Dear Prof Levin, Prof Rogge, and anonymous peer reviewers,

Thank you for your very helpful reviews of my manuscript.

I have provided point-by-point responses to each of your comments below. To aid your understanding of the original comment and my changes, I have labeled each actionable comment in square brackets, e.g., “[reviewer 1 comment 1]”. My response is then provided in italics below, followed by quotes from the changes made to the manuscript.

I have made a few substantive additions to the manuscript based on your excellent comments, and I think they have improved the manuscript.

Kind regards,

\*\*\*\*

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 \*\*\*\*,

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

[AE comment 1]

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.

*Author response: Thank you for noticing this. The pdfs of these plots do indeed have shaded regions, but something about the editorialmanager pdf creation process has rendered them transparent. In addition to embedding the plots inside the manuscript I have also attached them as separate PDFs, in which you can see the shaded regions.*

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.

[reviewer 1 comment 1]

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.

*Author response: I have added worked examples of power analyses for r = .20 (as used elsewhere in the manuscript) as well as Cohen’s d = .46, which represents the average transdiagnostic efficacy of ACT and therefore a potentially useful benchmark for intuitive comparison. Page 26-28 now reads:*

*“While keeping the above numerous qualifiers in mind as to why power calculations must be done for each study with respect to its design, measurement quality, and specific analysis, some readers may nonetheless wish for some discussion of the sample size needed to detect specific effects in future IRAP studies. I will therefore take two examples, both related to common use cases for the IRAP.*

*First, imagine a researcher who wishes to detect a correlation between IRAP scores and another continuous variable. Power analyses were again conducted using the R package pwr (Champely, 2016). To detect a correlation that is the magnitude of the average effect size observed in published psychological research (i.e., Pearson’s r = 0.20); with an α-level of 0.05 and a two-sided test (typical for correlations and regression); 194 participants would be needed to obtain the minimum recommended power (80%), or 319 participants for high power (95%). Comparing these sample sizes to the sample sizes observed in the published IRAP literature, 2% of published IRAP studies employed a sample size sufficient to detect the average effect size in the broader psychology literature (r = .20) with minimum recommended power (80%), and 0% could detect this with high power (95%). Note that half of all published effects are smaller than the average, and therefore would require even higher sample sizes to detect.*

*Second, imagine that a researcher wishes to detect a difference in mean IRAP scores between two groups, and they wish to power the study to detect effect sizes that are at least the average transdiagnostic efficacy of Acceptance and Commitment Therapy versus active and inactive control groups as reported in a recent review of meta-analyses (Hedges’ g = 0.46: Gloster et al., 2020). This effect size is no more or less meaningful than any other for the sake of illustration but is chosen to leverage intuitions the reader may have to define its Smallest Effect Size of Interest. That is, such a study would be powered to detect mean differences on the IRAP between groups that are no smaller than the general efficacy of ACT. Loosely speaking, I am inviting the reader to think about other effect sizes they have an intuition for to provide a benchmark. For example, if you are interested in between groups differences that are smaller than the average efficacy of ACT, you would need more participants than this again. With an effect size of Cohen’s d = .46 (i.e., ignoring the very small difference between Hedges’ g and Cohen’s d at this sample size), an α-level of 0.05, a two-sided test, and two equally sized groups; 152 participants would be needed to obtain the minimum recommended power (80%), or 248 participants for high power (95%). Comparing these sample sizes to the sample sizes observed in published IRAP research employing between-groups or mixed within-between designs, 2% of such published IRAP studies employed a sample size sufficient to detect an effect size the same size as the average transdiagnostic efficacy of Acceptance and Commitment Therapy (Hedges’ g = 0.46: Gloster et al., 2020) with minimum recommended power (80%), and 0% could detect this with high power (95%).*

*Please note again that the above power analyses are not specific recommendations for future IRAP research, and this article should not be cited as a source of such recommendations, but rather are worked examples of (a) how authors should begin to engage with conducting their own a priori power analyses, and (b) illustrative of how underpowered the published IRAP literature is to detect effect sizes of these magnitudes. I reiterate this caution due to The Law of Lakens’ Guidelines, which states that whenever you try to make the point that researchers should not follow certain guidelines, you will nonetheless sometimes be cited as a source of said guidelines (Rohrer, 2023).”*

[reviewer 1 comment 2]

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.

*Author response: Thank you for this suggestion. I retraced my thinking on this and this choice to "connect the dots" came down to the issue of accessibility and interpretability. To increase accessibility, I have chosen a color palette that is color-blind friendly and still interpretable when printed in black and white. However, interpreting whether each dot belongs to one group or the other still requires the reader to discriminate the color of individual dots, which can be difficult for some. This is particularly the case in Figure 2 where some of the green dots (‘all studies’) overlap with the blue dots in the other group (‘studies with between-subject comparisons’). I’ve tried combinations of removing the joining lines and elaborating the legend, but this always produces a plot that is ambiguous under at least some conditions or for some viewers. Given that the addition of the connecting lines is at worst redundant, I’ve therefore elected to retain them for accessibility and interpretability of the group membership of each data point.*

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.

[reviewer 2 comment 1]

--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.]

