**An aberrant abundance of Cronbach’s alpha values at .70**

**Response to reviewers letter**

Dear Prof Kisbu,

Thank you for the very constructive comments on our manuscript. We have addressed all points that you and the reviewers raised, and feel that the manuscript is stronger for it.

Our point-by-point author responses can be found below each comment, in italics, along with the updated quotes from the manuscript.

In addition to your requested changes, please note that we have also taken the time to include an analysis of a third large dataset: the APA’s PsychTests database, which claims to include all psychological measures published since 1896, and lists estimates of Cronbach’s alpha for many of them. The APA was kind enough to grant us full database access so that we could scrape information for 67,512 alpha values, from >60,000 measures published across >3000 journals. Results show the same pattern of excesses at alpha = .70 as the other two datasets, providing further evidence of alpha-hacking across a very wide range of literature. To illustrate the consistency of the magnitudes of the excesses, which all fell in the range of 12-15% between dataset and subsets, we have illustrated this in a new plot, see Figure 4.

Kind regards,

Ian Hussey

(on behalf of Taym Alsalti, Frank Bosco, Malte Elson, & Ruben Arslan)

## Editor’s comments:

10-Apr-2023

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, and the reviewer comments appear at the end of this letter.

There are a number of issues that require attention in the current version of the manuscript. These issues have been very competently covered by the expert Reviewers. After reading the manuscript carefully myself, I acknowledge the strengths in the manuscript, but am in agreement with the Reviewers’ comments. Therefore, I would like to invite you to submit a revision of your paper for continued consideration at AMPPS. I urge that you address all the issues raised by the Reviewers.

I would especially like to direct you to the following points:

*[editor point 1]*

1. Please strengthen your argument that the alpha-hacking is a prevalent problem in the literature. For example, you may provide the percent of studies that used item deletion in a randomly picked small subset of studies which reported .7 alpha values.

*Author response: Thank you for this point. We agree that our manuscript’s contribution can be strengthened by a greater degree of clarity that the practices which \*could\* function as alpha-hacking are indeed prevalent. Our analyses then test whether there is evidence of alpha hacking (of some form), and provide strong evidence that this is the case. We have therefore added discussion of several papers that show this is the case to our introduction. The manuscript now reads (p.6-7):*

*“Analogously to how analytic flexibility allows for p-hacking (Simmons, Nelson, & Simonsohn, 2011), measurement flexibility allows for α-hacking (Elson, 2019; Flake & Fried, 2020). Measurement flexibility has already been shown to be prevalent, for example Cortina et al. (2020) found that, among 3,334 multi-item self-report scales, 13% were reported to have been modified in some way, the most common of which was item dropping (37% of cases). 82% of modifications were rated as being major changes with problematic psychometric implications. Elsewhere, Heggestad et al. (2019) found that, among 2,088 scales, 46% were reported to have been adapted, yet evidence to support the validity of the adapted scales was presented in only 23% of modified scales. In their second study, Heggestad et al. (2019) also observed that even among five specific and well-established scales item-dropping occurred between 18 to 64% of the time. More worryingly, both of these studies can only speak to transparently reported alterations. Unreported or underreported modifications are both difficult to quantify, and have additional hidden and deleterious impacts on validity (Elson, 2019; Flake et al., 2022).*

*A review by Flake et al. (2017) concluded that items are often removed on the basis that doing so improves α. Taken at face value, it seems reasonable to remove poorly performing items. 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). In our experience, researchers do not routinely validate such scale alterations in new data. Despite not reliably selecting poorly performing items, item dropping can greatly increase α: for example, 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. Increases in individual cases can therefore be larger than this again, especially when a scale has few items. As such, post hoc changes to scales such as item-dropping represent a strong risk of α-hacking, even when researchers are well-intentioned.*

*[editor point 2]*

2. Please expand the discussion on the consequences/impact of alpha hacking. Also consider scale length in this discussion.

*Author response: Thank you for this opportunity to provide greater detail on the likely consequences and impact of alpha hacking. We have substantially expanded the discussion heading “Magnitude and consequences of α-hacking”, and included consideration of the length of scales and the exposure to alpha hacking it represents. To give our analysis a solid empirical basis, we have tabulated the frequencies of scale lengths according to the APA PsycTests database and highlighted them in the paper. Pages 29-32 now read:*

*“We suspect that Questionable Measurement Practices (Flake & Fried, 2020), including α-hacking, are currently perceived to be as permissible as some p-hacking practices were before the publication of Simmons et al.’s (2011) seminal article False-Positive Psychology. However, ad-hoc measures and ad-hoc modifications to existing measures may have more pernicious and further-ranging consequences than expected. Comparable to how many readers of Simmons et al. (2011) were at the time surprised that p-hacking could alter the apparent in-sample p value so much, readers of Kopalle and Lehmann’s (1997) simulations are often surprised that item dropping can modify apparent in-sample α by so much (e.g., increasing it by an average of .10, and in some instances up to .30). The potential magnitude of boost in observed alpha is enough to change a measure from being, in actuality, suitable only for “early stages of research” to being suitable “when making important decisions” (e.g., α > .90) according to Nunnally and Bernstein’s (1994) guidelines. Like p-hacking, there is also no guarantee that changes to the in-sample estimate bear any relation to a true increase in reliability. And as with p-hacking, it is exceptionally easy to fool oneself. Without replication in new samples, how can we know if, for example, a given item was genuinely performing poorly and should be dropped, or if one is simply conditioning analyses on their own results? We return to this point below when making recommendations for future research.*

*In addition to inflating the perceived reliability of our measures, α-hacking has multiple problematic downstream consequences. That is, reliability is not merely important in and of itself, but because of the relationship between reliability and other properties. For example, the maximum correlation that can be observed between two variables (x and y) is a function of not only the true correlation between the variables (ρ) but also the reliability of the measures of x and y* (i.e., and ), following the formula for correlation attenuation (e.g., Revelle, 2009):

