**Peer Review and Communication History**

**Ms Title:** Malleability of statistical perception: Impact of validity presentation on college admission test policy preferences

**Author Names:** Don C. Zhang& Jonathan Wai

**Submitted:** Jan 12, 2021

**Editor First Decision—Revise & Resubmit**

Mar 31, 2021

Dear Don Zhang,

I have now received all reviews of your manuscript, “Communication of validity evidence and admission test policy preferences” from qualified researchers. I also independently read the manuscript before consulting these reviews. I agree that your manuscript has important strengths and also that there are some issues that need to be addressed. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

The reviewer did an outstanding job in their review. I will highlight issues I think are particularly salient here. In your resubmission, please include a document with a point-by-point response to both the points I list here and the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

I was able to obtain a single review for your MS. As that review was very thorough and given my challenges in finding another willing reviewer, I am making a decision based on that one review and my own independent reading of your work. Both the reviewer and I reviewed your work favorably but also note several places where the manuscript might be strengthened.

I believe that a careful revision can address the reviewer’s concerns. In places where it cannot, simply acknowledging the critiques as limitations will place findings in the correct context.

The power analysis should specify the exact effect size. Additionally, since the primary predictions seem to focus on mediated effects in latent variable models, a power analysis focused simply on group means does not adequately address power for your analyses. As you are using a relatively complex analysis, I would recommend a tool like the R package simsem to build your power analysis. For the most part, canned power analyses (like you get from GPower) are useful only for the simplest of designs.

I do not believe it is necessary to explain why you are using SEM. You have multiple DVs and MANOVA is useless for tests of mediation (as it is for pretty much everything). Simply explaining the analysis strategy is more than enough information.

In several places, where there were non-significant results, there are no statistics presented. Statistics need to be presented in full for any test reported, not just the significant ones. Additionally, wherever you present ANOVA or similar analyses, there should be an effect size included. For effect sizes that can exceed 1.0 (like d), place a zero in front of the decimal. For values that cannot exceed 1.0 (like alpha), please exclude the leading zero (this is all APA Style minutiae). Please see the APA manual for presentation of scale anchors.

The Center for Open Science offers this text as part of their standard reviewer statement: “"I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. (Nosek et al., 2017).” This statement provides excellent guidance. Please add text to your paper that addresses these issues (where relevant).

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at editorialoffice@collabra.org.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Chris Aberson

**Reviewer 1**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I found this manuscript very interesting as a study in how presentation of effect size is related to perception of validity, and through that, how people use that information to make decisions. I believe it has shown that different choices of effect size are related to this perception, particularly as it relates to the case of admissions testing.

I will start with commenting on the statistical validity. I believe the choice of SEM was appropriate and reasonable, given that the mediation hypothesis was of primary interest. I believe some information was missing about how this was implemented that would clarify whether this was implemented appropriately. Standard SEM, if that is what was used, requires normally distributed outcomes, and I did not see any discussion of whether this assumption was appropriate. Particularly for study 2, where the outcome for the test-optional policy endorsement was ordinal and the test-optional policy petition was an indicator variable, this may be problematic and an adjustment to the SEM model may be required. This also means it might be more difficult to justify combining the data from the two studies, as was also done, since the variables were of different types.

With regard to construct validity, there were measures of reliability reported for each construct used, which is appropriate. A confirmatory factor analysis was performed for the measurement model for study 1 (and for study 2, although the writeup was not as comprehensive–were alternative measurement models explored here and found to be inappropriate?). Questions: were the measures of reliability obtained from the same sample that the analysis was conducted on, or was there previous testing (particularly for the newly developed test-optional policy endorsement measure)? Was there any additional validity testing done for the newly developed measure? This question also applies to the shortened versions of the scales used in study 2. Although it is good practice to validate instruments on samples taken specifically for that purpose, it is not uncommon to run these tests on the same sample used in the study–but it would be appropriate to clarify this.

With regard to internal validity, it appears that randomization was applied appropriately with respect to the experimental condition (the type of effect size that was presented to the respondent). This does allow for causal claims, although these were kept to a minimum in the conclusion, with the authors opting for words like “association.”

