Editor decision—Revise & Resubmit

May 7, 2020

Dear Dr. Brian Keith Miller,

Thank you for submitting your work to Collabra: Psychology. Foremost, I apologize for the delay in processing this paper. This was a busy semester and then the pandemic hit. In addition, Collabra switched editorial systems and I missed that the reviews were in. I am truly sorry for how long this paper has spent in review. Please accept my apology.

I was able to secure reviewers from two outstanding scholars who are experts in IER and careless responding. Both of these individuals are extraordinarily well qualified to review this paper and I thank them for their service to this journal. I independently read the paper and then consulted their comments.

As you will read below, the reviewers had mixed reactions to the paper. I was torn as well. I inherently find questions about how survey design features impact data quality to be important and worthwhile (as do the reviewers) and thus the basic experimental design used in the manuscript seemed interesting. On the other hand, there are some confusing aspects of the manuscript and some concerns about the logic underlying the hypotheses. This issue is also tied to the selection and justification of the DVs as operationalizations of reliability or internally consistent responding. In addition, I worried there was a missed opportunity to evaluate the impact of the experimental manipulations on means, SDs, and scale inter-correlations in an exploratory manner. I believe both reviewers struggled a bit to understand why the focus of the work was on reliability rather than on other features of data that can be influenced by survey design features that may or may not trigger certain reactions in respondents. The focus might be entirely appropriate, but the logic was not detailed enough in the text for the reviewers to draw that conclusion. In addition, to the “big” picture issues, there are also some basic design questions raised by the reviewers that might impact whether the paper is suitable for this outlet with respect to the consent process.

In short, I will extend a revise and resubmit decision because I think there is something interesting about the design and the overarching question. However, you should consider this a “high risk” R&R decision. I will send any revision to back to Reviewer #2 so factor that into your decision making about how to proceed. You might decide that the current package will face an easier road to publication at another outlet.

I think both reviewers raised excellent points and I suspect both of their views are representative of the reaction of other readers. You should address each of their concerns either in the revised text or in the response letter. Addressing their concerns will make this a stronger paper. I will comment on the most salient points and offer some of my own reactions when reading the paper in my letter. You might disagree with some (or all) of the points so feel free to pushback against any suggestions you believe will harm your work. Please just describe your counterpoints in the return letter.

1. Reviewer #2 raises important questions about the consent process of the study and nature of participants. Please clarify that issue for readers and address those questions. IRB approval (or an exemption determination from your IRB) is required for publication.

2. I had questions about the design and whether you needed to use a more complicated model to test for interactions between conditions. It seems as if there are 3 factors (Quality Control Items v. None; Scrambled v. Grouped Items; and Anonymity v. Confidentiality). The randomization was at the person level for the first two factors but at the classroom level for the anonymity factor which could (theoretically) complicate dependencies in the data. I am perhaps not too worried about the randomization level issue, but I do think testing interactions is worthwhile on an exploratory basis as in a 2x2x2 factorial. (And I might have missed something in the design so that could either be a main effect of my brain, the text description, or the interaction. If I missed something obvious, I apologize.)

3. I think it would help readers if they knew that the three scales analyzed in the paper were the only data collected. This should be noted in the paper and I think uploading a copy of each survey to the data repository would be useful to readers.

4. Explaining to readers how the quality control items were spread out in the various forms would be helpful. This is especially relevant in the group item design.

5. Both reviewers wanted greater justification for why the current manipulations are expected to impact internal consistency/reliability. I think this is an area of the paper that needs to be enhanced. To be sure, we are not asking you to hypothesize after the results are known but rather to flesh out the rationale for the predictions, so readers can evaluate the logic and how these results fit with the existing literature. I could also see value in looking at means and SDs for the scales and scale inter-correlations on an exploratory basis.

6. Reviewer #2 was especially interested in seeing greater justification for the specific DVs used in the paper.

7. I agree with the general sentiment expressed by Reviewer #2 (and perhaps implicit in reactions from Reviewer #1) about the need to clarify the point of the paper for readers and highlight the main take home messages and implications.

8. Reviewer #2 had a lengthy set of concerns about reporting exact p-values. I think adding these values will enhance the paper. I believe the reviewer’s intent was constructive, but this issue struck a chord with the reviewer as you will read. I will be honest and note that some of the passages surrounding this issue were perhaps too strongly worded for my taste. However, I saw them in the context of the whole review which was lengthy but provided in the spirit of helping to enhance the work. In other words, the language in that section might be pointed but I think you can report the exact values and address the concern. I took the reviewer as providing valuable reactions even if that section might have been intense.

9. I also think reporting effect sizes for the t-tests would be useful.

10. I thought there was some redundancy between the text and the tables in terms of reporting the same values for the tests in both places. I find tables easier to read so I don’t think it is necessary to have complete redundancy between the text and tables. This could streamline some sections. I would rather see effect size estimates reported and discussed.

Those were the major issues that I think require attention. I think clarifying the rationale, the nature of the data, and the major points will enhance the work. I hope you will agree and undertake a revision. You might also make sure the data file is usable by others by providing a short code book and including a .csv version so those without SPSS can access the data.

To access your submission account, follow the below instructions:
1) login to the journal webpage with username and password
2) click on the submission title
3) click 'Review' menu option
4) download Reviewed file and make revisions based on review feedback
5) upload the edited file
6) Click the 'notify editor' icon and email the confirmation of re-submission and any relevant comments to the journal.

Thank you for trusting us with your paper. I want to reiterate that this is a high-risk revision to be transparent about the nature of the task as I see it. If you have any questions or difficulties during the revision process, please do contact us. Good luck revising this work. Please also again accept my apology about the length of time this paper has been under review.

Sincerely,

Brent Donnellan
Michigan State University
donnel59@msu.edu

------------------------------------------------------
Reviewer A:

1) General comments and summary of recommendation
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
 - Are the methodologies used appropriate?
 - Are any methodological weaknesses addressed?
 - Is all statistical analysis sound?
 - Does the conclusion (if present) reflect the argument, is it supported by data/facts?
 - Is the article logically structured, succinct, and does the argument flow coherently?
 - Are the references adequate and appropriate?:
This manuscript reports an experimental study examining the influence of three survey design features on scale reliability. I agree with the importance of studying IER, and I see the merit in the use of the experimental approach. I hope my comments below may be useful as you further develop this paper.

1. The greatest challenge I had with the hypotheses of this paper is the assumption that “Pure content non-responsivity is completely random and results in a reliability of scores equal to zero”. Yet insufficient effort responding may not manifest in a true random fashion and may instead show up as repeated selection of the same response. When this occurs, Cronbach’s alpha may appear to increase instead of decrease (DeSimone, DeSimone, Harms & Wood, 2018). Thus, although I agree with the notion that surveys containing quality control items and confidential surveys may lead to lower IER, I am not sure one should expect higher reliability as a result. This issue may have contributed to the curious pattern of results found in the study.
2. Another question I had is the presence or absence of reversed coded items. One of the three measures (conscientiousness) had negatively keyed items, while the other two did not. Negatively keyed items can introduce method factor and presumably decrease reliability, if IER occurs (Schmitt & Stults, 1985).
3. One question I am curious about is whether IER may manifest as in reduced attention to the survey question, but instead of being nonresponsive to the questions, respondents simply determine if each question is desirable or undesirable, and respond favorably to desirable items and unfavorably to undesirable items. In other words, IER may result in social desirable responding. If (that is a big if) this happens, the hypotheses may become more complicated.
4. The exploratory factor analysis on conscientiousness circles back to my point #2 above. The presence of IER may have introduced the method factor for conscientiousness.
5. Please report effect sizes for t tests.

Reference
DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. Applied Psychology: An International Review, 67, 309-338.

2) Figures/tables/data availability:
Please comment on the author’s use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:
Use of tables is adequate. Underlying data is available.

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.:
I could not find a statement about ethical approval by an IRB.

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:
The text is generally well written.

------------------------------------------------------

------------------------------------------------------
Reviewer B:

1) General comments and summary of recommendation
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
 - Are the methodologies used appropriate?
 - Are any methodological weaknesses addressed?
 - Is all statistical analysis sound?
 - Does the conclusion (if present) reflect the argument, is it supported by data/facts?
 - Is the article logically structured, succinct, and does the argument flow coherently?
 - Are the references adequate and appropriate?:
I’ll start by saying that this is a good paper, and an interesting paper. I think it tells a story, and I have few problems with analysis or big picture design choices, at least as it’s set out now.

I’ll also be honest that I struggle with it in a few ways, also at the big picture. I’ll list out a number of these questions and critiques, as I think there’s still some good questions here that can perhaps simply be answered more fully with your data.

I’ll also admit that while I often feel fairly clear about where a paper should end up in terms of bins such as ‘minor revisions’, ‘major revisions’, etc., I’m struggling a bit more to pin this one down as well. Again, I think there’s some good content here, but I also have some real problems. The fact that I struggle to see where this paper needs to go exactly for it to be better doesn’t mean that place doesn’t exist. Hopefully the editor has the foresight to see that place with the help of these comments.

Anyway, you likely care a lot less about this actual hedging and setup than my actual comments, so here goes.

1) Big picture, maybe the biggest picture comment I have. As I finished reading the piece and have now thought about it a few days, I struggle as a reader to clearly elucidate a clear answer to the question ‘what is the point of this paper?’ Do not take this the wrong way – I’m not saying that this paper is pointless. What I’m saying is that I think this paper tries to have a number of points, but in doing so struggles to make any of them totally clear. Potential points flirt at the edges of my peripheral vision, only to escape when confronted as the point.

I review a lot of papers on IER and carelessness, and also on lots of methodological questions. I’m not saying that to sound like an expert, I’m saying that to make clear that it likely colors my opinions across the different points of the paper. The discussion of carelessness is also likely one of the main factors as to why I’m a reviewer on this paper. Again, I may thus be looking for more in that section than might otherwise be the case.

Whenever I read an intro that starts to talk about IER, I start mapping out in my head where it might be going. My thoughts here were that this would be something that you measure, perhaps something that mediates the relationship you eventually get to, between design and reliability. Don’t get me wrong, I’d have some other issues if this had been that mediation (which I’ll put out there if that’s where this paper ends up going), but it’s hard to read the introduction and then wonder why carelessness isn’t in this model somewhere.

The question of whether or not some of these design choices impacts carelessness is still a fairly open question. Some of these concepts have been examined to some degree, but I was still looking forward to some of your findings had those been the analyses in question.