*For example, when the true association between two variables is large (ρ = 0.50), and each variable is measured by a scale with α = 0.70, the maximum observable correlation (in the long run of highly powered samples) is r = 0.35. This has a direct bearing on the validity of statistical power analyses, which must specify an effect size (e.g., an expectation of the true effect size or a smallest effect size of interest, Lakens et al., 2018). While they typically do not explicitly involve quantifying the reliability of the measures used, power analyses are nonetheless dependent on accurate and stable estimates of it (Heo et al., 2015; Parsons, 2018). For example, imagine a researcher accurately judged the true association between the variables to be of large size (ρ = 0.50), but the estimates of the reliability of both measures had been α-hacked through item dropping. Following the results of Cortina et al.’s (2020) review demonstrating that poorly justified item dropping is common and Kopalle and Lehmann’s (1997) simulations, let us assume that item dropping had artificially increased the α of both scales in previous studies from 0.60 to an apparent α = 0.70 (i.e., the mean increase in α observed in their simulations, making this situation realistic). Even though the true population effect size has not changed, the observable effect size is actually r = 0.30 due to the measures’ lower than expected reliabilities. Our hypothetical researcher collected data from 62 participants, expecting that this would provide them with 80% power to detect a true observable correlation of r = .35. However, due to the prior α-hacking constraining the observable correlation more than they realized, these 62 participants only provide 66% power. And, when more severe α-hacking occurs, even more substantial reductions in power would result[3] [4] . As such, although it is distinct from p-hacking, there are good mathematical reasons to believe that α-hacking contributes to lower replicability and therefore weakens the credibility of our claims in a similar fashion. This is not limited to primary research. Several types of meta analysis, such as psychometric meta-analyses (Schmidt & Hunter, 2015; Wiernik & Dahlke, 2020), adjust effect sizes for the reliability of their measures (i.e., disattenuate for reliability). As a result, α-hacking would also bias the results of such meta-analyses.*

*α-hacking can also exacerbate issues of measurement (non)invariance and, in doing so, distort the homogeneity or heterogeneity of research findings. For example, imagine a 7-item scale that is used in two studies to study the same hypothesis. If each study dropped two items in their analysis, then as few as three items would overlap between studies. This may introduce “jingle” issues, where measures share identical names but measure different things (see Elson et al., 2023). That is, both studies state that they used the same scale and measured the same construct, but the overlap in the actual items is low: in this case, as little as 42% of the original items. To what degree are the two studies measuring the same construct any more? It may be the case that these modifications were perfectly appropriate. For example, perhaps these items were poorly translated and function poorly in this population. Or, perhaps the items functioned well, and their negative impact on the scales’ α is just due to sampling error in this sample. It is difficult to know without new data. If the two studies come to different conclusions, is it because of genuine differences in the observed effect, or just because they measured different things? Conversely, if the two studies come to similar conclusions, is this because the effect is replicable, or has the homogeneity of results been artificially increased by changing the measures? These concerns are not merely hypothetical, recent work has shown that replication studies often involve underappreciated degrees of modifications to measures (Flake et al., 2022; see also Elson, 2019 for further discussion of this general problem caused by flexible measures). In our opinion, potential improvements of the in-sample estimates of α are not worth the costs of decreased comparability to existing work, unlikely generalization of the increase in α in new samples, and less accurate estimates of the population value of α.*

*Of course, item dropping is likely to have a larger impact in shorter scales, because each drop represents a larger proportion of the total items. Dropping an item from a 100-item scale would have less influence than dropping an item from a 5-item scale. It is therefore worth asking: what proportion of psychological measures are relatively short, and therefore particularly to this form of α-hacking? We extracted this information from the PsycTests database: 15% of self-report measures have 5 items or fewer, 35% have 10 items or fewer, and 50% have 15 items or fewer. Less than 7% of measures have 50 items or more. Short scales that are quite susceptible to α-hacking therefore make up a substantial portion of all scales in psychology. While item dropping is the most common form of post hoc scale modification (Cortina et al., 2020), it is of course just one of many potential forms of α-hacking. Cortina et al. (2020) also identified several commonly reported methods of self-report scale modification which can be applied after data collection and may therefore be exploited for α-hacking, including item dropping, creating composites, dichotomizing, reverse coding items, or otherwise altering the scoring strategy. Other potential methods of α-hacking likely have direct analogues with p-hacking, such as inappropriate rounding, selective reporting (either of α values at all, or between multiple measures), outlier exclusion, favorable imputation, or subgroup analysis (Stefan & Schönbrodt, 2023).[5] Future research could examine the degree to which different α-hacking strategies bias in-sample α values (as has been done for p values and standardized effect sizes: Stefan & Schönbrodt, 2022).”*

*[editor point 3]*

3. Recommendations for the field can be added at the end of the manuscript (not only for authors but also for journal editors for example)

*Author response: Thank you for this suggestion. We have added a “Recommendations” section that more deeply discusses these. This includes reference to an article which fleshes out specific recommendations for both authors and editors/reviewers regarding how to decrease, detect, and mitigate questionable measurement practices (Elson et al., 2023: the Standardisation Of BEhavior Research (SOBER) guidelines). Page 30-32 now reads:*

*“****Recommendations***

*There are many circumstances under which it is appropriate and important to modify a measure. Measure development, translation, and use in new populations all require an ongoing process of validation. However, when measure development and measure use are conducted within the same dataset, for example by dropping an apparently poorly performing item that lowered α, it is extremely difficult to know whether one has either (a) appropriately refined the measure or (b) overfit on the data at hand. Even with the best of intentions, researchers may be overfitting more than they realize, given that item dropping increases the apparent (in-sample) α by a large degree. This risk is compounded by the fact that 43% of psychological measures are used just once, and indeed 80% of measures are used 10 times or less (Elson et al., 2023). With no, or limited, reuses in new samples, it is exceptionally difficult to avoid overfitting a measure on the data at hand, maximizing apparent α without knowing whether this represents a genuine increase in the reliability of the scale. Equally, when a scale has been reused many times and accumulated more validity evidence, it is difficult to justify modifying it, as this introduces measurement flexibility and heterogeneity into the literature. In order to balance the need to continually validate and refine measures with the risks of overfitting measurement choices, we argue that our field must move away from ad hoc changes to measures done within primary research and towards centralized measure repositories and versioned measurement standards (see Elson et al., 2023).*

