*Dear Prof Kisbu,*

*We have now revised the manuscript in light of your and the reviewers’ very useful comments. We feel that the manuscript is now improved, having corrected our own misreadings and places where the previous manuscript overreached. We are grateful to you and the reviewers for keeping us in check and guarding against over-claiming, and feel that the revised manuscript is more robust for it.*

***Summary of actions taken, following those specified in editor’s comments:***

* *We have substantially revised the manuscript in light of Reviewer 2’s useful comments about how realistically item dropping could account for the results. To do this, we have clarified the definition and descriptions of alpha hacking throughout the manuscript, so as to emphasise that there are a variety of forms that it might take, not only item dropping. In addition, the manuscript is now explicit that we are agnostic to what form(s) of alpha hacking may have occurred in the literature to date. The goal of our manuscript is to present evidence that some form of distorion is taking place, rather than to argue what specific practices may have given rise to these distortions.*
* *We have reworded our poorly phrased sentence about the reliability of item dropping results.*
* *We have corrected the three important miscitations that Reviewer 1 found. We no longer reference either Kopalle & Lehman (1996) (re its relevance to the impact of item dropping) or Heggestad et al. (re its relevance to scale modification practices). We have also corrected our miscitations of Cortina (2020): the manuscript is now clear that Cortina only examined pre-data collection scale modifications. These sections those references were in have been rewritten to be more cautious about what previous research has already shown, and what our results might indicate.*

*We provide point-by-point responses to the reviewers below, along with relevant quotes and page numbers. In all but the final point, we accept the issue the reviewers have raised and have appropriately changed the manuscript.*

*We have uploaded both a copy of the revised manuscript, as well as a separate copy with tracked changes to allow you to see what elements of the manuscript have been revised.*

*Kind regards*

*Ian Hussey*

*On behalf of the authors*

**Editor’s comments**

09-Jan-2024  
  
Dear Dr. Hussey:  
  
Thank you for submitting your Empirical Article (AMPPS-23-0020) entitled "An aberrant abundance of Cronbach’s alpha values at .70" to Advances in Methods and Practices in Psychological Science (AMPPS). The manuscript has now been reviewed a second time, and the reviewer comments appear at the end of this letter.  
  
I have thoroughly reviewed your revised manuscript along with the feedback provided by the reviewers. I am quite positive about the overall quality of your work, recognizing its numerous merits. However, it's important to note that while Reviewer 2 recommends acceptance, Reviewer 1 has raised some concerns that align with my own reservations. Given this sentiment, I would like to ask for a subsequent revision, with a particular focus on addressing Reviewer 2's comment #2 questioning the realistic nature of the conditions. Additionally, please review the sentence in the manuscript that states, "...common item removal strategies produce unreliable results due to low power," as there seems to be insufficient support for this claim.  
  
Please note that since both reviewers already reviewed the manuscript twice and provided valuable feedback, I intend to proceed with the decision-making process after receiving your revised manuscript, without the need for further peer review.   
  
We ask that you submit your revision within three months. Please let us know if you will not be able to meet this deadline. To submit your revision, log into <https://eur03.safelinks.protection.outlook.com/?url=http%3A%2F%2Fmc.manuscriptcentral.com%2Fampps&data=05%7C02%7Cian.hussey%40unibe.ch%7C33a73d3e2f1c4141ccec08dc11372839%7Cd400387a212f43eaac7f77aa12d7977e%7C1%7C0%7C638404175798946151%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=LHPm6YoGAQ%2FJb0rZTNn8MyiNwLhIom9IMyJHuzBesUg%3D&reserved=0> and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.  
  
IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant or outdated files before completing the submission.  
  
Once again, thank you for submitting your manuscript to AMPPS and I look forward to receiving your revision.  
  
Sincerely,  
  
Yasemin Kisbu  
Associate Editor, Advances in Methods and Practices in Psychological Science

**Reviewer 2**

**Reviewer 2 comment 1**  
I have no particular further comments on this paper.  
  
The only thing I would say is that some of the justification that has been added as to why alpha hacking is an issue seems to me to slightly overstate the importance of strict adherence to the exact specifics of a scale. There are many many instances of reduced scales being used when survey design has limits, and I think in many cases this is not necessarily an issue - they still measure the 'same thing', broadly speaking. This is especially the (conceptual) case if one takes a realist ontological position that is enforced by using reflective measurement such as a summative scale (as convincingly argued by Borsboom 2005). This, however, does not distract from the fact that fiddling with the items present to get a certain level (0.7, mainly in this case) is not bad practice. It clearly is, and the recommendations that are outlined by the paper about transparency are well put.