In terms of these choices impacting simple reliability (yes, I know you did some complex reliability analysis, but it’s still consistency at the end of the day), well, it feels like that’s something that is an older question that has been done before, to some degree. Has it been done in exactly this way with exactly this design? Probably not. But many of the papers you cite in the introduction are great pieces from the 80s and 90s where a lot of these questions were being asked. If the DV here is simply reliability, I think you need to make a better case as to why that’s new or innovative, or what this paper tells us that we don’t already know.

Let me make a concrete case for some of this. The first two sentences of your abstract are: “Innovative techniques have been designed to delete unreliable self-report survey responses post hoc. This study examines the a priori impact of three survey design and implementation tactics on score reliability.” I put forth that the second sentence adequately summarizes your paper, while the first sentence is a bit of a red herring.

Does carelessness have a place here? Sure, it could. You make maybe 70% of the argument for it, but then you don’t follow through in actually using metrics (which you clarify are calculable post-hoc) as a variable.

That said, the one way I could see this tying together in a semi-coherent way would be if you had conceptualized your individual reliability gained from your GPC model (there’s a point about this model I’ll bring up in a later point) as a measure of individual inconsistency, aka, carelessness. In fact, the fact that you’re fitting an IRT to this type of data would allow you to calculate one of the gold standards of carelessness, lzp (apologies for formatting, not sure how super/subscripts will come through the reviewer portal).

If you had done that, and continued this line of reasoning throughout, then that ‘reliability’ would actually be carelessness, and the whole concept of the paper would shift to ‘design choices -> carelessness’ instead of ‘design choices -> reliability’ (or design choices -> carelessness -> reliability).

The fact that you didn’t, though, leaves the path through your intro a bit more complex and hard to predict. At the end of the day, they sort of just don’t match.

Answer the question I started with. The point of this paper is \_\_\_\_\_\_\_\_. With that in mind, consider how that actual point informs the choice of focus and flow throughout the paper.

Sorry, that was all one point, I guess. I said it was big picture.

2) Alluded to in the last point, but not clearly highlighted, is your choice of which careless metrics you identify. Again, this is a bit of a lightning rod for me, and it’s not an uncommon question in this literature to be asked ‘which of these metrics should I use?’

That said, I’m not sure I’d ever likely come across a recommendation to use only the three you came up with. While I understand you’re describing things that are post-hoc calculable, there’s a lot that still fit that case.

I’d have to go back and read my Desimone, Harms, & Desimone, but first off I don’t believe that odd-even consistency is resampled, at least not classically. If it was, it would no longer be odd-even. There is a similar technique in Curran (2016) – which might also be what you’re seeing in Desimone – that does resample this, but it’s not odd-even reliability at that point. Perhaps you mean this other type, but it’s not fair to say that odd-even reliability is ‘sometimes referred to as’ something else that also exists.

While person-total correlations do also exist in this literature, you’re missing citations for this, just to be fair across all three of these. There are cites from the 60s or 70s, as well as some of the papers you already cite in other places.

That leaves ISDs. I want to say this as carefully as possible, as I can’t necessarily point at citations that show the points I’ll make, but ISDs are a terrifically unproven technique when it comes to carelessness. The only lab to have ever shown any use to it is the lab you cite, and their work doesn’t necessarily make the same case that you do here. You can also mathematically show that ISDs shouldn’t have a linear relationship as a proxy for carelessness, at least not in modern conceptualizations of the idea of carelessness.

Am I saying you shouldn’t talk about ISDs? No. I’m just saying that if you’re only going to use three things, there are better you left off the list, such as long-string analysis (also in many of the papers you cite). Frankly, again, there’s also lzp, the absolute gold standard of carelessness that most people don’t use because they can’t/haven’t fit a polytomous IRT model to their data in order to calculate it. That (lzp) seems to be a missing link in a lot of the flow of this paper, perhaps.

3) One more point on carelessness (no guarantees). Just as a big picture idea, my skepticism immediately turns on high when I see a paper talking about carelessness and reliability. While researchers in the past have conflated the two at times, it’s very clear in modern work that carelessness has a near zero relationship with reliability, as some types of carelessness are negatively related while others are positively related. I’m not saying you fell into this trap here, but I’m just pointing out that you’re walking around the perimeter of a hole that you might stumble into on any revision.

4) I worry that this data was more conveniently sourced than might be generally the case. Not only that it comes from only one type of class, but that it came so decidedly from that class that it was collected on paper and pencil, presumably in that class, presumably then under duress.

Perhaps you have a problem with my wording on that last part. If this data was collected in a classroom where students were already present, then it was collected under duress. Informed consent gives the right to not participate in a study, but if it’s as part of a class, or even during class time, you’ve leveraged social pressures of both classroom environment, peers, and social norms against your subjects.

Were your subjects unduly coerced? You got 98% response rate, so, uh, yeah. In many of my classes I don’t get 98% response rate on exams.

Any revision of this paper, should one be requested, should have a much deeper description of how this data was collected, why it was collected that way, what steps were taken to minimize the problems I just mentioned, and potential risks not to just participants just also to internal validity of the study, etc. Make a case as to why I should be less concerned than I am.

5) As an extension of this last point, I also worry that this data was also conveniently sourced to this researcher. That is, I worry that this paper currently under review was not the plan for the first or primary use of this data. If it was the case, I’m not sure why the researcher would have settled on only three scales, or three scales of different lengths, or three scales from different places as to create more differences between the scales.

I’m not sure that’s a fatal flaw, but certainly something that I’d like to hear more about regarding if this data had a primary purpose that has already been examined, if this data is essentially archival, or something else.

6) When you’re holding a hammer, everything looks like a nail, and right now I’m focused on data collection. The fully in-person part is also weird. If you could fully cross this design one more time with the same design but with online collection, do you think any of your findings would hold or change? I surmise the answer is unquestionably yes.

Again, this goes back to the last point. This is pretty conveniently sourced cross-sectional data from a cross section that is not the norm. What does this tell a researcher that predominantly does online work (or even just computerized work)? Who is the audience for this paper? I would put out there that I probably am, but not when the finding only applies to in-person data collection.

This is a big place where a fully-crossed design could easily be created, though, as online data collection is pretty easy. I’m not going to say that’s a necessity, but I’ll put that idea into the editor’s head. Maybe this is enough even fully in person, but it does leave a lot on the table.

7) While you perhaps have a case for calculating standard deviations for your ‘individual’ reliability from IRT, you have a much thinner case for being able to calculate the same for CTT. Yes, you have citations, and yes, these cases have been argued. My concern here is that many more assumptions have to fit into place for these sorts of calculations to work out, and so the results are cast in much more questionable light. I get that there just aren’t perfect solutions here, but this is just a place that gets tricky when you’re trying to use group statistics as a DV.

There are other options - resampling would be one. I’m also not going to push any of those here without support of the editor, but just highlighting that saying that alpha has gone up or down ‘significantly’ is not a non-trivial problem to wrestle with when your sample size is essentially 1.

8) You use a GPC here, and my curiosity is to why you didn’t use a GRM. This is polytomous personality data, not ability score data. It’s a small point, but I do believe a GRM would be the (only slightly different) better choice, unless you have some other reason for a GPC. It’s been a while since I ran either though, so feel free to make a case for GPC over GRM.

9) Allow me a moment to pull out a soapbox and then to stand upon it.

I have no reason to believe that you’re someone who does not understand statistics. I also have no reason to believe that you’re actively trying to fabricate your data or your results. I don’t assume that anyone is actively p-hacking or data-picking or any other sort of questionable research practice.

That said, you’re sort of standing up waving a flag warning that something is wrong with the inclusion of concepts like p<.05.

Now, are you wrong? Is your p-value in fact lower than .05? It probably is, and you probably aren’t, at least technically.

I don’t know what style guideline you’re following. APA, I assume, but this could be Chicago, or MLA, or Honolulu, or Deimos style. In fact, a new APA just came out, so maybe it’s even possible that it’s in that and they’re wrong on this point. Or, that they’re right in a way that escapes me and I’m deeply wrong, and if so, I apologize.

>
Update: I’ve checked in the new APA guidelines, and here is the relevant part:

“When reporting p values, report exact p values (e.g., p = .031) to two or three decimal places. However, report p values less than .001 as p < .001. The tradition of reporting p values in the form p < .10, p < .05, p < .01, and so forth, was appropriate in a time when only limited tables of critical values were available.” (p. 114)

Could I have just said this to start? Yes. I still could. For what it’s worth, I’d rather hit this point home.
>

You’re not in 1964, and I assume you’re not going to your university supercomputer with punch cards to calculate your statistics, or sitting up under candlelight to calculate p-values by hand using a slide rule or a pile of beans. You can know what all of these p-values are, and you demonstrate that you’re a competent statistician in many other regards in this paper. In fact, some of these are t-tests! I could find these p-values, by hand, with even just the data presented.

I’m not saying you’re a bad person, and don’t take this the wrong way. Maybe this is an honest mistake, or a simple point that has never been adequately made to you. Regardless of how it came to be, the simple fact is that artificially categorizing p-values into multiple categories and then not reporting the actual values is a questionable research practice in today’s world.

Perhaps this is the first time a reviewer (or any other colleague) has really dressed you down on this, so I’ll make sure I make it count so that only this one time is needed.

I’ll give you that many statistical programs, like SPSS, give you a whole bunch of stars for these different levels (something you also do – please don’t). They don’t do it for the purposes of a single researcher (or maybe they do – SPSS does a lot of dumb things). They do it because many different researchers use their software. In fact, maybe a researcher is in medicine and has decided to set their alpha level (from a power standpoint) at .01, instead of .05. Cool, now they have their own stars to look at (\*\*) while you’re over there looking at yours (\*). Having multiple stars for p-value thresholds is something that statistical programs do because there is between-researcher variance on chosen alpha level, and not because there is within-researcher variance on p-levels.

If you accidentally or purposefully carry over these many categories into a single paper, you’ve now committed the fallacy of conflating p-values with statistical power. Now your reader will see those effects with lots of stars and think ‘wow, those are the big ones’.

Perhaps they are. Holding n constant, as is the case in your study, p and power are linearly related. That’s not the point. The point is that this is not how p-values are supposed to work.

Beyond this, the only reason for someone to say p<.05 in this day and age, where anyone has access to p-values out to any number of decimal places, is to hide the fact that p=.047, or something of that sort. You might know that as part of the broader concept of p-hacking, and if you didn’t, now you do.