*Just as our understanding of the risks of p-hacking lead to a greater distinction being made between exploratory and confirmatory research (e.g., Munafò et al., 2017), we believe that the risk of α-hacking and other Questionable Measurement Practices require researchers to make a greater and more explicit separation between measure development and measure use in primary research. That is, modifications to measures should not be made post hoc; any proposed changes should be confirmed in new samples; and the details of measures should be fixed in preregistrations along with other elements of the design and analysis. Put another way, suggested modifications to scales based on a given dataset should not have those modifications be applied to that same dataset, as this is conditioning decisions on results. Instead, suggestions for modifications (e.g., item dropping) should be confirmed in new samples, and then applied consistently in future. The timing and rationale for such choices is only knowable with increased transparency (e.g., through preregistration). Of course, as with p-hacking, increased transparency through preregistration and a clear separation of exploratory and confirmatory research (aka measure development and use with regard to α-hacking), can increase the detectability of hacking but will not automatically prevent it. Increased transparency about the nature and timing of decisions is necessary but not sufficient to prevent hacking. These and other suggested changes to our measurement practices, how authors comply with them, and how editors and reviewers may enforce them, are discussed in greater detail in the Standardisation Of BEhavior Research (SOBER) guidelines (Elson et al., 2023). There are many differences between p-hacking and α-hacking, although the cure may often be the same: increased transparency about which researcher choices were planned (e.g., through preregistration) and which were data-dependent.*

*Just as with hypothesis tests, preregistration is a plan, not a prison. Preregistering details of a measure does not preclude making post hoc or data-informed decisions about that measure, such as whether to drop a truly badly performing item, it merely makes the timing of this decision visible to others. Deviations from preregistration should be clearly labeled, and the pre-registered analyses using data from the unmodified measures should be reported in addition to any exploratory analyses with modified measures.*

*Construct validation is difficult and often neglected (Schimmack, 2021). Hard questions such as the tradeoffs between internal consistency (which, when high, can represent a form of redundancy), participant time, and construct breadth are important and should be explicitly investigated, and the resulting measures be validated in independent data. For such work to become more commonplace, field norms likely need to change. Currently, the incentive structures in academic psychology tend to reward primary research that attempts to test hypotheses over validation of measures, even when the ability to test those hypotheses relies on valid measurement (Scheel et al., 2021).”*

*[editor letter continues]*

If you choose to submit a revision, please include a letter detailing your point-by-point responses to each reviewer comment and indicating how you changed the manuscript to address them. The revised manuscript may undergo 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 http://mc.manuscriptcentral.com/ampps 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: 1

*[R1 point 1]*

I do not question the analyses done in the article. But I was not convinced by the arguments that an important problem has been uncovered.

First, the article did not convince me that “alpha hacking” is real. For “p-hacking,” a number of common practices have been identified: dropping conditions, choosing a DV, conditioning on a covariate, adding more participants, etc. In contrast, the only specific “alpha hacking” practice identified is dropping an item to increase alpha. I would say this is quite uncommon with established scales.

If I understand correctly, the authors hypothesize that the following occurs: if a paper is using the 10-item Roserberg Self-Esteem Scale, after the data are collected, they drop one item and use a 9-item sub-composite. To me this sounds unusual. I think the onus is on the authors to show this behavior is common. Should it appear in at least some method sections, or is the authors’ implication that it is done entirely in secret? We still need some evidence even for secret behaviors—maybe an anonymous survey, similar to how the prevalence of p-hacking practices was identified.

*Author response: Thank you for this point - we agree that (1) specific examples of methods of alpha hacking/Questionable Measurement Practices and (2) discussion of the prevalence of these practices were not covered in sufficient depth in the previous version of the manuscript. Excellent evidence for both of these points comes from Cortina et al. (2020), which we cited in the previous version of the manuscript but did not adequately discuss. For instance, Cortina et al. (2020) find that 41% of the studies in their sample dropped one or more items from an already established scale. We have therefore expanded our discussion of these points in the introduction.*

*Regarding the already-established widespread prevalence of just one form of alpha hacking (item dropping), page 6-7 now reads:*

*“Analogously to how analytic flexibility allows for p-hacking (Simmons, Nelson, & Simonsohn, 2011), measurement flexibility allows for α-hacking (Elson, 2019; Flake & Fried, 2020). Measurement flexibility has already been shown to be prevalent, for example Cortina et al. (2020) found that, among 3,334 multi-item self-report scales, 13% were reported to have been modified in some way, the most common of which was item dropping (37% of cases). 82% of modifications were rated as being major changes with problematic psychometric implications. Elsewhere, Heggestad et al. (2019) found that, among 2,088 scales, 46% were reported to have been adapted, yet evidence to support the validity of the adapted scales was presented in only 23% of modified scales. In their second study, Heggestad et al. (2019) also observed that even among five specific and well-established scales item-dropping occurred between 18 to 64% of the time. More worryingly, both of these studies can only speak to transparently reported alterations. Unreported or underreported modifications are both difficult to quantify, and have additional hidden and deleterious impacts on validity (Elson, 2019; Flake et al., 2022).*

*A review by Flake et al. (2017) concluded that items are often removed on the basis that doing so improves α.”*

*Regarding the many forms the alpha hacking might take, page 32 now reads:*

*“Cortina et al. (2020) also identified several commonly reported methods of self-report scale modification which can be applied after data collection and may therefore be exploited for α-hacking, including item dropping, creating composites, dichotomizing, reverse coding items, or otherwise altering the scoring strategy. Other potential methods of α-hacking likely have direct analogs with p-hacking, such as inappropriate rounding, selective reporting (either of α values at all, or between multiple measures), outlier exclusion, favorable imputation, or subgroup analysis (Stefan & Schönbrodt, 2023). Future research could examine the degree to which different α-hacking strategies bias in-sample α values (as has been done for p values and standardized effect sizes: Stefan & Schönbrodt, 2022).”*

*[R1 point 2]*

As a side note, the above practice to me would not seem to be very effective, at least for psychological scales of at least 10 items. Alpha tends to increase with the length of the scale if new items have similar intercorrelations to existing items. For alpha to increase – rather than decrease -- when an item is dropped, that item would have to be much less correlated with all the other items. For this increase to be noticeable (e.g., more than .01), the item would have to be quite poorly functioning, or the scale quite short. As an aside to an aside, I believe in continual psychometric validation of scales in every application, and would not find anything wrong with a researcher removing a poorly functioning item even from a well-established scale.