ref:  
Borsboom, D. (2005). Measuring the mind: Conceptual issues in modern psychometrics. Cambridge University Press. <https://www.cambridge.org/core/books/measuring-the-mind/1DB84F33B196C4F2658209B7BC8806E1>   
  
***Author response***

***Summary of action taken:***

*(1) We have further clarified our definition of alpha hacking in the manuscript:*

*“When measures are altered ad hoc after seeing the results with the goal or effect of increasing in-sample α estimates, this conditions the analytic choices on the results and overfits to the in-sample data, making it less likely that increases in α will be reproduced in new samples. That is what we term α-hacking.” (p. 5)*

*(2) the manuscript is now explicitly agnostic about the forms that alpha hacking might take (i.e., item dropping has been deemphasised throughout).*

*(3) the manuscript is also now explicit that we are not arguing for strict adherance to an original scale, but rather we are arguing that not all scale modifications are aligned with genuine scale improvement. E.g.,*

*“For example, while practices like item dropping, alterations to reverse scoring, and subscale redefinition might all be legitimate and necessary forms of scale improvement in some contexts, these practices can also be abused in order to artificially increase in-sample alpha (i.e., ways that overfit on the data at hand, without producing increases in reliability that would be reproduced in new samples). The line between legitimate and problematic use of these methods can be a very fine one, and this situation is sometimes not helped by their unthinking use.” (p. 26-27)*

**Reviewer 1**

**Reviewer 1 comment 1**  
I was Reviewer 1 on the previous version of the article. Unfortunately, I largely stand by my original points and do not feel that they have been sufficiently addressed. Again, I do not disagree with the authors that measurement should be done more carefully and more transparently in psychology, and that many poor practices abound, including, possibly, “alpha-hacking”. But I don’t think we should be doing bad science to point out bad science, and for this reason I must insist that my original objections and concerns have not been addressed.

***Author response***

*We thank the reviewer for insisting. In the previous review round, we did not fully appreciate this concern in the previous round, and are grateful to have the opportunity to correct these issues in this revision. We discuss these in detail below, under the reviewer’s individual points.*

**Reviewer 1 comment 2**

1. As the first point in my original review, I said that the onus is on the authors to provide evidence of “alpha-hacking”, in the form of item-dropping to improve coefficient alpha. In response, the authors provided some references which according to them provide “excellent evidence” of this, such as Cortina et al. (2020). They say that “Cortina et al. (2020) find that 41% of the studies in their sample dropped one or more items from an already established scale.” However, the evidence needed is that item-dropping occurs post-data collection in order to improve alpha. I believe the evidence in Cortina (and several other cited papers) simply points to the fact that researchers often shorten scales prior to use; in fact, this alarming trend is well-documented in this article. Thus, we still do not have evidence that it is common for researchers to collect data on the entire existing scale, and then, based on item analyses, drop items from this scale.

***Author response***

*We are genuinely grateful to the reviewer for keeping us in check and requiring us to read and tread more carefully. We would like to be clear that our argument is that alpha hacking (distortions in alpha) is occurring in the literature - but we are agnostic to what form it takes. Previous versions of this manuscript leaned too hard on the idea that item dropping was a possible mechanism. We see now that this was a misstep, and have changed the manuscript accordingly.*

*The manuscript is now clear that we are agnostic to what form alpha hacking takes, and are explicit that it is not tied uniquely to item dropping.*

*We have corrected all miscitations of Cortina et al. (2020) and Heggestad et al., (2019). E.g.,*

*“Cortina et al. (2020) observed that the most common form of pre-data collection self-report scale modification is item dropping, which could also be applied post hoc as a form of α-hacking. Several other forms of scale modification identified by Cortina et al. (2020) could also be exploited post hoc for α-hacking, including creating composite scores, dichotomizing scores, opportunistically reverse coding items, or otherwise altering the scoring strategy.” (p. 28).*

*The only remaining point of disagreement surrounds the comment 'the onus is on the authors to provide evidence of “alpha-hacking”, in the form of item-dropping to improve coefficient alpha'. We absolutely agree the onus is on us to provide evidence of alpha hacking, ie. opportunistic scale alterations to increase alpha through overfitting to the in-sample data. We present the deviations from a smooth distribution that as this evidence, as only opportunistic scale adjustments can explain it. Our mistake in previous versions of the manuscript was to cite (misreadings of) previous research that might also speak to the occurrence of alpha-hacking or the form it might take. We have rectified this in the current manuscript. The manuscript is now clear that we are agnostic to what form alpha hacking takes, and that further research is necessary to consider what forms it may take and what impact and prevalence they might have. The “Possible explanations” heading in the discussion discusses this in detail.*

**Reviewer 1 comment 3**

My second point (sidenote) was that it is actually fairly difficult to manipulate alpha under realistic conditions, and certainly difficult to get it to increase by dropping an item, particularly for well-established scales. The authors respond by citing simulations in (Kopalle & Lehmann, 1997; not in the references but I was able to find it). Specifically, they say: “Kopalle and Lehmann's (1997) simulations demonstrate that when true α = .63, item dropping increases the apparent α by an average of .10, and in specific instances of up to .30.” This article is cited heavily throughout the response letter.  
  
