate.  
  
I think this paper addresses an important topic and does so in a way that seems systematic, comprehensive, and credible. As such, it is my guess that this paper would make a fantastic contribution to the literature. I don't have any substantive criticisms. I have a couple of constructive suggestions below; use them if you find them helpful.  
  
1. The author is reluctant to make specific recommendations for sample sizes. I understand why. Nonetheless, I do think the natural arc for a paper like this is to end with recommendations (i.e., identify a problem, explain why it is important, make some recommendations) that people can easily follow. If the bottom line is "think about power in a way that you think is sensible for the problem you're studying," most researchers (many of whom are, ironically, not adept at statistics and methods) will make poor decisions. They will, for example, assume the effects they are studying are likely to be much larger than that "other" people study. I can't think of any good reason why researchers in 2023 should not--as a bare minimum--power their studies to be capable of detecting an effect of r = .20 or higher. The costs of doing underpowered research are simply too high to try to take short-cuts based on unknown or assumed population parameters.  
  
2. This is just a preference issue: I think it makes sense to "connect the dots" in a time series graph it no other interpolation is going to be used. But, if one is fitting linear regressions to the data points, I'd rather just see the points and the regression line; the "connect the dots" line isn't really needed in such a situation.  
  
REVIEWER 2 COMMENTS:  
This is a very well written and important paper about a central measure used in the CBS literature. It uses well articulated strategy to demonstrate important cultural practices inside the CBS community that are likely very harmful to the science being conducted. These results show that it is likely that many IRAP findings are not replicable and that effect sizes are likely overestimated in the existing literature due to file drawer effects. The paper uses a straightforward approach to assessing the literature, that while imperfect, is strong and has valid conclusions. The data is available for other researchers to verify whether conclusions are warranted. In addition, the author draws well formulated and sober conclusions from the findings and does not exaggerate nor draw conclusions that are overly broad. In all, this is an excellent paper that is a service to the field and much needed to hopefully start to correct these systemic problems.  
Below I note a number of minor issues with the paper in order to further strengthen it.  
  
--They need to better describe what a multiway ANOVA is. I'm not completely confident I know what the author is saying in using that term and so I suspect other readers may not be either. It would also be useful to demonstrate/explain why it inflates FP rates, which is not obvious from the current description.  
[AE NOTE – I think you might be referring to 2-way ANOVAs, 3-way ANOVAs, etc. as a group. Please clarify this and add narrative as requested.]  
  
--Typo here: "inductive manner (Lakens, 2021) or in an inductive manner"?  
[AE NOTE – from the rest of the paragraph, it would seem that the phrase prior to the Lakens citation was intended to be “deductive manner” – please verify and correct this]

--The writer should describe the rationale for "Variant procedures such as the Mixed-Trials IRAP (MTIRAP: Levin et al., 2010) and the Training IRAP (T-IRAP: Kilroe et al., 2014) were excluded." This is important for readers less familiar with the IRAP literature.  
  
--This statement seems incorrect or at least I am interepreting it to be incorrect, "median sample sizes in IRAP studies are small (range 12 to 64)" as the Figure 1 shows samples in the 200s.  
  
-- I don't think this statement is correct, "Results demonstrated that the implied statistical power to detect the average published effect size (Cohen's d = 0.408, equivalent to Pearson's r = 0.20) was increasing from an estimated .142, 95% CI [.108, .177] in 2006 (the model intercept) by an average of .009, 95% CI [.005, .012], p < .001 participants per year." I believe the "participants per year" should be deleted?  
  
--There are numerous typos on the manuscript that should be corrected by careful proofreading. I'd recommend the author have someone else read the manuscript with the eye of catching potential typos lest this weakness take away from the perceived intellectual contribution of the manuscript.  
  
-- One additional weakness that should be noted is that it appears that only the author participated in the coding, lending the possibility of systematic bias or inaccurate coding. I realize other people can check all the coding, but this is an arduous process that is not likely to occur, so this remains a weakness that should be noted.