*Author response: We agree that it is important to address readers’ intuitions about the “efficacy” of alpha hacking (i.e., the degree of bias that it can create). We have added discussion of a simulation study by Kopalle & Lehmann (1997) which is very informative on these points. Page 7 now reads:*

*“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). In our experience, researchers do not routinely validate such scale alterations in new data. Despite not reliably selecting poorly performing items, item dropping can greatly increase α: for example, 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. Increases in individual cases can therefore be larger than this again, especially when a scale has few items. As such, post hoc changes to scales such as item-dropping represent a strong risk of α-hacking, even when researchers are well-intentioned.”*

*We also agree that there are of course situations where scale modification is legitimate. We are now explicit about this in the manuscript. However, in light of the aforementioned simulation study results, we encourage authors to be cautious.*

*Page 7 now reads:*

*“Despite not reliably selecting poorly performing items, item dropping can greatly increase α: for example, 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. Increases in individual cases can therefore be larger than this again, especially when a scale has few items. As such, post hoc changes to scales such as item-dropping represent a strong risk of α-hacking, even when researchers are well-intentioned.*

*And pages 34-35 now read:*

*“There are many circumstances under which it is appropriate and important to modify a measure. Measure development, translation, and use in new populations all require an ongoing process of validation. However, when measure development and measure use are conducted within the same dataset, for example by dropping an apparently poorly performing item that lowered α, it is extremely difficult to know whether one has either (a) appropriately refined the measure or (b) overfit on the data at hand. Even with the best of intentions, researchers may be overfitting more than they realize, given that item dropping increases the apparent (in-sample) α by a large degree. This risk is compounded by the fact that 43% of psychological measures are used just once, and indeed 80% of measures are used 10 times or less (Elson et al., 2023). With no, or limited, reuses in new samples, it is exceptionally difficult to avoid overfitting a measure on the data at hand, maximizing apparent α without knowing whether this represents a genuine increase in the reliability of the scale. Equally, when a scale has been reused many times and accumulated more validity evidence, it is difficult to justify modifying it, as this introduces measurement flexibility and heterogeneity into the literature. In order to balance the need to continually validate and refine measures with the risks of overfitting measurement choices, we argue that our field must move away from ad hoc changes to measures done within primary research and towards centralized measure repositories and versioned measurement standards (see Elson et al., 2023).”*

*Importantly, our pattern of results across our three datasets cannot be explained by legitimate scale modification alone. We are now more explicit about this point. Pages 25-26 now read:*

*“It is useful to first set aside some potential explanations of our results as implausible or impossible. First, whether α values were changed due to justifiable modification of measures. For example, dropping an item from a translated scale that, on reflection, maybe poorly performing. While there may be good reasons to modify a measure after data collection (e.g., dropping items, changing the scoring method, etc.), there is no reason to assume that all such changes are conducted exclusively under such circumstances. Numerous sources of evidence show that Questionable Research Practices are prevalent in psychology (for review see Lakens, 2022, section 15.1), and there is no reason to believe that α estimates are somehow uniquely immune to such practices. Indeed, Flake and Fried (2020) recently argued that, more generally, Questionable Measurement Practices are prevalent and concerning. Importantly, well-justified changes to measures cannot explain the excesses of α values at the .70 threshold that we observed: improvements due to post hoc changes to measures would raise α values generally, not just to .70.”*

*[R1 point 3]*

This leads me to my second objection. Alpha-hacking is not at all similar to p-hacking in its impact. When researchers p-hack, they completely invalidate the outcome of the statistical test. There is a wide variety of p-hacking behaviors to choose from, and selecting a few at a time can shoot the Type I error rates up to 80% from 5%. The result of the experiment is invalidated. In contrast, nothing is invalidated to a great extent if researchers “alpha hack”. The authors do talk about some small effects, such as inflation of sample reliability estimates, and I agree, but these effects are tiny (again it’s hard to “move” alpha). Changing a scales reliability from .68 to .71 is unlikely to affect anything else much. Removing a single poorly functioning item from a scale does little damage to the actual analyses using that scale; e.g., the effects on correlations with other scales will be minimal, etc. If the item somehow “malfunctioned” (e.g., translated badly to a new language), removing it will actually increase the validity of subsequent analyses.

*Author response: We disagree with the reviewer’s points here, and we feel there is a useful analogy to draw between p-hacking and alpha-hacking. Importantly, the same researcher behavior (e.g., item dropping) can function as both alpha-hacking (i.e., it can bias alpha values) but also can be a form of p-hacking (i.e., it can bias p values; Stefan & Schonbrodt, 2022). The two are therefore necessarily interrelated.*

*Second, alpha is not hard to move. As is discussed in our reply to [R1 point 2] above, questionable measurement practices can bias alpha by quite large amounts: e.g., Kopalle and Lehman (1997) show biases of more than .30 from the true alpha in some conditions. As we now say in text (p. 6): “The potential magnitude of boost in observed alpha is enough to change a measure from being, in actuality, suitable only for “early stages of research” to being suitable “when making important decisions” (e.g., α > .90) according to Nunnally and Bernstein’s (1994) guidelines.”*

*The degree to which individual alpha can be biased by such hacking should also not be confused with the magnitude of the distortions we observe in the literature in our results - these are two very different phenomena. As we discuss in our discussion section under the new heading of “Prevalence of alpha hacking” (beginning p. 26), our estimates represent only an extreme lower bound of the prevalence of hacked alpha values. Our analyses cannot answer the questions of (a) the true prevalence of alpha hacking, or (b) the magnitude of the degree of bias away from the true alpha in any given case of a hacked alpha value. As such, the small excesses observed in our plots should not be mistaken for either (a) a low prevalence of alpha hacking in the population, or (b) a low severity of the impact of alpha hacking in specific instances, which is already known to be high (Kopalle and Lehman, 1997; see changes to the manuscript referred to previously). Our results represent a hypothesis test of whether alpha hacking is occurring, which has never been tested before.*

*[R1 point 4]*

Lastly, while the authors claim that their effect remained even for a subgroup of extracted coefficients that used well-established scales, I am not sure that it is so easy to identify the sub-sample of articles, without reading them, that aren’t engaged to some degree in scale development/validation. It is common for articles about well-established scales to propose shorter versions of these scales, e.g., the goal may be to create a scale half the length but that has the minimum reliability of .7. Or, well established scales may be tested in other cultures, etc. This is really common. The bump at .7 the authors discuss appears to be quite small. It seems to me that both publication bias and some mix of articles focusing on or performing additional scale development/extension where the .7 criterion is being reached explicitly could explain it. I’m not convinced some other nefarious unreported practices are needed to explain it. They may exist, but we need some evidence of that first, and we have to be really careful about defining what is and isn’t nefarious when it comes to examining item statistics for a well-established scale.