2. I had to look up the conditions in this paper because this summary contradicts my own simulations and my own wealth of experience computing and comparing alphas on scales and subscales. This finding is one of the lines in their Table 2, and the conditions for the average bias of .10 are as follows: true loadings are .5, number of initial items is either 10, 20, or 30 (averaged across), the number of selected items is either (3, 5, or 10 (averaged across), and sample size is either 3 or 10 times the number of items. Thus, this average bias is across conditions with many small sample sizes (half are 30-90), and the number of selected items is often quite small (e.g., two thirds of the time, we are selecting 3-5 items out of 10-30). This is not the same as dropping an item or two from an existing scale. To boot, the result that the bias can be up to .3 I believe comes from simulations in their Figure 1, when 3 out of 10 items get chosen at very small samples (n=30). To claim that this article provides evidence relevant for applied practice is disingenuous. I invite you to run your own simulation under realistic conditions (e.g., N = 100, 10 items on a scale, or whatever you think is realistic), trying to see if an item that otherwise has a good loading in the population has such a low loading in a particular sample that it lowers alpha. This is difficult to do. The simulations you cite do not create such conditions.

***Author response***

*We apologize for neglecting to include Kopalle & Lehmann (1997) in the reference section. More importantly, we agree that our previous version of the manuscript over-reached in its citation of Kopalle & Lehmann. We have removed citations of Kapalle & Lehmann from the manuscript, and de-emphasised item dropping throughout the manuscript as an example of one form that alpha hacking may take.*

*With that said, we still believe that Kopalle & Lehmann is relevant and claiming so is not disingenuous. As we see it, the difference is probably what we consider "applied practice". We believe that we and the reviewer have different foci: Yes, for well-established, well-validated scales whose items each load relatively well onto the factor, any changes to item composition are more likely to decrease alpha rather than increase it (due to the relationship between scale length and alpha). However, the datasets we employ in our manuscript are not limited to established scales, and our goal is to speak to more general measurement practices outside of this subset of work. E.g., our introduction now states:*

*“While excellent measure development and refinement work is of course done by assessment specialists, it is an unfortunate fact that the majority of the primary research literature is produced by substantive researchers who are relatively less expert in measurement. Measurement choices in the primary literature often remain unreported or severely underreported (Flake & Fried, 2020; Flake et al., 2017; Hussey & Hughes, 2020), which undermines the ability to quantify their effects on test reliability and validity (Elson, 2019; Flake et al., 2022). Hence, to reiterate, here we are focused on measurement as it occurs in the primary literature, and* *use the term α-hacking to refer to this type of ad-hoc, weakly justified and underreported practices that overfit to in-sample data and do not increase the test’s true reliability.” (p.7)*

*And:*

*“Analogously to how analytic flexibility allows for p-hacking (Simmons et al., 2011), measurement flexibility may allow for α-hacking (Elson, 2019). Flexibility and under-reporting have been argued to be prevalent issues (Flake & Fried, 2020). For example, Cortina et al. (2020) report that out of 101 self-developed scales, only 25% reported any sort of scale development process in detail. Self-developed, poorly validated, single-use scales are extremely common and appear to make up the majority of scales in psychology (Anvari et al., 2024; Elson et al., 2023), increasing the potential for α-hacking. In addition to this, a large proportion of psychological scales contain few items, and are therefore particularly vulnerable to α-hacking.” (p.5-6)*

**Reviewer 1 comment 4**

3. Related, there is now a sentence in the paper that cites this article in the following manner: “Unfortunately, it remains an underappreciated fact that common item removal strategies produce unreliable results due to low power, and there is little evidence that they can accurately identify poorly performing items (Kopalle & Lehmann, 1997).” This claim is not true, and the cited article provides no support for it. Samples of a few hundred (or at least N  > 100) are sufficient to accurately estimate correlations without a reasonable band such that items with low item-total correlations can be reliably identified. This is an amazingly inaccurate claim that ignores much of the scale development literature and research.