My biggest concern with the study is the extent of the generalization. This is not a concern for the samples of individuals; the researchers did an excellent job in describing their samples and discussing their generalizability. Including two separate studies, one using an online crowd-sourcing platform and one using a community-based sample, was an impressive aspect of this manuscript. The concern is regarding the connections between the effect sizes presented to the participants, the perceived validity of the admissions tests, and the true usefulness of the admissions tests. I have broken this discussion down into several points that follow.

There is a difference between the “traditional” and the “alternative” measures of effect sizes concerning their subjectivity, and this is not commented on by the authors. The r and the R2 can each only be defined one way (although that does not necessarily make their meaning clear or make them “good” effect sizes). The other effect sizes (expectancy chart, BESD, and CLES) require selection of cutoffs that could be chosen in such a way as to potentially bias perceived validity. For example, the expectancy chart reports “proportion above 3.5 in GPA” as an outcome. Would the results for perceived validity be different if a cutoff of 3.0 had been chosen instead? Similarly, the CLES reports on “a randomly chosen student with an ACT score greater than 26”; would the results for perceived validity be different if a cutoff of 30 had been chosen instead?

Additionally, at the core of the paper is the idea that standardized tests *should be used* as part of college admission criteria. It is not clear to me that this is widely agreed upon, although this is not an area of expertise for me (I am only aware of this issue through a media lens, not as a researcher, and this should be evaluated by another expert). However, if one were to take the alternative view, one could easily come to opposite conclusions about the appropriateness of the effect sizes presented; would an R2 be considered more appropriate because it is associated with *not* endorsing an admissions test? Even if it is not considered debatable that standardized tests should be used in college admissions–does the use of the alternative effect sizes always lead to better alignment with *actual* validity and/or usefulness, or does it always make people *believe* that a measure is more valid, whether it is or not? If the latter, it might be considered a manipulation of perception, rather than better communication of scientific evidence.

I think this could be addressed by re-orienting the context of the paper to showing that different presentation of effect sizes *changes* perceptions, rather than claiming that certain effect sizes result in better communication of scientific evidence for decision making. For example, the last two sentences of the manuscript are “We showed in our paper that simply presenting statistical evidence differently can affect how evidence is interpreted and acted upon. Indeed, when evidence is presented clearly and effectively, people will listen and decide according to the facts.” I believe the first of these two sentences is entirely appropriate; the second has not been shown by this manuscript, as the conclusions are at best situationally dependent (and potentially event debatable within the context of admissions tests, although again, this is not my area of expertise).

Two more specific comments: at the end of p. 4 in my pdf (first section of the paper), I would recommend describing study 1 as a *crowd-sourced* online experiment specifically (it makes it more clear how it differs from study two at an earlier point in the paper). I would also recommend to remove the comment about the one-sided t-test in the supplemental analysis for study 1 (it seems like this was only added because it supported the results, as no other one-sided results were reported).

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  | ✔ |  |  |  |

**Author Response**

**Apr 6, 2021**

Dear Dr. Aberson,

On behalf of my co-author, Dr. Jonathan Wai, I am submitting a revised manuscript for consideration for Collabra: Psychology. In the attached cover letter, I have included my responses to you and the reviewer’s comments. We also like to point out a revision to the paper’s title. We feel the revised title is more reflective of the changes in the paper’s framing based on reviewer 2’s suggestions. We believe that the manuscript has significantly improved as a result of these revisions. We are grateful for your guidance and remain available to address any further suggests you and the reviewing team has.

Regards,

Don C. Zhang
Assistant Professor
Department of Psychology
Louisiana State University
zhang1@lsu.edu

EDITOR COMMENTS

The power analysis should specify the exact effect size. Additionally, since the primary predictions seem to focus on mediated effects in latent variable models, a power analysis focused simply on group means does not adequately address power for your analyses. As you are using a relatively complex analysis, I would recommend a tool like the R package simsem to build your power analysis. For the most part, canned power analyses (like you get from GPower) are useful only for the simplest of designs.