I don’t disagree with the authors on any of their conclusions that measurement should be done more carefully and more transparently.

*Author response: Thank you for this point, and we’re glad you agree with our conclusions.*

*Our intuitions diverge from yours, in that we do not think that a large percentage of studies using an established measure are engaged purely in scale development/validation but rather use for primary research. We have now gone to some lengths to add a third dataset which can aid us in assessing this question of whether alpha hacking is still present in established scales (vs. those in development). Both the I/O dataset (reported in the previous version of the manuscript) and the newly added PsycTests dataset provide additional information regarding whether a measure was newly developed or already established (I/O dataset), and whether the measure was newly developed or a modification or translation. As well as testing for alpha hacking in the full samples, we test for alpha hacking in each of these subsets. So, these different subsets represent different stages of the scale development pipeline (e.g., first construction, modification, revision, translation and use in the field). If most excesses at thresholds were driven by legitimate optimization of the criterion during scale development, we would expect the excesses to be strongest at earlier stages of the pipeline. Instead, we consistently find evidence of alpha hacking at the .70 threshold, and the magnitude of the excesses is quite consistent across datasets and subsets (all 12-15%). This is also in line with Cortina et al.'s finding that researchers frequently alter scales for use in primary research.*

*As we discuss in the response to previous questions above, the behaviors that can be said to represent alpha hacking have already been demonstrated to be common (e.g., item dropping in already established scales); and the impact of alpha hacking has already been shown via simulation study to be quite large (e.g., boosting alpha by an average of .10 under many circumstances, and up to .30 in some). The final thing, which was until now unexamined, was for us to check if the necessary consequences of these behaviours were visible in the wider literature. Our results demonstrate robustly that they are. At the same time, the manuscript is explicit that (a) like p-hacking, alpha hacking is agnostic to intent (i.e., does not imply “nefarious” behavior), and (b) that these excesses in threshold values can likely be caused by one or both of publication bias and hacking. Just as with early investigations into p hacking and excesses of barely significant p values, it will take future work beyond the scope of our manuscript to determine exactly what systems and behaviors give rise to this. The exact contributions of hacking vs publication bias are not known for p values either, but this is no reason to not talk about the fact that there is evidence of either.*

*In the sections on “Influence of construct frequency” and “Influence of measure revision and translation” (pp 22 to 24), we analyze with greater depth the robustness of the results in established measures. The “Possible explanations” section (pp 25-28) expands on the logic the possible (vs. implausible) explanations for the results. Other modifications relevant to this comment have also been documented in the above replies to other comments.*

## Reviewer: 2

Thank you for the opportunity to review this very clear and well put together paper. The supporting documentation and supplementary information are particularly clear. I have very little by way of comments, the paper speaks for itself and I cant find much to criticise it on. I recommend publication as is.

*[R2 point 1]*

If I were forced to make some comments, they would be:

• Maybe add a little nuance to the 0.7 cut off – it isn’t quite as hard and fast rule as 0.05 in frequentist statistics. The work does however provide a strong rationale for why this value requires greater attention, but the discussion could bring out this difference.

*Author response: Thank you - we agree and have therefore added some nuance on this point. Page 5 now reads:*

*“While this threshold is the most common, previous work has demonstrated that a wide range of descriptive labels are used to describe an even wider range of α values. For example, Taber (2018) found that α values ranging from 0.45 to 0.98 have all been described as “acceptable” by authors. As such, while α > .70 is a sufficiently common threshold for our analyses here, it is not as ubiquitous as an α value of .05 for p values, and does not preclude authors from describing their α as "acceptable" (or other descriptors) in a looser everyday sense.”*

*[R2 point 2]*

• Following on from this – the 0.6 cut off seems quite prominent in the beta regression approach, this could be subject of slightly more discussion.

*Author response: While we agree and were somewhat intrigued by this result too, out of an abundance of caution against overfitting on our own data, we have elected not to delve into this point. The .70, .80, and .90 thresholds have the benefit of being preregistered in the I/O dataset and the same code being used for the PsycTests dataset. This provides a strong test of the hypothesis that alpha hacking is occurring. As we now discuss in the “Prevalence of α-hacking” section on page 28 onwards, our aim here is to provide a useful hypothesis test rather than to estimate the prevalence of alpha hacking. The preregistered thresholds provide the strongest test of that, without risking weakening our evidence in the eyes of sceptics. We have therefore elected not to further expand the discussion of the results at other thresholds, although the supplementary materials do include plots for the Beta regression and the calliper tests which may suggest evidence of an excess at .60 too.*

*[R2 point 3]*

• A little more information on the number of publications from each journal would be interesting, and some metrics indicating the prevalence of articles reporting Cronbach’s alpha values within each.

*Author response: Thank you for this suggestion. We have expanded our reporting of the number of journals, articles, and estimates for each dataset. For example, the PsycTests dataset is now described on pages 10-11 as:*

*“The full PsycTests dataset contains data from 60,491 scales reported in 71,692 articles published in 3,159 journals between 1896 and 2023.” (continued page 14) “38,885 out of the original 106,397 instances were excluded. 67,512 α estimates were extracted that were deemed to be valid. 70.1% of DOIs in the database produced at least one α estimate.”*

*In addition to this, the titles of Figures 1 to 3 now include the number of alpha values that were included in that analysis (see Figure 1 to 3).*

*[R2 point 4]*

• Going further than this, some levels of suspected ‘alpha hacking’ in each of the separate journals would also be of interest. I do see the risks with this however – sample size issues and unnecessary criticism of each individual journal is maybe best avoided.

*Author response: We would dearly have loved to conduct such analyses, in addition to changes over time and other questions of who and when. Unfortunately, as you correctly point out, the count of alpha values at the thresholds which are used in our analyses preclude meaningful analyses at the level of the journal or over time. If only!*

## Reviewer: 3

The authors present an interesting study that looks at the distribution of Cronbach’s alpha coefficients in the psychology and I/O literature. They find evidence for a “bump” at several desired thresholds, which may indicate some form of publication bias and/or questionable research practices.

I think this is an interesting paper and a nice extension of the work that has been done on “bumps” in p-value distributions. Automatically extracting statistical information from text can be painstaking and error-prone process (I speak from experience), and I am impressed by the way the authors built in several rounds of manual checks and safeguards to make sure they only extracted actual Cronbach’s alphas.