***Author response***

*We agree this sentence was poorly worded. We have removed it, references to Kopalle & Lehman, and the emphasis on item dropping.*

**Reviewer 1 comment 5**

4. My last point was that “alpha-hacking” is not at all similar to p-hacking in its impact. P-hacking completely invalidates the outcome of a statistical test. A p-value is not a population parameter estimate (in fact, it does not converge to a point estimate and has a uniform distribution under the null), whereas alpha is. We can remove all sampling fluctuations from this question by considering that research is only done with very large samples. We would still want to build scales that have some minimal amount of reliability, and so we would want to “hack” them upwards, if they do not. If appropriate tools are used to do this, this is called scale improvement/scale development.

There is also probably a peculiar prevalence of articles with SEM fit indices just below the cutoffs (e.g., RMSEA < .08) and factor analytic solutions with loadings just greater than .3, and any other reasonable cutoff on a population parameter that measures quality in some way. All can be abused, but all can also be used well. That is, p-hacking is a problem of finite samples; alpha’s cutoff carries into the population. I did not receive a good response to this point.

***Author response***

*We thank the reviewer for raising these points, and agree that metrics other than p values that employ common cut-off values are likely to be both appropriately used and abused. A recent article by White et al. (2023), which we cite in the manuscript, has shown this to also be the case in ROC curves in clinical prediction models. We completely agree that when appropriate tools are used to refine a scale, this is appropriately referred to as scale improvement. However, when appropriate tools (e.g., large samples, replication) are used, this would not be what we would call alpha-hacking, as these legitimate increases in reliability would be likely reproducible in new samples, unlike those which were obtained by overfitting on in-sample data. The analogue here is that attempts to decrease p values by increasing sample size would be considered both acceptable and desirable even by critics of p-hacking.* *The problem is similar in both regards: capitalizing on chance variation, overfitting to the data at hand. We seem to agree that these coefficients can be abused. We have rephrased throughout the manuscript to further emphasise that we do not take issue with proper scale development practices (e.g., in large samples or validated in new data), and that these do not constitute alpha hacking. E.g.,*

*“there is more debate about whether other practices which can be benign or even helpful in some contexts but can be abused in others. For example, while practices like item dropping, alterations to reverse scoring, and subscale redefinition might all be legitimate and necessary forms of scale improvement in some contexts, these practices can also be abused in order to artificially increase in-sample alpha (i.e., ways that overfit on the data at hand, without producing increases in reliability that would be reproduced in new samples). The line between legitimate and problematic use of these methods can be a very fine one, and this situation is sometimes not helped by their unthinking use. For example, when calculating α, many common statistics packages including SPSS, the popular R package psych (Revelle, 2018), as well as the open-source programs JASP (2024) and jamovi (2024), all also provide suggestions for what α would instead be if items was dropped. This increases the probability that a researcher may drop items post hoc without providing deeper engagement with the relative performance of the items or the appropriateness of doing so.” (p. 26-27)*

*With regard to the issue raised about p values not being population estimates: our data show clear inflation at alpha = 0.70, a common threshold of acceptability. This is analogous to confidence intervals on effect sizes (i.e., population estimates) barely excluding 0 (i.e., p < .05) and the effect sizes themselves being biased upwards by p-hacking. As such, we feel that there is no such clear cut distinction between p values and alpha estimates, and no a priori reason why one is undermined and the other is not.*

**Reviewer 1 comment 6**

Again, I find myself torn when writing this review because I do completely agree that poor practices abound in scale development and scale use. But I worry much more about people writing near duplicate items to create meaningless scales with artificially high alphas, than I do about people dropping a poorly functioning item from a scale (established or not) based on at least a medium sized sample—I worry less about this both in terms of its severity and impact and its prevalence in applied practice (though I remain open-minded on this point, until further evidence).

***Author response***

*We agree that there are other ways to produce high alpha estimates that do not constitute capitalizing on chance, but still serve to produce artificially high reliabilities. Essentially, it is a case of Campbell's or Goodhart's law. We let a metric stand in for a desirable property of a study or measure. Researchers under pressure find ways to make the number go up without correspondingly improving the desirable property. Here, we are focused on approaches that have to do with capitalizing on chance variation, overfitting. Do we need to worry about them less than near-duplicate items? We'd argue that it could be easier to underestimate alpha hacking than "Modern Talking items" (see Markus Bühner, 2021: situations where the items have the same melody but the text varies slightly), because the former might be less visible than the latter. Regardless, while we completely share the reviewer’s concerns about this form of artificial boosting of reliability, we do not think we need to find out which problem is worse here, as long as we are clear which problems we are concerned with in the present manuscript.*