Thank you for the recommendation. We chose to use G\*Power initially because although we used a complex analysis (SEM) to test the indirect effects, the central hypothesis (the effect of the treatment on perceived validity) is best conceptualized as a between-subject treatment effect.

With that said, we have now added details on the exact effect sizes we used in our sample size planning, as well as power analysis for the entire SEM model.

I do not believe it is necessary to explain why you are using SEM. You have multiple DVs and MANOVA is useless for tests of mediation (as it is for pretty much everything). Simply explaining the analysis strategy is more than enough information.

The explanation has been removed.

In several places, where there were non-significant results, there are no statistics presented. Statistics need to be presented in full for any test reported, not just the significant ones. Additionally, wherever you present ANOVA or similar analyses, there should be an effect size included.

We have also added effect sizes for all ANOVAs results and t-tests, regardless of significance. In some cases, when results are presented in tables (path estimates), we chose not to repeat the results in-text.

For effect sizes that can exceed 1.0 (like d), place a zero in front of the decimal. For values that cannot exceed 1.0 (like alpha), please exclude the leading zero (this is all APA Style minutiae).

Thank you, this has been fixed throughout.

Please see the APA manual for presentation of scale anchors.

Fixed, thank you.

The Center for Open Science offers this text as part of their standard reviewer statement: “"I request that the authors add a statement to the paper confirming whether, for all experiments, they have reported all measures, conditions, data exclusions, and how they determined their sample sizes. The authors should, of course, add any additional text to ensure the statement is accurate. (Nosek et al., 2017).” This statement provides excellent guidance. Please add text to your paper that addresses these issues (where relevant).

We have added this statement in the “Overview of Experiments” section.

“We also confirm that all measures, conditions, data exclusions, and sample size planning is accurately reported for both experiments.” (p. 12)

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Thank you for the opportunity to revise the manuscript.

REVIEWER COMMENT

I found this manuscript very interesting as a study in how presentation of effect size is related to perception of validity, and through that, how people use that information to make decisions. I believe it has shown that different choices of effect size are related to this perception, particularly as it relates to the case of admissions testing.

We are grateful for the reviewer’s constructive feedback.

I will start with commenting on the statistical validity. I believe the choice of SEM was appropriate and reasonable, given that the mediation hypothesis was of primary interest. I believe some information was missing about how this was implemented that would clarify whether this was implemented appropriately. Standard SEM, if that is what was used, requires normally distributed outcomes, and I did not see any discussion of whether this assumption was appropriate. Particularly for study 2, where the outcome for the test-optional policy endorsement was ordinal and the test-optional policy petition was an indicator variable, this may be problematic and an adjustment to the SEM model may be required.

We appreciate the reviewer’s insight. We originally used maximum likelihood estimation, which assumes normally distributed continuous indicators. We have now revised our model estimation approach, in particular for Study 2, to use diagonally weighted least squares (WLSMV) estimation, which is more robust against violation of normality is better suited for modeling of latent variables with ordinal/categorical indicators (Li, 2016). Model estimates and fit indices have been updated accordingly. Overall, the changes in parameter estimates were in the magnitude of 0.01 ~ 0.02 and did not in any way change the interpretation of the results. On the issue of non-normality: in both studies, the model parameters were estimated using bootstrapped resamples, which is robust against violations of normal distributions (Barber & Thompson, 2000).

Barber, J. A., & Thompson, S. G. (2000). Analysis of cost data in randomized trials: a

n application of the non‐parametric bootstrap. *Statistics in medicine*, *19*(23), 3219-3236.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949.

This also means it might be more difficult to justify combining the data from the two studies, as was also done, since the variables were of different types.

To clarify, we only combined data from Studies 1 and 2 to test our moderation hypotheses. Here, we did not include the dichotomized DV (petition signature), which was not part of Study 1. Instead, our dependent variables (perceived validity, test preference, test-optional policy preference) were all measured using Likert scales, which were transformed to POMP, and then combined.