There were a few points I would like to authors to clarify.

*[R3 point 1]*

1. It would be useful to have a bit more detail about the two databases. Of course, these weren’t the authors’ own projects and referring to the original sources for details is fine, but at the moment, the description in the manuscript is not sufficient to get a clear idea of what these databases looked like.

a. For the both data sets, I would like to see more information on the included journals. In the psychology dataset: mention [all apa journals? Zoek op].

*Author response: Thank you, these were very useful suggestions. The supplementary materials now contain complete lists of the journals included in the (APA) Psychology and (metaBUS) I/O datasets (see Tables 1S and 2S).*

*[R3 point 2]*

b. For the psychology data set, some sources seem to have been mixed up. The list of journals in the supplementary materials does not match the journals analyzed in Nuijten et al. 2015; in the latter publication, only 5 APA journals were analyzed (JPSP, JAP, JCCP, DP, and JEPG). This project included about 30,000 articles (not >70,000 as the authors describe). Please update the description of the sample.

*Author response: Thank you for catching this! We obtained the psychology dataset from Chris Hartgerink a few years ago. At the time he said it was the statcheck dataset from Nuijten et al. 2015, but after checking ourselves and with him we now see that there were multiple statcheck datasets across different publications. The dataset used here is the APA journals curated as part of Hartgerink (2016, “688,112 Statistical Results: Content Mining Psychology Articles for Statistical Test Results”). That full dataset, including data from other journal publishers, appears to no longer be publicly available. I gather that this may be due to the publishers. We have made use of the largest subset of it that we could access: that from APA psychology journals. Our number of articles (down to double digit difference due to some missing cases) and number of journals matches that reported in Hartgerink 2016, and he has confirmed this is the dataset we made use of and that it is now cited appropriately.*

*References have been corrected throughout the manuscript.*

*[R3 point 3]*

b. In the metaBUS set: how representative is the set of selected journals for the field of I/O?

&

c. For the metaBUS data set, I would like a summary of 1) what this project is, 2) what the full database looks like (the authors refer to “variables” and “correlations on diagonals”, but without more context, I don’t know how to envision this). At the moment, the lack of detail made me struggle to understand the sentence “To prepare for our analyses, the database was reduced to its variable-level analogue where each row represented one variable rather than one effect, resulting in 208,369 unique variable instances.” (p. 9).

*Author response: Thank you for this suggestion. On reflection, we see that our previous descriptions of the metaBUS dataset were not sufficiently clear. Part of the description we did give, as you quoted, probably served to muddy the waters more than to clarify them. We have substantially revised the way the I/O dataset is described. Page 9 to 10 of the manuscript now reads:*

*“Distortions in the distributions of α-estimates in the I/O literature were assessed using the metaBUS database (version 2018.09.09). In short, the metaBUS project (Bosco et al., 2017, 2020) seeks to curate (i.e., extract, classify, and make available) all zero-order effects from primary studies in I/O research (e.g., applied psychology and management) to facilitate future meta-analyses. Over 90% of zero-order effects reported in I/O are correlation coefficients, so the extracted values are highly representative of results reported in the field of I/O. Extraction of numerical values is semi-automated, but classification of the values into their constructs is fully manual. The metaBUS project covers 27 journal titles selected based on their impact factor in the areas of applied psychology and management according to ISI’s Web of Science Journal Citation Reports in the year the project began. Years of coverage vary by journal and range from 1980-2017. The full metaBUS dataset contains data from 14,038 articles published in 27 journals between 1980 and 2017. A list of all journals included in the dataset can be found in Table 2S in the Supplementary Materials. Full details of the dataset’s curation and utility can be found in the original publications (Bosco et al., 2017, 2020).*

*Each row of the database represents one numerical value extracted from a correlation matrix in a published article. In general, correlation matrices report the zero-order correlations between a given number of variables. Most fields of psychology report only either the upper or lower triangle of correlations in order to avoid redundancy. Unlike some other fields of psychology, articles in the I/O field often also report reliability estimates in the diagonal of correlation matrices (e.g., in the correlation matrix, rather than leaving it blank, the element representing the association between a given measure and itself reports the Cronbach’s α for that measure). Entries in the metaBUS dataset were extracted from correlation matrices reported in the I/O literature. To date, uses of the metaBUS dataset have made use of the correlations reported in the non-diagonal elements. This study is the first to make use of the large number of reliability estimates reported in the diagonals of these matrices that are included in metaBUS. It is these values of Cronbach’s α estimates that were used in the present analyses. A variety of other meta-data is available in the database, including each original study’s sample size, sample type, country of origin, publication year, and construct classification. For details on the metaBUS database architecture see Bosco et al. (2017); for information about the method and reliability of extractions see Bosco and colleagues (Bosco, Aguinis, et al., 2015; Bosco, Steel, et al., 2015). Two journals were included in both the psychology dataset (1985 to 2013) and the I/O dataset (1980 to 2017): the Journal of Applied Psychology and Journal of Occupational Health Psychology. We discuss this overlap later.”*

*[R3 point 4]*

2. I had some questions about some of the statistical analyses that the authors used.

a. p. 12: in order to judge whether the sensitivity analysis of the I/O articles with the overlapping articles removed led to the same conclusion, I would also report here whether you still found significant “bumps”; not only whether the effect size estimates changed.

*Author response: Thank you for this nudge to report fully - I think we were trying to be concise, but as you point out there is a risk that this lack of detail may have been obscuring something. I have added the full results for the robustness tests, which show both the % bumps and the results of the permutation tests (all were significant). Page 18 now reads:*

*“At the time of preregistration, the two datasets were understood to be non-overlapping. Upon obtaining the I/O dataset we discovered that two APA psychology journals were included in both datasets (Journal of Applied Psychology and Journal of Occupational Health Psychology; see Tables S1 and S2 in the Supplementary Materials), albeit using a wider range of years and a very different extraction method in the I/O dataset. We elected not to deviate from our preregistered analyses of the I/O dataset. As a robustness test, we report the results of the same analyses applied to a non-overlapping dataset (i.e., removing all DOIs from the I/O dataset that were already present in the psychology dataset) in the Supplementary Materials. The conclusions of the preregistered analyses in the I/O dataset were not affected by the removal of these articles: the proportion of excess α values differed by ≤ 1% between analyses (14% excess at .70, Z = 4.68, p = .00007; excesses across the three bins, Z = 4.84, p < .00001, 3% excess at .80, 1% excess at .90). See Note 2S and Figure 2S in the Supplementary Materials.”*

*[R3 point 5]*

b. p. 14: the fact that you found significant excesses in each of the subgroups, does not necessarily mean that the subgroups did not differ from each other. The test you describe does not seem to answer the question.