What you’re doing is (maybe) not p-hacking, but at the very least it’s p-mishandling.

Again, you can cite any style guide you want to justify any method you’re using, and at the end of the day I simply don’t care. There’s a (methodologically) right way to do this, and a (methodologically) wrong way to do this, and right now you’re on the wrong side of that line.

I’m not someone out to destroy p-values or their use. In fact, I often find myself defending p-values. Using p-values in this way makes them harder to defend, as this is the misuse that advocates for their destruction point to.

Am I being a bit mean here? Perhaps. I’ll let the editor chide me – privately or in the letter back to you – if they believe I’m being too harsh here. If I’m coming off harsher than you might expect a reviewer to, it’s because I want you to remember this point long after you have a decision either way on this paper. I want you to remember this and take this lesson to heart.

It’s an easy fix, and a small one. I’ve put much more in this review about it than is warranted, by any reasonable stretch of the imagination. It’s why it’s point whatever it is instead of point 1.

I’ll get off my soapbox now, and thank you for hearing me out if you’ve read this whole diatribe. Nothing against you personally, and I hope you take it as constructive, not destructive, as with all my points.

10) As a follow up to the last point, you should also be reporting effect sizes where they’re easy to calculate. In modern times it’s borderline unacceptable for a t-test to be reported, for instance, without a Cohen’s d. Another very simple fix that would strengthen this reporting.

11) I’ll point out to the author and editor that I think there enough points above that this paper needs at the bare minimum some revisions. As such, I’ve not looked as hard as I otherwise would at the actual data or tables/charts. If this paper does move into R&R, I would at that time dig more into those areas, and thus may have comments/concerns at that point that I don’t have at the moment.

>
Okay. Those are the big things, and I think it’s a fairly non-trivial list. At the end of it, though, I’m very much in the place I was at the beginning. While I’ll support whatever decision the editor comes to, I’m not sure I’d wager much on guessing what it might be. I do think there’s good stuff in here (despite how much I might have belabored this point or that). I do believe I’m always happier to see a paper undergo revisions than be rejected outright, but again I’ll leave that to the editor.

2) Figures/tables/data availability:
Please comment on the author’s use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:
As noted above, I've made a cursory pass at these and found no large problems, but if issues I was concerned with were revised I'd dig deeper into this specific point.

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.:
As noted above, I'd like to see a little more about the particulars of this collection. I don't have any reason to believe that anything was amiss, but it could be described more deeply to make that clearer.

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:
Yes, well written.

Auhtor response

June 27, 2020

Dear Dr. Donnellan,
My co-author, Dr. Marcia Simmering of Louisiana Tech University, and I thank you very much for the opportunity to revise and resubmit our paper to Collabra. Consistent with the characterization of this opportunity as a major revision we have implemented a major overhaul of it. To that end, as we state in our response to the reviewers, we have dropped the item response theory (IRT) analysis from the paper. Our research is about factors that survey designers implement to improve classical test theory (CTT) internal consistency reliability. As you no doubt are aware, IRT-based conditional reliability and CTT internal consistency are not the same thing. Just to clarify, a consistently low score or a consistently high score on an inventory can both show high internal consistency but the IRT conditional reliability is truly dependent or conditional on the magnitude of the score on the inventory and low scores are likely to have different conditional reliability than high scores. We hope you agree that this major change allows us to more finely focus our analysis and to adhere to the suggestions and comments of the reviewers. We also thank you for agreeing not to send our paper back to Reviewer B given his or her strong tone. However, we do respond appropriately to every comment and suggestion from both reviewers and you.

Please note that we have had a bit of a struggle to figure out where to submit our Response to Reviewers. Please advise us if we need to do something different but we have submitted our paper, this cover letter, and our response to reviewers to the submission site. We look forward to hearing back from you.

Best Regards,
Brian K. Miller

**RESPONSE TO REVIEWERS**

Note to the editor: We thank the editor and the reviewers for the insightful comments and suggestions which we address in a point-by-point manner below. We especially thank the editor for agreeing to not send our paper back to Reviewer B whose comments were a bit "strong". However, we respond to both reviewers' comments below. We think it prudent to state at the outset that we have completely discarded the item response theory (IRT) conditional reliability analysis from the paper and instead focus on the classical test theory (CTT) application of internal consistency reliability. Our experimental manipulations were designed to affect internal consistency and IRT conditional reliability is simply not the same thing as CTT internal consistency reliability.

1. Reviewer #2 raises important questions about the consent process of the study and nature of participants. Please clarify that issue for readers and address those questions.  IRB approval (or an exemption determination from your IRB) is required for publication.

Authors' response: We have added the appropriate verbiage to the manuscript. Our study was determined to be exempt and the exemption approval number 2658 is now included in the text on page 11.

2. I had questions about the design and whether you needed to use a more complicated model to test for interactions between conditions.  It seems as if there are 3 factors (Quality Control Items v. None; Scrambled v. Grouped Items; and Anonymity v. Confidentiality).  The randomization was at the person level for the first two factors but at the classroom level for the anonymity factor which could (theoretically) complicate dependencies in the data.  I am perhaps not too worried about the randomization level issue, but I do think testing interactions is worthwhile on an exploratory basis as in a 2x2x2 factorial. (And I might have missed something in the design so that could either be a main effect of my brain, the text description, or the interaction. If I missed something obvious, I apologize.)

Authors' response: Our original use of t-tests only allowed for main effect tests. We have now run the analysis using three-way factorial ANOVAs (with a different ANOVA for each instrument upon which we collected data and therefore calculated alpha reliability). To avoid the complicated computation by hand of factorial ANOVA with summary data we used SPSS syntax provided by David Nichols to simulate sample data based upon the group-level Cronbach's alpha and the standard deviation of alpha that we calculated using formulae from Hakstian and Whalen (1976). The cell sample sizes were the necessary third component of the summary data. Then we ran ANOVA on the generated/simulated data so we could test the three different two-way and one three-way interaction. The results of our analysis using simulated data based upon sample characteristics from our raw data were validated using a R program developed and run on our behalf by David Rindskopf. The Rindskopf program uses only the summary data. David's results correspond exactly to the results using syntax-generated simulated data. We hope this makes sense and we have described the process to simulate the data in the text and included an acknowledgement in our paper to Drs. Nichols and Rindskopf.

3. I think it would help readers if they knew that the three scales analyzed in the paper were the only data collected.  This should be noted in the paper and I think uploading a copy of each survey to the data repository would be useful to readers.

Authors' response: Thank you for this opportunity. We have added to page 11 the following text: "The instruments described below were the only scales in the survey and were selected because of their likely discriminant validity yet were not so dissimilar as to affect the effects induced by the scrambling versus grouping manipulation." Additionally, the actual scales themselves are now in the same data repository and include a code book with variable names.

4. Explaining to readers how the quality control items were spread out in the various forms would be helpful. This is especially relevant in the group item design.

Authors' response: We have added the following text to page 10: "One quality control item was embedded at the one quarter point into the survey, at the halfway point into the survey, and at the three-fourths point in the survey." For the grouped manipulation this placed the items about half way into each of the three scales in the survey.

5. Both reviewers wanted greater justification for why the current manipulations are expected to impact internal consistency/reliability.  I think this is an area of the paper that needs to be enhanced.  To be sure, we are not asking you to hypothesize after the results are known but rather to flesh out the rationale for the predictions, so readers can evaluate the logic and how these results fit with the existing literature.  I could also see value in looking at means and SDs for the scales and scale inter-correlations on an exploratory basis.

Authors' response: We hope that our justification for the manipulations and the study in general are more clear now that we focus exclusively on internal consistency reliability. We got a bit off track with our IRT-based calculations of conditional reliability (which is decidedly not the same as CTT-based internal consistency reliability) and our discussion of both careless responding and insufficient effort responding (IER). We apologize for that mix-up in our original submission but thanks in large part to the requirements of a more fine-tuned focus on the purpose of the paper we now see that our original analytic plan was misguided. With that in mind, we hope that our much more focused attention on internal consistency reliability offsets any concerns of the editor.

6. Reviewer #2 was especially interested in seeing greater justification for the specific DVs used in the paper.

Authors' response: We spent ample time reading and re-reading Reviewer 2's lengthy comments and feel much better about our DVs now. The DVs are the internal consistency reliability of scores on three commonly used personality scales. As the editor no doubt knows, alpha reliability is only one number which is meant to encapsulate the reliability of scores for an entire group of respondents. It is almost impossible to surmise that every respondent in a sample responds with the same level of internal consistency (which is why we no longer use IRT conditional reliability which is purely a function of trait level or magnitude and differs between scores). Thus, we use Hakstian and Whalen's (1976) formulae to estimate the standard deviation of alpha. Then when we combine alpha (or mean reliability for a group), the standard deviation of alpha, and the sample size we can run analysis based on these summary statistics. However, calculating factorial ANOVA by hand based upon on summary data is rather complicated. Thus, we used the aforementioned SPSS syntax and validated those results with Rindskopf's R analysis.

7. I agree with the general sentiment expressed by Reviewer #2 (and perhaps implicit in reactions from Reviewer #1) about the need to clarify the point of the paper for readers and highlight the main take home messages and implications.

Authors' response: We thank both reviewers for helping us clarify our points. Without them, our paper was indeed a bit of a mess.

8. Reviewer #2 had a lengthy set of concerns about reporting exact p-values.  I think adding these values will enhance the paper. I believe the reviewer’s intent was constructive, but this issue struck a chord with the reviewer as you will read.  I will be honest and note that some of the passages surrounding this issue were perhaps too strongly worded for my taste.  However, I saw them in the context of the whole review which was lengthy but provided in the spirit of helping to enhance the work.  In other words, the language in that section might be pointed but I think you can report the exact values and address the concern.  I took the reviewer as providing valuable reactions even if that section might have been intense.

Authors' response: We thank the editor for acknowledging the tone of Reviewer 2. However, we politely and succinctly addressed his concerns and must note that we were abiding by the sixth edition of the APA manual and the submission guidelines of the journal. Reviewer 2 was suggesting that we abide by the seventh edition of the manual. We have now done exactly that. It was an easy fix.

9. I also think reporting effect sizes for the t-tests would be useful.

Authors' response: We now include effect sizes (partial eta squared) for every manipulation and interaction resulting from the factorial ANOVA.