With regard to construct validity, there were measures of reliability reported for each construct used, which is appropriate. A confirmatory factor analysis was performed for the measurement model for study 1 (and for study 2, although the writeup was not as comprehensive–were alternative measurement models explored here and found to be inappropriate?). Questions: were the measures of reliability obtained from the same sample that the analysis was conducted on, or was there previous testing (particularly for the newly developed test-optional policy endorsement measure)?

The reported reliabilities were obtained from the same sample. The items we used in Study 1 were drawn from a previous paper (Kim & Berry 2015). We now include results for alternative CFA models in Study 2 to provide further evidence for construct validity (see supplemental materials).

Was there any additional validity testing done for the newly developed measure? This question also applies to the shortened versions of the scales used in study 2. Although it is good practice to validate instruments on samples taken specifically for that purpose, it is not uncommon to run these tests on the same sample used in the study–but it would be appropriate to clarify this.

There was no separate validation done with the items outside of the samples in this study.

We recognize that post-hoc measures can suffer from potential construct validity issues particularly with complex constructs. In our case, we feel that the constructs of interest are fairly well-defined and that the items we chose are face-valid. We acknowledge that, to some extent, our measure of test-optional policy preference may suffer from some coverage issues, particularly for Study 2, where a single item was used. But the psychometric shortcomings, we argue, is somewhat vindicated by the external realism of using actual signatures from a field experiment as a real behavioral outcome. Nevertheless, we note the limitation in our discussion.

*“Our study relied primarily on ad-hoc measures that were adapted for our research, which may potentially suffer from construct validity. Although we took care to ensure that the items used in our study meaningfully distinguished the constructs of interest, future research could improve on the measures used in this study.” (p33)*

With regard to internal validity, it appears that randomization was applied appropriately with respect to the experimental condition (the type of effect size that was presented to the respondent). This does allow for causal claims, although these were kept to a minimum in the conclusion, with the authors opting for words like “association.”

Although the direct effect of the treatment on the outcomes can be interpreted as causal due to random assignment, the associations amongst the measured variables (e.g., perceived validity -> test endorsement) is still associative in nature. Nevertheless, we appreciate the reviewer pointing this out and have ensured to use causal language only when justified.

My biggest concern with the study is the extent of the generalization. This is not a concern for the samples of individuals; the researchers did an excellent job in describing their samples and discussing their generalizability. Including two separate studies, one using an online crowd-sourcing platform and one using a community-based sample, was an impressive aspect of this manuscript. The concern is regarding the connections between the effect sizes presented to the participants, the perceived validity of the admissions tests, and the true usefulness of the admissions tests. I have broken this discussion down into several points that follow.

There is a difference between the “traditional” and the “alternative” measures of effect sizes concerning their subjectivity, and this is not commented on by the authors. The r and the can each only be defined one way (although that does not necessarily make their meaning clear or make them “good” effect sizes). The other effect sizes (expectancy chart, BESD, and CLES) require selection of cutoffs that could be chosen in such a way as to potentially bias perceived validity. For example, the expectancy chart reports “proportion above 3.5 in GPA” as an outcome. Would the results for perceived validity be different if a cutoff of 3.0 had been chosen instead? Similarly, the CLES reports on “a randomly chosen student with an ACT score greater than 26”; would the results for perceived validity be different if a cutoff of 30 had been chosen instead?

This is a good point. Our study did not examine every possible iteration of these choices (e.g., proportion with 4.0 GPA; proportion below 3.0 GPA, etc.). We now discuss this as a limitation in our paper:

*“Relatedly, we the presentation of alternative effect sizes requires a selection of cut-off values for the criteria (e.g., proportion of students with GPA > 3.5). Depending on the choice of cut-off, the resulting evidence will likely differ. Future research, therefore, should more comprehensively examine the different ways in which alternative effective sizes can be presented and how these choices may affect interpretation of evidence.” (p.33)*

Additionally, at the core of the paper is the idea that standardized tests should be used as part of college admission criteria. It is not clear to me that this is widely agreed upon, although this is not an area of expertise for me (I am only aware of this issue through a media lens, not as a researcher, and this should be evaluated by another expert). However, if one were to take the alternative view, one could easily come to opposite conclusions about the appropriateness of the effect sizes presented; would an be considered more appropriate because it is associated with not endorsing an admissions test? Even if it is not considered debatable that standardized tests should be used in college admissions–does the use of the alternative effect sizes always lead to better alignment with actual validity and/or usefulness, or does it always make people believe that a measure is more valid, whether it is or not? If the latter, it might be considered a manipulation of perception, rather than better communication of scientific evidence.