*Author response: Thank you for pointing out this inaccuracy. To put our cards on the table: we have given this a significant amount of thought and cannot come up with a method to directly compare the magnitudes of the excesses between the groups, as we’d ideally like to do. Our analysis here isn’t merely (the absence of) moderation by group type, as the excesses are themselves derived from the residuals on the kernel smooths in each subgroup. This makes testing (lack of) moderation extremely non trivial. Nonetheless, as you point out, it is important that the reporting of these results does not commit a common statistical fallacy. We’ve therefore changed the language to more accurately reflect the analyses and appropriate conclusions. Page 23 now reads:*

*“In each subgroup, we applied kernel smoothing using the same method as previously and calculated the residuals at α = .70. Statistically significant excesses were found in each subgroup, all ps < .013, indicating that α-hacking was present in ad hoc measures, reused measures, and highly reused measures. Descriptively, the excesses were of similar magnitudes across subsets (i.e., 12-15%), perhaps suggesting that α-hacking was not more prevalent in ad hoc measures. However, the equivalence of excesses between the subsets could not be tested directly (i.e., no statistical method of doing so was known to us), and so this must be interpreted with caution as a descriptive comparison.”*

*[R3 point 6]*

c. p. 14: if, a priori, you could already expect that distributional differences would make the Caliper test less suitable for your situation, why use one at all? How can we interpret the outcomes?

*Author response: Thank you for this point. You’re right that we needed more detail here. We have expanded the rationale for (1) our choice of the kernel smoothing method over the caliper method, and yet (2) our desire to also include the caliper method. Our choice to include them was, I suppose, an attempt to try to preempt readers’ desire to see the same method that was applied to p values in the past, even if we found that method to be imperfect. We also overcome some of the limitations of caliper tests by calculating caliper ratios between all neighbouring bins and applying a permutation tests assess if ratios are larger at thresholds. Pages 24 to 25 now read:*

*“Previous research on the overabundance of barely significant p values has employed caliper tests, which count the number of estimates in two bins of equal width on either side of a cut-off (Hartgerink et al., 2016). We judged these tests to be less suitable than the kernel smoothing method above on the basis that there are plausible distributional differences between adjacent bins (i.e., the distribution of α values is non-uniform, see Figures 1 to 3). Of course, this was also the case when this analysis was applied to p values, and is not specific to our analyses (i.e., in the presence of non-zero effects, the distribution of p values is also non-uniform). Regardless of whether they are applied to p values or α values, the logic of the caliper ratio test is the same: (1) in the absence of any distortions in the distributions caliper ratios should not be expected to be zero; (2) nonetheless, substantial deviations from zero can be usefully interpreted as evidence of distortions., The caliper tests retain utility here because we calculate a caliper ratio for each bin (e.g., 70 vs. .69, .69 vs. .68, etc.). Although we cannot expect any ratio to be exactly zero given a non-uniform distribution of α values, it is still useful to ask whether the ratios at threshold values are larger than non-threshold values. Permutation tests were therefore used to test this, similar to the residuals from the kernel smoothing method. For the sake of comparability with previous work on distortions in the distributions of p values, and as a secondary test to assess robustness to the analytic method, we therefore also implemented caliper tests. See Note 3S and Figures 6S-12S in the Supplementary Materials. In summary, the pattern of excesses at α = .70 was robust to the choice of analytic method (.69 vs. .70 caliper ratios: psychology = 1.71, I/O = 1.64, PsycTests = 1.60). The collective excesses at all three thresholds were not robust in the I/O dataset (.79 vs. .80 caliper ratios: psychology = 1.16, I/O = 1.13, PsycTests = 1.19; .89 vs. .90 caliper ratios: psychology = 1.02, I/O = 0.96, PsycTests = 0.98).”*

*[R3 point 7]*

3. The authors explain in the Discussion that alpha hacking may easily lead to alphas “overshooting” the exact .70 threshold. So what could be an explanation for the abundance of the exactly .70 instances? Could it also have something to do with reporting and/or wrongly rounding up (intentionally or not). To illustrate: in the Hartgerink et al. paper the authors cite, there was a bump in \*reported\* p-values just below .05, but not in \*computed\* p-values just below .05. This suggested that authors wrongly round down p-values so that they fall on the desired side of the cut-off. Could something like that have happened here as well?

*Author response: We think the overabundance at .70 exists because researchers aim to meet or exceed this threshold. Our analyses can detect the results that meet it, but not any additional ones that exceed it. An explanation for the abundance of excesses at exactly .70 is that there is a larger degree of alpha hacking occurring and our analyses detect just some of it.*

*As you suggest, simply rounding up the observed alpha value to the desired threshold could be one form of alpha hacking. Unfortunately, Cronbach's alphas are not reported redundantly with test statistics as is the case with p values, so this is not possible to check in the way it is for p-values. We now acknowledge this and other forms of alpha hacking on pages 32:*

*“While item dropping is the most common form of post hoc scale modification (Cortina et al., 2020), it is of course just one of many potential forms of α-hacking. Cortina et al. (2020) also identified several commonly reported methods of self-report scale modification which can be applied after data collection and may therefore be exploited for α-hacking, including item dropping, creating composites, dichotomizing, reverse coding items, or otherwise altering the scoring strategy. Other potential methods of α-hacking likely have direct analogs with p-hacking, such as inappropriate rounding, selective reporting (either of α values at all, or between multiple measures), outlier exclusion, favorable imputation, or subgroup analysis (Stefan & Schönbrodt, 2023). Future research could examine the degree to which different α-hacking strategies bias in-sample α values (as has been done for p values and standardized effect sizes: Stefan & Schönbrodt, 2022).”*

*In addition to this, we acknowledge the possibility of rounding followed by inappropriate reporting of trailing zeros. In our discussion of data extractions, pages 11-12 now read:*

*“Fourth, we extracted potential α estimates from each character string such that it must follow one of several variations of “α = .XX” where at least two numerical characters were reported. We did not extract α values reported to just one decimal place because they could too easily be affected by (questionably appropriate) rounding (e.g., observing α = .65, rounding and reporting this as α = .7). We however noted that it was very uncommon to observe α values reported to only one decimal place. Nevertheless, this choice of selecting only α values reported to at least two decimal places can pick up on inappropriate rounding followed by inappropriate reporting of trailing zeros (e.g., observing α = .65, rounding this to α = .7, and then reporting as α = .70). This is a feature rather than a bug: such forms of rounding followed by inappropriate reporting of trailing zeros would represent a clear form of α hacking that we would want to detect, agnostic to whether it was accidental or intentional.”*

*[R3 point 8]*

4. In the Discussion, a solution is proposed that validation research and primary research should be explicitly separated, to avoid alpha-hacking. This point I don’t understand: given that the authors also found an excess of .70 in often used measures (that are likely already validated), doesn’t this indicate that validated scales are also checked again in primary research? And in such cases, wouldn’t the authors of the primary research still reach the “.70” threshold?