10. I thought there was some redundancy between the text and the tables in terms of reporting the same values for the tests in both places.  I find tables easier to read so I don’t think it is necessary to have complete redundancy between the text and tables.  This could streamline some sections.  I would rather see effect size estimates reported and discussed.

Authors' response: Because we now employ factorial ANOVA the effect sizes are in each ANOVA table and briefly summarized in the Results section of our paper. Our in-text results of our new analysis now summarize the results with verbiage only but the full results are in each (new) ANOVA table. In addition to this change we have provided the marginal means for alpha on the three scales in the main effect test results. For example, on page 14 we now state, "Alpha in the anonymous condition for scores on the conscientiousness scale, entitlement scale, and work ethic scale was .808, .786, and .881, respectively. In the confidential condition the alpha was .761, .828, and .856 for these scales, respectively." We think this is an important addition because despite two of the three manipulations resulting in significant F-scores in some instances the difference was significant but not in the direction which was hypothesized and therefore not technically supportive of our hypotheses.

11. Those were the major issues that I think require attention.  I think clarifying the rationale, the nature of the data, and the major points will enhance the work.  I hope you will agree and undertake a revision.  You might also make sure the data file is usable by others by providing a short code book and including a .csv version so those without SPSS can access the data.

Authors' response: We thank the editor very much for this opportunity. The editor's requirements and the comments and suggestions of both reviewers have undoubtedly made this a much better and more finely focused paper. In addition to replacing the SPSS data with .csv data we have also included copies of the scales themselves which also serve as a code book. Thanks!

**REVIEWER A COMMENTS**

This manuscript reports an experimental study examining the influence of three survey design features on scale reliability. I agree with the importance of studying IER, and I see the merit in the use of the experimental approach. I hope my comments below may be useful as you further develop this paper.

Authors' response: We thank the reviewer for the comments below.

1. The greatest challenge I had with the hypotheses of this paper is the assumption that “Pure content non-responsivity is completely random and results in a reliability of scores equal to zero”. Yet insufficient effort responding may not manifest in a true random fashion and may instead show up as repeated selection of the same response. When this occurs, Cronbach’s alpha may appear to increase instead of decrease (DeSimone, DeSimone, Harms & Wood, 2018). Thus, although I agree with the notion that surveys containing quality control items and confidential surveys may lead to lower IER, I am not sure one should expect higher reliability as a result. This issue may have contributed to the curious pattern of results found in the study.

Authors' response: We have discarded mention of IER in the paper except as an area of study that sometimes entails reliability assessment. The study of IER is a complex one and likely well-beyond the scope of our paper.

2. Another question I had is the presence or absence of reversed coded items. One of the three measures (conscientiousness) had negatively keyed items, while the other two did not. Negatively keyed items can introduce method factor and presumably decrease reliability, if IER occurs (Schmitt & Stults, 1985).

Authors' response: We thank the reviewer for this. We had previously cited Schmitt and Stults (1985) and now added this in-text citation to our overview of the EFA results for conscientiousness that revealed a two-factor structure best explains the variance in scores on the scale.

3. One question I am curious about is whether IER may manifest as in reduced attention to the survey question, but instead of being nonresponsive to the questions, respondents simply determine if each question is desirable or undesirable, and respond favorably to desirable items and unfavorably to undesirable items. In other words, IER may result in social desirable responding. If (that is a big if) this happens, the hypotheses may become more complicated.

Authors' response: As noted above and in an effort to not overly complicate our study (which we think has some beauty in its simplicity) we no longer use IER as a framework for analysis or explanation.

4. The exploratory factor analysis on conscientiousness circles back to my point #2 above. The presence of IER may have introduced the method factor for conscientiousness.

Authors' response: We agree. See our response to points 2 and 3 above.

5. Please report effect sizes for t tests.

Authors' response: Because we now use a full-fledged factorial ANOVA, we now report the effect sizes for each factor as partial eta-squared, summarized in the text and reported in detail in the three newly added ANOVA tables.

5. Figures/tables/data availability: Use of tables is adequate. Underlying data is available.

Authors' response: Thank you.

6. Ethical approval: If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared. If not, please detail where you think a further ethics approval/statement/follow-up is required. I could not find a statement about ethical approval by an IRB.

Authors' response: Our study was approved as exempt by our IRB and we now include the application number and the appropriate statement to that regard in the text.

4) Language: Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.: The text is generally well written.

Authors' response: Thank you for this comment and the others. We hope that we have properly addressed your concerns.

**REVIEWER B COMMENTS**

I’ll start by saying that this is a good paper, and an interesting paper. I think it tells a story, and I have few problems with analysis or big picture design choices, at least as it’s set out now. I’ll also be honest that I struggle with it in a few ways, also at the big picture. I’ll list out a number of these questions and critiques, as I think there’s still some good questions here that can perhaps simply be answered more fully with your data. I’ll also admit that while I often feel fairly clear about where a paper should end up in terms of bins such as ‘minor revisions’, ‘major revisions’, etc., I’m struggling a bit more to pin this one down as well. Again, I think there’s some good content here, but I also have some real problems. The fact that I struggle to see where this paper needs to go exactly for it to be better doesn’t mean that place doesn’t exist. Hopefully the editor has the foresight to see that place with the help of these comments. Anyway, you likely care a lot less about this actual hedging and setup than my actual comments, so here goes.

Authors' response: We thank the reviewer very much for the comments above and below and agree that the big picture painted initially by us was a bit out of focus. Our focus is on internal consistency reliability instead of IRT-based conditional reliability. The two forms of reliability are not the same and our experimental manipulations were designed to affect internal consistency. This entailed major revisions and we hope that we adequately address each comment from the reviewer below. In that vein, the IRT analysis is no longer part of our paper.

1a. Big picture, maybe the biggest picture comment I have. As I finished reading the piece and have now thought about it a few days, I struggle as a reader to clearly elucidate a clear answer to the question ‘what is the point of this paper?’ Do not take this the wrong way – I’m not saying that this paper is pointless. What I’m saying is that I think this paper tries to have a number of points, but in doing so struggles to make any of them totally clear. Potential points flirt at the edges of my peripheral vision, only to escape when confronted as the point.

Authors' response: We agree that we tried to do too many things with this paper. As we note above, the main intent of our experimental manipulations was to affect internal consistency reliability. Our focus is now solely on that. To that end (in addition to the IRT analysis) we have discarded the concept of IER, which is a much more complicated concept than our paper can address.

1b. I review a lot of papers on IER and carelessness, and also on lots of methodological questions. I’m not saying that to sound like an expert, I’m saying that to make clear that it likely colors my opinions across the different points of the paper. The discussion of carelessness is also likely one of the main factors as to why I’m a reviewer on this paper. Again, I may thus be looking for more in that section than might otherwise be the case.

Authors' response: We thank the reviewer for their expertise.

1c. Whenever I read an intro that starts to talk about IER, I start mapping out in my head where it might be going. My thoughts here were that this would be something that you measure, perhaps something that mediates the relationship you eventually get to, between design and reliability. Don’t get me wrong, I’d have some other issues if this had been that mediation (which I’ll put out there if that’s where this paper ends up going), but it’s hard to read the introduction and then wonder why carelessness isn’t in this model somewhere.

Authors' response: As we note above, we no longer use IER as a framework for understanding responses to our manipulations.

1d. The question of whether or not some of these design choices impacts carelessness is still a fairly open question. Some of these concepts have been examined to some degree, but I was still looking forward to some of your findings had those been the analyses in question.

Authors' response: Our design choices most properly address internal consistency reliability and respondents' motives for providing internally consistent responses are beyond the scope of our paper.

1e. In terms of these choices impacting simple reliability (yes, I know you did some complex reliability analysis, but it’s still consistency at the end of the day), well, it feels like that’s something that is an older question that has been done before, to some degree. Has it been done in exactly this way with exactly this design? Probably not. But many of the papers you cite in the introduction are great pieces from the 80s and 90s where a lot of these questions were being asked. If the DV here is simply reliability, I think you need to make a better case as to why that’s new or innovative, or what this paper tells us that we don’t already know.

Authors' response: We thank the reviewer for helping us realize that, as the reviewer says, "...it's still consistency at the end of the day". We agree and because of that we now focus on Cronbach's alpha. As we noted in our paper, ours is the first that we know of to implement and test three design differences at the same time as well as the first to statistically test them rather than an "eyeball" test of differences in alpha for such machinations.

1f. Let me make a concrete case for some of this. The first two sentences of your abstract are: “Innovative techniques have been designed to delete unreliable self-report survey responses post hoc.  This study examines the a priori impact of three survey design and implementation tactics on score reliability.” I put forth that the second sentence adequately summarizes your paper, while the first sentence is a bit of a red herring.

Authors' response: We agree and have now deleted the first sentence. Our initial purpose in including it in the abstract was to contrast post hoc and a priori techniques for deleting/detecting/affecting unreliable responses. Thank you.

1g. Does carelessness have a place here? Sure, it could. You make maybe 70% of the argument for it, but then you don’t follow through in actually using metrics (which you clarify are calculable post-hoc) as a variable.

Authors' response: We have now excised mention of careless responding from the paper. Despite the use of carelessness metrics by others we believe that focusing on a more narrowly tailored story helps us more properly define our big picture, an issue with which the reviewer expressed some concern above.

1h. That said, the one way I could see this tying together in a semi-coherent way would be if you had conceptualized your individual reliability gained from your GPC model (there’s a point about this model I’ll bring up in a later point) as a measure of individual inconsistency, aka, carelessness. In fact, the fact that you’re fitting an IRT to this type of data would allow you to calculate one of the gold standards of carelessness, lzp (apologies for formatting, not sure how super/subscripts will come through the reviewer portal).

Authors' response: We agree that a graded response model may have been more appropriate for the data here. However, the IRT-based calculation of conditional reliability was developed based upon a graded partial credit model. We thought it beyond the scope of our paper to alter the initial purpose of the IRT calculations as they were developed for our experiment. All of this is moot now as we have eliminated the IRT analysis from the paper.

1i. If you had done that, and continued this line of reasoning throughout, then that ‘reliability’ would actually be carelessness, and the whole concept of the paper would shift to ‘design choices -> carelessness’ instead of ‘design choices -> reliability’ (or design choices -> carelessness -> reliability). The fact that you didn’t, though, leaves the path through your intro a bit more complex and hard to predict. At the end of the day, they sort of just don’t match. Answer the question I started with. The point of this paper is \_\_\_\_\_\_\_\_. With that in mind, consider how that actual point informs the choice of focus and flow throughout the paper. Sorry, that was all one point, I guess. I said it was big picture.