*Author response: Thank you – I agree this point needed fleshing out. Page 6 now reads:*

*“One specific class of statistical methods, multiway Analyses of Variance (ANOVAs, i.e., those with more than one independent variable such as 2-way ANOVAs, 3-way ANOVAs), are almost ubiquitous in IRAP research. Due to familywise error rates, the use of multiway ANOVA in exploratory or inductive research inflates false positive rates much higher than the 5% rate implied by the standard alpha level of 0.05 (Cramer et al., 2016). In the case of a simple 2X2 between groups ANOVA, this can be illustrated with simple math: if a researcher is willing to accept the result of any of the three p values generated by the ANOVA (i.e., either main effect or the interaction effect) as evidence of an effect, as would be common when applying the ANOVA in a an exploratory or inductive manner, then the false positive rate for the ANOVA as a whole is not equal to the alpha value (e.g., 5%), but a higher value. Specifically, False Positive Rate = , where k is the number of p values. Using alpha = 0.05 and k = 3 (i.e., two main effects and one interaction effect), False Positive Rate = 14.3%. Cramer et al. (2016) note that the false positive rate implied by larger ANOVA designs, such as those often employed in IRAP research (e.g., 4X2X2 mixed within-between ANOVAs), are higher again, but would require specific simulation studies to estimate.”*

[reviewer 2 comment 2]

--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]

*Author response: Thank you for catching this – you are correct. Page 7 now reads:*

*“Testing in a deductive manner (Lakens, 2021) or in an inductive manner (e.g., to generate new hypotheses rather than test existing ones).”*

[reviewer 2 comment 3]

--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.

*Author response: Thank you for this suggestion. I originally didn’t get into detail here for the sake of brevity. I have explicated this logic now. Pages 9-10 now read:*

*“Variant procedures such as the Mixed-Trials IRAP (MT-IRAP: Levin et al., 2010) and the Training IRAP (T-IRAP: Kilroe et al., 2014) were excluded on the basis that, although these tasks share similar names with the IRAP, they specifics diverge so substantially from the IRAP as to represent a strong risk of a jingle fallacy: the mistaken assumption that two measures sharing the same name measure the same thing (e.g., Lilienfeld & Strother, 2020). For example, the IRAP requires participants to provide responses that are both notionally consistent and inconsistent with their pre-experimentally established learning history (e.g., to respond to “White people” and “positive” with “true” on some blocks and “false” on others). In contrast, the Training IRAP requires responding consistent with only one of these patterns in order to establish that pattern of responding rather than assess it. Despite its name, the Training IRAP is therefore more closely related to the Relational Evaluation Procedure (e.g., J. Hayes et al., 2016) than the IRAP.*

*The risk of a jingle fallacy also applies to treating the IRAP and MT-IRAP as if they are meaningfully similar. Whereas the IRAP alternates between response patterns between blocks, the MT-IRAP does it between trials through the inclusion of an additional stimulus that indicates whether participants should tell the "truth" (provide a history-consistent response) or "lie" (provide a history-inconsistent response) on that trial. There are thus parallels between the MT-IRAP and the Recoding Free version of the Implicit Association Test (IAT-RF: Rothermund et al., 2009) as variants of their respective original tasks. To the best of my knowledge, no work to date has assessed the correlation between IRAPs and MT-IRAPs designed to assess the same domain. More broadly, it is important to note that although the IRAP and several other tasks including the Implicit Association Test are collectively labeled "implicit measures", scores on these tasks are typically found to correlate poorly with one another, even when the tasks share some procedural features and are intended to measure the same domain (e.g., Clayton et al., 2023; Schimmack, 2021; for a detailed conceptual critique see Corneille & Hütter, 2020). As such, in the absence of evidence for convergent validity between the IRAP and MT-IRAP, the MT-IRAP was excluded out of an abundance of caution against introducing jingle fallacy into the analysis.”*

[reviewer 2 comment 4]

--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.

*Author response: Thank you for catching this lack of clarity. Figure 1 does indeed illustrate the actual sample sizes observed, but the paragraph you quote is from the section on “Change in sample size per study over time”. The quoted text has been adjusted to correspond to the paragraph’s point more closely:*

*“As can be seen in Figure 2, across years, median sample sizes in IRAP studies have been small (range of medians 12 to 64).”*

[reviewer 2 comment 5]

-- 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?

*Author response: Thank you for catching this error. “participants per year” was deleted from this quote on page 18.*

[reviewer 2 comment 6]

--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.

*Author response: Thank you – I have corrected several dozen typos and suboptimal word choices throughout the manuscript and had it thoroughly proofread.*

[reviewer 2 comment 7]

-- 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.

*Author response: Good point. I would have preferred a second coder, but this project evolved quickly and I didn’t find an interested collaborator to do this second coding. I have added this limitation to the limitations section. Page 28 now reads:*

*“Extraction of sample sizes*

*Sample sizes were extracted from the original articles by a single coder and no estimates of inter-rater reliability were therefore produced. All data and code for the current manuscript are public, therefore making the accuracy of these extractions testable in principle. However, given this is a non-trivial task, and inaccurate data extractions could lead to bias, the lack of a second scorer must be acknowledged as a limitation.”*