*Author response: This is a good point regarding the separation being necessary but not sufficient. What we meant was that post hoc modifications of scales are especially problematic when decisions about these modifications are made using only in-sample data without confirmation in new samples (e.g., does dropping a given item actually improve reliability?). Of course, the benefits of such separations are only possible when the new studies are not also hacked. We have expanded our recommendations section in the Discussion to discuss what might limit alpha hacking in future, and that separation of development and use, like preregistration, is necessary but not sufficient. Pages 35 to 36 now read:*

*“Just as our understanding of the risks of p-hacking lead to a greater distinction being made between exploratory and confirmatory research (e.g., Munafò et al., 2017), we believe that the risk of α-hacking and other Questionable Measurement Practices require researchers to make a greater and more explicit separation between measure development and measure use in primary research. That is, modifications to measures should not be made post hoc; any proposed changes should be confirmed in new samples; and the details of measures should be fixed in preregistrations along with other elements of the design and analysis. Put another way, suggested modifications to scales based on a given dataset should not have those modifications be applied to that same dataset, as this is conditioning decisions on results. Instead, suggestions for modifications (e.g., item dropping) should be confirmed in new samples, and then applied consistently in future. The timing and rationale for such choices is only knowable with increased transparency (e.g., through preregistration). Of course, as with p-hacking, increased transparency through preregistration and a clear separation of exploratory and confirmatory research (aka measure development and use with regard to α-hacking), can increase the detectability of hacking but will not automatically prevent it. Increased transparency about the nature and timing of decisions is necessary but not sufficient to prevent hacking. These and other suggested changes to our measurement practices, how authors comply with them, and how editors and reviewers may enforce them, are discussed in greater detail in the Standardisation Of BEhavior Research (SOBER) guidelines (Elson et al., 2023). There are many differences between p-hacking and α-hacking, although the cure may often be the same: increased transparency about which researcher choices were planned (e.g., through preregistration) and which were data-dependent.”*

*[R3 point 9]*

5. p. 7: “Only the first apparent α estimate was extracted from each instance of a character string to avoid duplication.” But could it still happen that multiple alphas were extracted from the same paper? There might be some dependency there, it might be good to mention this in the text.

*Author response: yes, multiple estimates could be extracted from each paper, and therefore there is some dependency in the data (same as the statcheck papers or other papers that you or Christ Hartgerink have written using similar datasets). You raise a good point - we have acknowledged this explicitly in the manuscript now. Page 12 now reads:*

*“Only the first apparent α estimate was extracted from each instance of a character string to avoid duplication, although multiple character strings could be detected and extracted from each article. As such, there were dependencies among the extracted α estimates.”*

*[R3 point 10]*

6. p. 9: “282 rows were removed for which erroneous coding information was identified (i.e., the country in which the data was collected)”. What does this mean? What happened to the country?

*Author response: We thank you for bringing this to our attention. The metaBUS database was filtered by its curator prior to being provided to use for our analyses, in order to only provide us with the relevant subset of the dataset (i.e., reliability estimates reported in the diagonals of correlation matrices, and not non-diagonal values representing zero order correlations). In doing this, he hopefully removed rows with missing data for variables that thought might been useful as moderators in our planned or post-hoc analyses, including country of data collection. However, no such moderator analyses were, in actuality, either planned or conducted by the analytic team, due to our concerns that there were too few counts at the thresholds to break results down further (e.g., by country). This line therefore merely represents full reporting of all our exclusions.*

*As you point out, this may be overly honest reporting: the line probably serves to confuse readers more than it informs them. These 282 of 89,926 rows represent 0.31% of the database. By the time the authors responsible for the analysis fully understood this fine grain detail of the (unnecessary) exclusion, we felt it was more appropriate to simply use this dataset rather than go back and do another database pull and filter to include these additional 282 cases rows.*

*The revised manuscript now simply states on page 13-14:*

*“To prepare for our analyses, only a subset of the metaBUS dataset was used: those rows of the database that referred to reliability values reported in the diagonals of correlation tables. 92,725 reliability values were extracted and sent to us. Of them, 89,926 (97.0%) were of the coefficient α type. Only α estimates from psychometric scales relating to psychological constructs, subjective reports, performance measures, behaviors, and attitudes were employed in the current analyses (i.e., but not for demographic variables or other non-psychological constructs). Finally, 282 rows were removed due to erroneous or missing data as identified by the metaBUS curators. This resulted in an analyzable dataset of 89,644 α values.”*

*[R3 point 11]*

7. p. 11: typo in stats? Double p-value: “Z = 3.15, p = .01042, p < .00001”

*Author response: Corrected*

*[R3 point 12]*

8. Also, when I ran statcheck on your paper, I found three inconsistent p-values:

Z = 3.15, p = .01042; computed p = .00163

Z = 3.94, p = .00035; computed p = .00008

Z = 4.53, p = .00016; computed p = .00001

*Author response: Thank you for your diligence here. Note for the editor: I liaised with Prof Nuijten about why statcheck might be detecting these as errors despite the fact that (having triple checked) they appear to be correct. The answer seems to be that the Z scores returned by the coin package’s permutation tests are approximated rather than exact. This causes false positives in statcheck. A footnote on page 17 now reads:*

*“Note that the Z scores returned by the coin package’s permutation tests are approximated rather than exact. These may cause false positives in statcheck’s reporting error detection.”*

*We have also added a thanks to Prof Nuijten in our acknowledgements for helping us understand statcheck’s behavior here.*