Authors' response: The answer to the reviewer's concern is "Our point is to examine the impact of three survey design and administration techniques that can be implemented a priori on the internal consistency reliability of commonly used personality scales." We thank the reviewer for helping us to focus on the true purpose of our paper.

2. Alluded to in the last point, but not clearly highlighted, is your choice of which careless metrics you identify. Again, this is a bit of a lightning rod for me, and it’s not an uncommon question in this literature to be asked ‘which of these metrics should I use?’ That said, I’m not sure I’d ever likely come across a recommendation to use only the three you came up with. While I understand you’re describing things that are post-hoc calculable, there’s a lot that still fit that case. I’d have to go back and read my Desimone, Harms, & Desimone, but first off I don’t believe that odd-even consistency is resampled, at least not classically. If it was, it would no longer be odd-even. There is a similar technique in Curran (2016) – which might also be what you’re seeing in Desimone – that does resample this, but it’s not odd-even reliability at that point. Perhaps you mean this other type, but it’s not fair to say that odd-even reliability is ‘sometimes referred to as’ something else that also exists. While person-total correlations do also exist in this literature, you’re missing citations for this, just to be fair across all three of these. There are cites from the 60s or 70s, as well as some of the papers you already cite in other places. That leaves ISDs. I want to say this as carefully as possible, as I can’t necessarily point at citations that show the points I’ll make, but ISDs are a terrifically unproven technique when it comes to carelessness. The only lab to have ever shown any use to it is the lab you cite, and their work doesn’t necessarily make the same case that you do here. You can also mathematically show that ISDs shouldn’t have a linear relationship as a proxy for carelessness, at least not in modern conceptualizations of the idea of carelessness. Am I saying you shouldn’t talk about ISDs? No. I’m just saying that if you’re only going to use three things, there are better you left off the list, such as long-string analysis (also in many of the papers you cite). Frankly, again, there’s also lzp, the absolute gold standard of carelessness that most people don’t use because they can’t/haven’t fit a polytomous IRT model to their data in order to calculate it. That (lzp) seems to be a missing link in a lot of the flow of this paper, perhaps.

Authors' response: As we note above, our focus in not on carelessness nor IER per se but rather it is on alpha reliability. Granted, carelessness and IER can affect alpha reliability but such constructs are beyond the focus of our paper. Our renewed focus would probably not have been possible without the comments of the reviewer, whom we thank very much!

3. One more point on carelessness (no guarantees). Just as a big picture idea, my skepticism immediately turns on high when I see a paper talking about carelessness and reliability. While researchers in the past have conflated the two at times, it’s very clear in modern work that carelessness has a near zero relationship with reliability, as some types of carelessness are negatively related while others are positively related. I’m not saying you fell into this trap here, but I’m just pointing out that you’re walking around the perimeter of a hole that you might stumble into on any revision.

Authors' response: Thank you for this information.

4. I worry that this data was more conveniently sourced than might be generally the case. Not only that it comes from only one type of class, but that it came so decidedly from that class that it was collected on paper and pencil, presumably in that class, presumably then under duress. Perhaps you have a problem with my wording on that last part. If this data was collected in a classroom where students were already present, then it was collected under duress. Informed consent gives the right to not participate in a study, but if it’s as part of a class, or even during class time, you’ve leveraged social pressures of both classroom environment, peers, and social norms against your subjects. Were your subjects unduly coerced? You got 98% response rate, so, uh, yeah. In many of my classes I don’t get 98% response rate on exams. Any revision of this paper, should one be requested, should have a much deeper description of how this data was collected, why it was collected that way, what steps were taken to minimize the problems I just mentioned, and potential risks not to just participants just also to internal validity of the study, etc. Make a case as to why I should be less concerned than I am.

Authors' response: We have now included on page 9 the following: "The data came from one course in two very large lecture sections taught by the same professor. The in-class survey was announced two weeks ahead of time as well a one-week reminder by the professor that extra credit could be earned by participating. An alternative extra credit assignment was also offered on both announcement occasions. That assignment was to find an article published by that professor and type the abstract and send it as an email to that professor. No student chose that option." The 98% response rate was not actually 98% of all possible students in the two sections. It was the percentage of students who provided complete data. We are not sure how else to offset any coercion concerns of the reviewer but would be pleased to address it further in a revision if one is offered to us.

5. As an extension of this last point, I also worry that this data was also conveniently sourced to this researcher. That is, I worry that this paper currently under review was not the plan for the first or primary use of this data. If it was the case, I’m not sure why the researcher would have settled on only three scales, or three scales of different lengths, or three scales from different places as to create more differences between the scales. I’m not sure that’s a fatal flaw, but certainly something that I’d like to hear more about regarding if this data had a primary purpose that has already been examined, if this data is essentially archival, or something else.

Authors' response: This data examined here were for the purpose of this very study. The study was submitted for funding as an internal grant at the first author's university. The grant was not funded but the study proceeded without funding as the grant application was for supplemental summer salary not for any expenses. The three scales used were selected because they have been used scores of times in published studies and are of general interest to the first author. The fact that they are only moderately correlated ranging from r = -.23 to r = .33 provides some evidence of discriminant validity. Because they are related is a plus in our opinion. To have scales measuring decidedly different constructs like, say, preference for designer clothing, recollections of childhood trauma, and current romantic partners would have added a layer of unnecessary complexity to our efforts. In sum, respondents would clearly have seen that such hypothetical items as very different and might have confounded the findings when the items were scrambled instead of grouped. In the grouped condition, such items would have likely yielded artificially inflated alphas because of their similarity to each other and dissimilarity to items in the other scales. We admit that our actual manipulation using the scales we chose may have masked some of the effects and have added the following text to page 18:

"The bootstrapped confidence intervals for the three scale score correlations overlapped for both conditions, indicating no significant difference whether the items were grouped or scrambled. It is notable that the correlation between entitlement and work ethic was non-significant in both conditions. The lack of difference between the conditions among these relationships suggests either that (a) respondents were adept at changing their focus in order to provide similar responses to items from the same scale whether grouped or scrambled, or (b) the nature of the items was so similar that it did not matter that they were scrambled or grouped. Both of these suggests a potentially weak manipulation. On the other hand, the non-significant correlation between entitlement scores and work ethic scores in both conditions suggests ample discriminant validity exists between scores on the scales."

We hope that is explanation addresses the reviewer's concerns.

6. When you’re holding a hammer, everything looks like a nail, and right now I’m focused on data collection. The fully in-person part is also weird. If you could fully cross this design one more time with the same design but with online collection, do you think any of your findings would hold or change? I surmise the answer is unquestionably yes. Again, this goes back to the last point. This is pretty conveniently sourced cross-sectional data from a cross section that is not the norm. What does this tell a researcher that predominantly does online work (or even just computerized work)? Who is the audience for this paper? I would put out there that I probably am, but not when the finding only applies to in-person data collection. This is a big place where a fully-crossed design could easily be created, though, as online data collection is pretty easy. I’m not going to say that’s a necessity, but I’ll put that idea into the editor’s head. Maybe this is enough even fully in person, but it does leave a lot on the table.

Authors' response: It is our experience that online surveys sometimes yield low response rates. By announcing the survey twice in advance of its implementation we engendered as many responses as we think possible. If another variable was introduced (e.g. an online survey versus the face-to-face survey) would be interesting indeed as the experiment expands to 2x2x2x2. Of course, there are even more manipulations that could be implemented as well. We think the simplicity of our design and the control we had over the manipulations add value to the paper not detract from it. We think it likely that an online survey would yield less reliable data perhaps because of the simplicity of clicking on radio buttons instead of actually writing a number on paper. Bear in mind that such a manipulation as an online administration could be contaminated by careless responding which is a concept no longer of focus in our paper. We thank the reviewer for this suggestion however.

7. While you perhaps have a case for calculating standard deviations for your ‘individual’ reliability from IRT, you have a much thinner case for being able to calculate the same for CTT. Yes, you have citations, and yes, these cases have been argued. My concern here is that many more assumptions have to fit into place for these sorts of calculations to work out, and so the results are cast in much more questionable light. I get that there just aren’t perfect solutions here, but this is just a place that gets tricky when you’re trying to use group statistics as a DV. There are other options - resampling would be one. I’m also not going to push any of those here without support of the editor, but just highlighting that saying that alpha has gone up or down ‘significantly’ is not a non-trivial problem to wrestle with when your sample size is essentially 1.

Authors' response: We agree that the IRT analysis would be easier to analyze because of the more accurate calculation of the standard deviation. Please note that we now employ a full-fledged factorial ANOVA that tests interactions. Computing such interactions by hand based on summary data is incredibly difficult. Therefore, we used some SPSS syntax to create sample data for the eight experimental groups (total n = 435) based upon alpha, the standard deviation of alpha, and the cell size of the eight groups/cells. This allowed us to use the simulated data with n = 435 in SPSS to run the ANOVA for each of the three scales. These results were validated by a colleague with a customized program in R that does so simply using summary data. In sum, the results of our syntax generated sample data conducted in SPSS and the R results were exactly the same. We have added an appropriate explanation of this to the text and an acknowledgement to both researchers who helped us.

8. You use a GPC here, and my curiosity is to why you didn’t use a GRM. This is polytomous personality data, not ability score data. It’s a small point, but I do believe a GRM would be the (only slightly different) better choice, unless you have some other reason for a GPC. It’s been a while since I ran either though, so feel free to make a case for GPC over GRM.

Authors' response: See our response to point 1h above. Also please note that we no longer do any analysis in IRT because IRT-based conditional reliability and CTT-based internal consistency reliability are not the same thing and our revised focus in the paper is on the latter.