This is a poignant observation and unfortunately, there is no objective/analytical way for determining, a priori, whether ‘perceived’ validity aligns with ‘actual’ validity. Ultimately, perception of evidence and the decisions on which evidence is acted upon will always be subject to human judgment. For example, some people may perceive 70% efficacy of a COVID-19 vaccine as very effective whereas others may perceive the same evidence as not effective enough. We are not sure if there is way to decide which perception better aligns with actual evidence, without some degree of human judgment.

In practice, other factors such as test cost and adverse impact are also considered in conjunction with validity when deciding whether tests are used. Our study focused on perceived validity as a focal factor because our research focused on presentation and perception of validity evidence. We note, in our limitation section, the importance of considering other non-validity related reasons to justify decisions to adopt/reject admission tests.

*“Finally, our study focused primarily on interpretation of validity evidence affects policy preferences related to admission testing. In practice, attitudes toward standardized testing and decisions toward its use are multi-determined. Other factors such as cost of the test and adverse impact are also relevant for understanding whether standardized tests are used. Although an experimental design allows us to rule out these factors as potential confounds for our study, future research should take these factors into considering when modeling the decision processes related to admission test policies.” (p36)*

I think this could be addressed by re-orienting the context of the paper to showing that different presentation of effect sizes changes perceptions, rather than claiming that certain effect sizes result in better communication of scientific evidence for decision making. For example, the last two sentences of the manuscript are “We showed in our paper that simply presenting statistical evidence differently can affect how evidence is interpreted and acted upon. Indeed, when evidence is presented clearly and effectively, people will listen and decide according to the facts.” I believe the first of these two sentences is entirely appropriate; the second has not been shown by this manuscript, as the conclusions are at best situationally dependent (and potentially event debatable within the context of admissions tests, although again, this is not my area of expertise).

We think this is a good suggestion. We have revised our introductory paragraphs to present a more balanced framing. Specifically, we reframed the intro to place less focus on ‘science-practice gap’ and more on how different presentation of validity evidence can affect policy preferences with admission testing.

We have also revised our discussions to be more descriptive and neutral with respect to whether tests *should* be used. The concluding sentence noted by the reviewer has been changed to:

*“We showed in our paper that* *when the validity of standardized tests is presented using alternative effect size displays instead of traditional statistics (e.g., Pearson’s r), people perceive the test to be more predictive and are more willing to endorse its use for admission decisions.” (p33)*

Two more specific comments: at the end of p. 4 in my pdf (first section of the paper), I would recommend describing study 1 as a crowd-sourced online experiment specifically (it makes it more clear how it differs from study two at an earlier point in the paper).

Done.

I would also recommend to remove the comment about the one-sided t-test in the supplemental analysis for study 1 (it seems like this was only added because it supported the results, as no other one-sided results were reported).

Done.

We are grateful for the reviewer’s thoughtful comments.

**Editor Final Decision—Accept**

May 13, 2021

Dear Don Zhang,

I have now had a chance to read over your manuscript “Malleability of statistical perception: Impact of validity presentation on college admission test policy preferences”, along with the letter describing the changes you made. Thank you for your responsiveness to the concerns that the reviewers and I raised. I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

My apologies for the delay in reviewing this manuscript. I must have overlooked the email informing me that the revisions was received (I’m an incomming department chair, so I get 100s of extra emails a day now!).

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. Our managing editor will contact you in case there are any pre-prodution file related questions. You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely, Chris Aberson