9. Allow me a moment to pull out a soapbox and then to stand upon it. I have no reason to believe that you’re someone who does not understand statistics. I also have no reason to believe that you’re actively trying to fabricate your data or your results. I don’t assume that anyone is actively p-hacking or data-picking or any other sort of questionable research practice. That said, you’re sort of standing up waving a flag warning that something is wrong with the inclusion of concepts like p<.05. Now, are you wrong? Is your p-value in fact lower than .05? It probably is, and you probably aren’t, at least technically. I don’t know what style guideline you’re following. APA, I assume, but this could be Chicago, or MLA, or Honolulu, or Deimos style. In fact, a new APA just came out, so maybe it’s even possible that it’s in that and they’re wrong on this point. Or, that they’re right in a way that escapes me and I’m deeply wrong, and if so, I apologize. Update: I’ve checked in the new APA guidelines, and here is the relevant part: “When reporting p values, report exact p values (e.g., p = .031) to two or three decimal places. However, report p values less than .001 as p < .001. The tradition of reporting p values in the form p < .10, p < .05, p < .01, and so forth, was appropriate in a time when only limited tables of critical values were available.” (p. 114). Could I have just said this to start? Yes. I still could. For what it’s worth, I’d rather hit this point home. You’re not in 1964, and I assume you’re not going to your university supercomputer with punch cards to calculate your statistics, or sitting up under candlelight to calculate p-values by hand using a slide rule or a pile of beans. You can know what all of these p-values are, and you demonstrate that you’re a competent statistician in many other regards in this paper. In fact, some of these are t-tests! I could find these p-values, by hand, with even just the data presented. I’m not saying you’re a bad person, and don’t take this the wrong way. Maybe this is an honest mistake, or a simple point that has never been adequately made to you. Regardless of how it came to be, the simple fact is that artificially categorizing p-values into multiple categories and then not reporting the actual values is a questionable research practice in today’s world. Perhaps this is the first time a reviewer (or any other colleague) has really dressed you down on this, so I’ll make sure I make it count so that only this one time is needed. I’ll give you that many statistical programs, like SPSS, give you a whole bunch of stars for these different levels (something you also do – please don’t). They don’t do it for the purposes of a single researcher (or maybe they do – SPSS does a lot of dumb things). They do it because many different researchers use their software. In fact, maybe a researcher is in medicine and has decided to set their alpha level (from a power standpoint) at .01, instead of .05. Cool, now they have their own stars to look at (\*\*) while you’re over there looking at yours (\*). Having multiple stars for p-value thresholds is something that statistical programs do because there is between-researcher variance on chosen alpha level, and not because there is within-researcher variance on p-levels. If you accidentally or purposefully carry over these many categories into a single paper, you’ve now committed the fallacy of conflating p-values with statistical power. Now your reader will see those effects with lots of stars and think ‘wow, those are the big ones’. Perhaps they are. Holding n constant, as is the case in your study, p and power are linearly related. That’s not the point. The point is that this is not how p-values are supposed to work. Beyond this, the only reason for someone to say p<.05 in this day and age, where anyone has access to p-values out to any number of decimal places, is to hide the fact that p=.047, or something of that sort. You might know that as part of the broader concept of p-hacking, and if you didn’t, now you do. What you’re doing is (maybe) not p-hacking, but at the very least it’s p-mishandling. Again, you can cite any style guide you want to justify any method you’re using, and at the end of the day I simply don’t care. There’s a (methodologically) right way to do this, and a (methodologically) wrong way to do this, and right now you’re on the wrong side of that line. I’m not someone out to destroy p-values or their use. In fact, I often find myself defending p-values. Using p-values in this way makes them harder to defend, as this is the misuse that advocates for their destruction point to. Am I being a bit mean here? Perhaps. I’ll let the editor chide me – privately or in the letter back to you – if they believe I’m being too harsh here. If I’m coming off harsher than you might expect a reviewer to, it’s because I want you to remember this point long after you have a decision either way on this paper. I want you to remember this and take this lesson to heart. It’s an easy fix, and a small one. I’ve put much more in this review about it than is warranted, by any reasonable stretch of the imagination. It’s why it’s point whatever it is instead of point 1. I’ll get off my soapbox now, and thank you for hearing me out if you’ve read this whole diatribe. Nothing against you personally, and I hope you take it as constructive, not destructive, as with all my points.

Authors' response: We followed the submission guidelines of Collabra which requires embedded tables and a series of statements regarding funding information, data accessibility, etc. as well as APA formatting (sixth edition) of headings, citations, and references. Upon initial submission to Collabra of our paper, the seventh edition of the APA guidelines were not yet available. We thank the reviewer for bringing to our attention that exact p-values are now required by the APA in most reporting instances. Please understand that we were in no way trying to obfuscate our results, withhold any pertinent information, or otherwise disguise any misdeeds and apologize for implying their possibility.

10. As a follow up to the last point, you should also be reporting effect sizes where they’re easy to calculate. In modern times it’s borderline unacceptable for a t-test to be reported, for instance, without a Cohen’s d. Another very simple fix that would strengthen this reporting.

Authors' response: We thank the reviewer and because we now use factorial ANOVA we report partial eta-squared in all three ANOVA tables for each factor and in the text for each main effect on each personality scale.

11. I’ll point out to the author and editor that I think there enough points above that this paper needs at the bare minimum some revisions. As such, I’ve not looked as hard as I otherwise would at the actual data or tables/charts. If this paper does move into R&R, I would at that time dig more into those areas, and thus may have comments/concerns at that point that I don’t have at the moment. Okay. Those are the big things, and I think it’s a fairly non-trivial list. At the end of it, though, I’m very much in the place I was at the beginning. While I’ll support whatever decision the editor comes to, I’m not sure I’d wager much on guessing what it might be. I do think there’s good stuff in here (despite how much I might have belabored this point or that). I do believe I’m always happier to see a paper undergo revisions than be rejected outright, but again I’ll leave that to the editor.

Authors' response: We thank the reviewer for this and the other useful comments above.

12. Figures/tables/data availability: As noted above, I've made a cursory pass at these and found no large problems, but if issues I was concerned with were revised I'd dig deeper into this specific point.

Authors' response: We have removed existing tables and added five new ones. Any reader should be able to reproduce our results in our new Tables 2, 3, and 4 based solely upon the data in our new Table 1. Table 5 provides the scale correlations (not the correlation between alpha on the scales).

13. Ethical approval: As noted above, I'd like to see a little more about the particulars of this collection. I don't have any reason to believe that anything was amiss, but it could be described more deeply to make that clearer.

Authors' response: Please note our responses above and the changes noted above to the manuscript.

14. Language: Yes, well written.

Authors' response: Thank you!

Editor second decision—Revise & Resubmit

Sep 28, 2020

Dear Dr Brian Keith Miller,

Thank you for submitting your revised work to Collabra: Psychology. I sent the paper back to one of the previous reviewers and that review is appended below. I want to thank this reviewer for their time. I also independently read the revision before consulting the review. I think this revision is clearer than the previous version and I would like to move forward with a revise decision. I do not plan to send a revision back out for review and thus the time lag to the final decision should be greatly reduced if/when you submit the revised manuscript. My plan would be to review the new version and the letter of response and then make a final up or down decision.
The reviewer did an outstanding job responding to this revision and you should address each of their concerns either in the revised text or in the response letter. Addressing their concerns will make this a stronger paper. I will highlight some issues that came up as I read your paper. I acknowledge that you might disagree with some (or all) of these points so feel free to pushback against any suggestions you believe will harm your work. Just describe your counterpoints in the letter.

1. I think Reviewer’s point #3 about Hypothesis #2 is worth considering.

2. I agree with the Reviewer’s point #4 about the Interaction effects. I think it is fine to just pursue them as an exploratory set of analyses. I was not sure I could give a clean interpretation of those results and I think reporting the shape of the interactions as suggested by the Reviewer would be useful. At the end of the day, the alphas seemed pretty similar and this seems to be the general conclusion of the paper. I had originally asked about this issue to help clarify the design and to make sure the analytic model was properly specified.

3. I think readers might have the section on pages 10 and 11 under “Calculation of Dependent Variables and Standard Deviations” a little challenging. It might be useful to walk readers through Table 1 early in the Results section. I worry that some readers may question why there is not just a single alpha estimate for the 8 different groups.

4. I think it would be helpful to readers if you could refer back to the explicit hypotheses (esp. 1 to 3) in the Discussion and make it clear when effects were counter to expectations. It is fine that occurred but it would be useful to spell out those out for readers. I think this may help resolve some of the Reviewer’s concerns about the Discussion. However, I also think it is fine if you have to acknowledge that the results themselves were surprising.

5. I think the bottom line conclusion might be too strong just given that the current results are from one study and are perhaps constrained by the sample and design of the study. Thus, I am asking you to consider a more modest conclusion that is tied to the current design. Moreover, like to see formal statements of constraints on generalizability in papers at this outlet. Please strongly consider formally including that section and citing the Simons et al. paper.
Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6), 1123-1128.

6. Likewise, when considering the results issues of statistical power and precision are relevant so that might be worth some additional consideration in the Discussion.

7. Minor Issues

a. Reference to Costa & McRae in page 5 should likely by Costa & McCrae.
b. Can you report the first three eigenvalues for each scale on page 14? I doubt this is a big issue but would a formal parallel analysis be worth implementing?
c. It might help readers if you interpret the effect size measures and provide them with a rubric for partial eta squared on page 21.

Those were the most salient issues that occurred to me in reading the paper and the reviews. Please feel free to contact me for any clarifications.

To access your submission account, follow the below instructions:
1) login to the journal webpage with username and password
2) click on the submission title
3) click 'Review' menu option
4) download Reviewed file and make revisions based on review feedback
5) upload the edited file
6) Click the 'notify editor' icon and email the confirmation of re-submission and any relevant comments to the journal.

Good luck revising this work. Thank you for trusting us with your paper.

Sincerely,

Brent Donnellan
Michigan State University
donnel59@msu.edu

------------------------------------------------------
Reviewer A:

I appreciate your effort in incorporating the previous round of feedback in this paper, and I can see its improvements as the result. However, I still have several questions about this paper. Again, just like last review, I hope my comments may be useful as you further improve this paper (either for this journal or for the next outlet).

1. This relates back to my previous comment #1, which I don’t think is resolved: Your hypotheses about the improvement in reliability associated with (1) quality control items; (2) confidential administration; and (3) grouped items are rooted in the expectation that these conditions will reduce overly random responding, which negatively impacts reliability. I can accept this argument, but you are ignoring how other forms of responses (e.g., straightlining) may also be impacted by the three conditions, which can increase reliability under some circumstances (DeSimone, DeSimone, Harms, & Wood, 2018).

2. An implicit assumption about the influence of quality control items (pp. 5-6) is that respondents would read these items and then care about their responses to these items. If someone is responding in a random fashion, I am not sure they would even see these items.

3. In developing Hypothesis 2 on confidential data collection, you bring up employment testing and psychological/psychiatric testing. The reason scores in these testing situations are expected to be more reliable is not due to confidential data collection, but rather the important consequences these tests may lead to. Thus, you are confounding the stakes in testing with the lack of anonymity here.

4. I understand the Editor asked you to explore interaction effects, but I don’t think you need to set up hypotheses (Hypotheses 4-7) just for these tests. In fact, it is quite odd to see these hypotheses with no justification at all. Can you simply report exploratory analyses in the Results section without H4-H7?

In terms of results of these interactions, it is difficult for me to follow your report (pp. 20-21). It might be helpful to simply report the shapes of the interactions, as opposed to saying the hypotheses were not supported.

5. I am not sure about your decision to report that you generated sample data “based upon the (mean) alpha and the standard deviation” (p. 11), as a reader is likely to challenge this decision (e.g., whether this approach is valid). In the response letter, you reported that the Rindskopf program based on summary data led to the exact same results as your simulated data. It seems more straightforward to simply report that you used the Rindskopf program to generate the current results.

6. I don’t think the results provide partial support for Hypothesis 2 (p. 19), as two of the three tests resulted in significant results in the opposite direction. Shouldn’t you reject Hypothesis 2 instead?

7. A sentiment I had in my last round of review was you could have done more in this paper, given the study design and data, and the same sentiment still lingers. I understand that you chose to focus on alpha, but this does not preclude other investigations of measurement properties in exploratory analysis, such as the scale means and SDs suggested by the Editor. I appreciate your comparison of scale intercorrelations, but I would probably want to see a bit more, such as the factor structure across conditions (see Schell & Oswald, 2013).

8. The Discussion leaves much to be desired. Some of your hypotheses yielded significant results in the opposite direction. Can you come up with potential explanations?

9. A minor issue: Schell and Oswald was published in Personality and Individual Difference.

**Author response**

October 16, 2020

Dear Dr. Donnellan,

My coauthor, Dr. Marcia Simmering, and I thank you for the opportunity to revise and resubmit our paper. We also thank you for not sending it back to Reviewer 2 from the first round of reviews. We believe that we have diligently addressed each of your concerns and those of the reviewer and address each in a point-by-point manner below.

Best Regards,

Brian and Marcia

**Editor's Comments**

Thank you for submitting your revised work to Collabra: Psychology.   I sent the paper back to one of the previous reviewers and that review is appended below. I want to thank this reviewer for their time.  I also independently read the revision before consulting the review.  I think this revision is clearer than the previous version and I would like to move forward with a revise decision. I do not plan to send a revision back out for review and thus the time lag to the final decision should be greatly reduced if/when you submit the revised manuscript. My plan would be to review the new version and the letter of response and then make a final up or down decision.

The reviewer did an outstanding job responding to this revision and you should address each of their concerns either in the revised text or in the response letter.  Addressing their concerns will make this a stronger paper. I will highlight some issues that came up as I read your paper. I acknowledge that you might disagree with some (or all) of these points so feel free to pushback against any suggestions you believe will harm your work. Just describe your counterpoints in the letter.

 *Authors' response: Thank you!*

1. I think Reviewer’s point #3 about Hypothesis #2 is worth considering.

*Authors' response: We thank the editor and reviewer for this suggestion and have made appropriate changes to the paper and addressed this more thoroughly in our response to the reviewer's point 3 below.*

2.  I agree with the Reviewer’s point #4 about the Interaction effects. I think it is fine to just pursue them as an exploratory set of analyses.  I was not sure I could give a clean interpretation of those results and I think reporting the shape of the interactions as suggested by the Reviewer would be useful.  At the end of the day, the alphas seemed pretty similar and this seems to be the general conclusion of the paper.  I had originally asked about this issue to help clarify the design and to make sure the analytic model was properly specified.

*Authors' response: We thank the editor for this bit of a reprieve. However, as we detail in our response below to the reviewer we think that describing the shape of interactions will not add to the paper given that none of them were significant or when significant they were not in the direction implied by the confluence of the main effect hypotheses. Given this reprieve of sorts, we re-wrote the next to last sentence in the abstract as follows with changes in all caps: "Although most of the EXPLORATORY interaction tests for each scale were statistically significant, none categorically WERE IN THE DIRECTION IMPLIED BY THE CONFLUENCE OF MAIN EFFECT HYPOTHESES." Please see our more detailed response to the reviewer's point 4.*

3. I think readers might have the section on pages 10 and 11 under “Calculation of Dependent Variables and Standard Deviations” a little challenging.  It might be useful to walk readers through Table 1 early in the Results section. I worry that some readers may question why there is not just a single alpha estimate for the 8 different groups.

*Authors' response: We must note that in the version of our manuscript to which the editor's and reviewer's suggestions refer, we included the following sentence in the Measures section on Conscientiousness: " Cronbach's coefficient alpha of internal consistency reliability for the whole sample was .79 with slightly different values in the eight treatment conditions." To more prominently bolster this notion we have now amended each overview of the other two scales being analyzed to include the following: "...with slightly different values in the eight treatment conditions or subsamples." Additionally, we added "...(i.e. sub-samples created experimentally)" to the end of our first mention of Table 1 to alert the reader that the treatment conditions were experimentally created sub-samples. Next we moved the statement described above as well as Table 1 to AFTER the overview of the measures. Then we have added the following detail to the very first part of the results section to directly address the editor's suggestion: "As seen in Table 1 the range of alphas in each of the experimentally created sub-samples was .719 to .825 for conscientiousness, .769 to .863 for entitlement, and .836 to .910 for entitlement. The mean alpha for the entire sample for these three scales was .79, .82. and .87, respectively." Lastly, we amended the note for Table 1 to indicate that M = mean alpha. We hope the editor agrees that this minor rearrangement and additional text helps the reader follow the paper and we thank the editor for this very good suggestion.*

4. I think it would be helpful to readers if you could refer back to the explicit hypotheses (esp. 1 to 3) in the Discussion and make it clear when effects were counter to expectations.  It is fine that occurred but it would be useful to spell out those out for readers.  I think this may help resolve some of the Reviewer’s concerns about the Discussion. However, I also think it is fine if you have to acknowledge that the results themselves were surprising.

*Authors' response: We agree and added verbiage about the hypotheses to the beginning of each paragraph providing an overview of the results. This required that we rearrange the paragraphs so as to flow from H1 to H2 to H3. We added text (in all caps here) to the newly arranged first paragraph on hypothesis one as follows: "THE FIRST HYPOTHESIS ABOUT THE EXPECTED IMPROVEMENT TO RELIABILITY WITH THE INCLUSION OF QUALITY CONTROL ITEMS YIELDED the overall weakest effect AS THERE WAS no impact on score reliability on any of the three instruments." Similar text was added the beginning of each of the two other overview paragraphs. We thank the editor for this suggestion.*

5. I think the bottom line conclusion might be too strong just given that the current results are from one study and are perhaps constrained by the sample and design of the study.  Thus, I am asking you to consider a more modest conclusion that is tied to the current design.  Moreover, like to see formal statements of constraints on generalizability in papers at this outlet.  Please strongly consider formally including that section and citing the Simons et al. paper.
Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6), 1123-1128.

*Authors' response: We thank the editor for helping us more properly focus on the limitations of our paper and have added a concluding paragraph to the manuscript that reads as follows in its entirety: "These caveats are in line with recommended constraints on the generality of findings in the social sciences (Simons, Shoda, & Lindsay, 2017) and 'we have no reason to believe that the results depend on other characteristics of the participants, materials, or context' (pp. 1126-1127). The target population is adult survey respondents in general and our sample was comprised of undergraduate students at one university which limits the generalizability of the results. The sample size also limits generalizability in that the statistical power to detect an effect where one truly exists fell short of the standard of .80. Regarding the instruments being used here, it should be noted that they may have suffered from multidimensionality and it is well-known that Cronbach's alpha is best suited for unidimensional measures. Our exploratory factor analysis revealed that a two-factor solution fit the conscientiousness data best, either a one or two-factor solution depending on whether one gives more credence to eigen values or to the scree plot fit the work ethic data best, and a one-factor solution was supported for the entitlement data regardless of whether interpreting the eigen values or the scree plot. The procedure was quite common in that paper and pencil surveys were administered in both confidential and anonymous situations mimicking the real-world implementation. However, given this sample's likely familiarity with surveys frequently administered in exchange for extra credit in the college classroom, the results could be affected by both demand characteristics and previous experience with being allowed to engage in random responding with no consequence."*

6. Likewise, when considering the results issues of statistical power and precision are relevant so that might be worth some additional consideration in the Discussion.

*Authors' response: We ran power analysis using G\*power software and found that our power fell short of the standard. We report the power in the section on effect sizes on page 20 and cite the G\*power software creators in our paper. Additionally, in our discussion section we added the following caveat on page 22: "It should be noted that the statistical power in this study was short of the standard of .80 and these results should be interpreted in this light."*

7. Minor Issues
a. Reference to Costa & McRae in page 5 should likely by Costa & McCrae.

 *Authors' response: Oops. Thanks.*

b. Can you report the first three eigenvalues for each scale on page 14?  I doubt this is a big issue but would a formal parallel analysis be worth implementing?

*Authors' response: We added the following sentences to each paragraph in the measures section on our three scales. For conscientiousness the results suggested a two-factor solution so we added: "The first three eigen values were 3.56, 1.28, and .93." For entitlement the results suggested a one-factor solution so we reported the first two eigen values with the following: "The first two eigen values were 3.55 and .85." For work ethic the results were mixed in that the eigen values suggested a two-factor solution and the scree plot suggested a one-factor solution so we reported the first three eigen values as follows: "The first three eigen values were 4.72, 1.06, and .89." Because Cronbach's alpha is best suited for unidimensional measures we provided the exploratory factor analysis solely in an effort at being complete in our reporting. However, it is our experience that it is not uncommon for scales designed to measure only one trait to have more than one underlying dimension. We think that parallel analysis will not provide anything extra to the paper in view of the fact that we proceeded with the computation of Cronbach's alpha despite the possibility that the scales were not unidimensional. However, we think this is worthy of inclusion in the discussion section where we now include it as a limitation. On page 22 we added the following: "Additionally, Cronbach's alpha is best suited for unidimensional measures which may be problematic here. Our exploratory factor analysis revealed that a two-factor solution fit the conscientiousness data best, either a one or two-factor solution depending on whether one gives more credence to eigen values or to the scree plot fit the work ethic data best, and a one-factor solution was supported for the entitlement data regardless of whether interpreting the eigen values or the scree plot." We hope the editor agrees that because we proceeded with the analysis on the data despite the possibility of some structure problems in the scales the provision of results from parallel analysis will add very little to the paper.*

c. It might help readers if you interpret the effect size measures and provide them with a rubric for partial eta squared on page 21.

*Authors' response: We used Cohen's (1988) guidelines for each set of effects and characterized the effects as being loosely construed as...very small or small, as the case may be, in view of the consensus among researchers that these were merely offered as very general rules of thumb by Cohen and were not intended to be strict rules for interpreting effect sizes. That verbiage is included in the section on effect sizes on page 20.*

Reviewer A Comments

I appreciate your effort in incorporating the previous round of feedback in this paper, and I can see its improvements as the result. However, I still have several questions about this paper. Again, just like last review, I hope my comments may be useful as you further improve this paper (either for this journal or for the next outlet).

1. This relates back to my previous comment #1, which I don’t think is resolved: Your hypotheses about the improvement in reliability associated with (1) quality control items; (2) confidential administration; and (3) grouped items are rooted in the expectation that these conditions will reduce overly random responding, which negatively impacts reliability. I can accept this argument, but you are ignoring how other forms of responses (e.g., straightlining) may also be impacted by the three conditions, which can increase reliability under some circumstances (DeSimone, DeSimone, Harms, & Wood, 2018).

*Authors' response: Two of our hypotheses are technically about techniques designed to reduce overly random responding and therefore improve reliability. The third, about grouped items, might actually encourage overly random responding in that grouping similar items together in a survey will increase the ease of providing consistent responses. We do not refer to the extreme form of overly consistent responses referred to by the reviewer as "straightlining", but are pleased to use that terminology in this version of our paper and to include a statement about it using the citation that is recommended. Thus, on page 7 we rewrote the first paragraph with the changes below in all caps:*

*"...can be matched to respondents. AN EXTREME FORM OF OVERLY CONSISTENT RESPONDING IS KNOWN AS "STRAIGHTLINING" WHEREBY A SURVEY RESPONDENT PROVIDES THE EXACT SAME RESPONSE DOWN A STRAIGHT LINE OF ITEMS ON A SURVEY (DESIMONE, DESIMONE, HARMS, & WOOD, 2018). STRAIGHTLINING IS NOT AN UNCONSCIOUS HUMAN TENDENCY LIKE THE CONSISTENCY MOTIF NOR IS IT AN EFFORT AT AVOIDING DETECTION FOR INCONSISTENCIES BUT RATHER IT IS FORM OF INSUFFICIENT EFFORT RESPONDING. Of note is the difference between the consistency motif and acquiescence bias. With the latter, respondents engage in yea-saying or nay-saying and researchers typically use reverse scored items to discourage choosing all high or all low responses to the items in an instrument. REVERSE SCORED ITEMS USUALLY HELP ALLEVIATE THE TENDENCY OF SOME RESPONDENTS TO ENGAGE IN STRAITLINING AS WELL."*

*We hope this adequately captures the reviewer's excellent suggestion.*

2. An implicit assumption about the influence of quality control items (pp. 5-6) is that respondents would read these items and then care about their responses to these items. If someone is responding in a random fashion, I am not sure they would even see these items.

*Authors' response: We agree. If someone is truly randomly responding then they probably won't read any of the items, let alone a few randomly inserted quality control items. However, for those not completely randomly responding but simply engaging in less-random but nevertheless careless responding, we think the quality control items might serve as a bit of an eye-opener thus indicating that their responses do matter. We have added the following to pages 5-6 about this point with our changes in all caps here: "...and content non-responsivity (Nichols, Greene, & Schmolk, 1989). HOWEVER, TO A RESPONDENT KEENLY INTENT ON TRULY RANDOM RESPONDING, EVEN QUALITY CONTROL ITEMS MAY NOT BE SALIENT. Quality control items can be worded..." We thank the reviewer for this observation.*

3. In developing Hypothesis 2 on confidential data collection, you bring up employment testing and psychological/psychiatric testing. The reason scores in these testing situations are expected to be more reliable is not due to confidential data collection, but rather the important consequences these tests may lead to. Thus, you are confounding the stakes in testing with the lack of anonymity here.

*Authors' response: We regret that we did not more properly highlight this issue. However, we did write on page 6, " Such situations are more common in employment testing and psychological/psychiatric testing where it is highly critical to match responses to respondent and, perhaps equally so, that the scores on such tests be reliable." We have added the following to this passage with our changes in all caps: "...perhaps equally so, BECAUSE OF THE PURPOSES OF THESE TESTS, THE scores on such tests MUST be HIGHLY reliable." We thank the reviewer for this suggestion.*

4. I understand the Editor asked you to explore interaction effects, but I don’t think you need to set up hypotheses (Hypotheses 4-7) just for these tests. In fact, it is quite odd to see these hypotheses with no justification at all. Can you simply report exploratory analyses in the Results section without H4-H7? In terms of results of these interactions, it is difficult for me to follow your report (pp. 20-21). It might be helpful to simply report the shapes of the interactions, as opposed to saying the hypotheses were not supported.

*Authors' response: We thank the reviewer for this suggestion and have retitled that sub-section as "Exploratory Analysis of Interaction Effects" and have added the following passage to the first sentence of the introductory paragraph: " Because the creation of eight independent groups of respondents was made possible by the two-by-two-by-two experimental design, we engaged in some exploration of the three different two-way interaction effects and the single three-way interaction. Regarding..." Then in each subsequent paragraph in that section we dropped mention of hypotheses and simply explain the results of tests on three different two-way interactions and one three-way interaction. Because none of these interactions supported the directionality of the exploratory analyses on interactions we think it is inappropriate to then discuss their shape. In essence, there were no significant interactions in the expected direction so we think the discussion should about them should probably end there and hope that the reviewer agrees.*

5. I am not sure about your decision to report that you generated sample data “based upon the (mean) alpha and the standard deviation” (p. 11), as a reader is likely to challenge this decision (e.g., whether this approach is valid). In the response letter, you reported that the Rindskopf program based on summary data led to the exact same results as your simulated data. It seems more straightforward to simply report that you used the Rindskopf program to generate the current results.

*Authors' response: To be clear, the Rindskopf program and David Nichols' SPSS syntax was used to conduct factorial ANOVA (i.e. main effects and interaction effects) tests. We also used it to test the main effects but could have done the main effects tests without the help of either Rindskopf or Nichols given that we had n, sd, and mean alpha for each sub-group created in the 2x2x2 experiment. Because the interactions tests are technically separate from the main effects tests and reported in a different section of the paper we removed mention of generating sample data as recommended by the reviewer from page 11 and simply referred to the R code on page 19 and, of course, continue to acknowledge David Rindskopf in the acknowledgement section of the paper. We thank the reviewer for helping us streamline our paper in this manner.*

6. I don’t think the results provide partial support for Hypothesis 2 (p. 19), as two of the three tests resulted in significant results in the opposite direction. Shouldn’t you reject Hypothesis 2 instead?

*Authors' response: We can find no clear-cut guidance on this topic in the extant research. There are certainly some blog posts on the topic, but it seems to be a bit of a contentious topic. On the one hand, we might argue that if we had done the analysis on, say, 100 different instruments or scales and on 99 of them the results supported our hypothesis could we then not claim partial support or even claim full support because only one of them failed our test? On the other hand, if we use 100 instruments and the hypothesis was supported on only one of them would that still be partial support or would that qualify as no support? We think it is safe to err on the side of partial support as we have no way of knowing if the findings would generalize to other scales or not and therefore provides fodder for our discussion section as well as some newly added caveats on the generalizability of our findings. We hope this is acceptable to the reviewer.*

7. A sentiment I had in my last round of review was you could have done more in this paper, given the study design and data, and the same sentiment still lingers. I understand that you chose to focus on alpha, but this does not preclude other investigations of measurement properties in exploratory analysis, such as the scale means and SDs suggested by the Editor. I appreciate your comparison of scale intercorrelations, but I would probably want to see a bit more, such as the factor structure across conditions (see Schell & Oswald, 2013).

*Authors' response: We agree with the reviewer but think it prudent to focus our paper on measurement reliability and the various machinations that survey designers and implementers use to improve reliability. Straying too far from this beaten path may muddy the waters (two metaphors in one sentence!) for the reader.*

8. The Discussion leaves much to be desired. Some of your hypotheses yielded significant results in the opposite direction. Can you come up with potential explanations?

*Authors' response: We apologize for this. Our responses to the above suggestions as well as those from the editor have allowed us to make some major changes to the discussion section. First, the stating of the hypotheses in the discussion section was added to assist the reader. Second, an entire new paragraph on the limits to the generalizability that partly incorporates some limitations to the study was added to the discussion section. To more directly address the query about other potential explanations we added the following to pages 22-23: "When items are grouped the reliability of scores may suffer because some respondents engage in some purposeful deviations from consistency so as not to be seen as straightlining thereby providing downward pressure on internal consistency reliability. That is, they may occasionally provide a slightly different response to an item in a set of like items so as not to be caught in, or accused of, straightlining. On the other hand when items are scrambled straightlining is not likely as it would require a search for similar items randomly placed in a survey. Thus, a more truthful response may be provided to those items that is not dependent on remembering one's responses to other items measuring the same construct placed elsewhere in the survey. In sum, recall is not necessary when responding consistently truthfully to scrambled items from a particular scale thus applying upward pressure to reliability. When lying or faking becomes the task, remembering one's previous lies is a problem and the responses are likely to become somewhat random when not truthful."*

9. A minor issue: Schell and Oswald was published in Personality and Individual Difference.

*Authors' response: Thank you! We are sort of embarrassed by this oversight and apologize.*

**Editor final decision—Accept**

Octobr 22, 2020

Dear Dr Brian Keith Miller,

Thank you for submitting the revised version of your manuscript and response letter. I appreciate your attention to the remaining issues and I am happy to accept this paper for publication at Collabra. I look forward to seeing this cited in the literature. Congratulations!

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. The accepted submission will now undergo final copyediting. You will be contacted once this is complete to answer any queries that may have arisen during copyediting and to allow a final chance to edit the files prior to typesetting. If you wish to view your submission during this time, you can log in via the journal website.

I hope you will continue to submit work to our journal.

Sincerely,

Brent Donnellan
Michigan State University
donnel59@msu.edu