An aberrant abundance of Cronbach's alpha values at .70

Ian Hussey, Taym Alsalti, Frank Bosco, Malte Elson*, & Ruben Arslan*

Cronbach's alpha (α) is the most widely reported metric of the reliability of psychological measures. Decisions about an observed α's adequacy are often made using rule-of-thumb thresholds, such as α of at least .70. Such thresholds can put pressure on researchers to make their measures meet these criteria, similar to the pressure to meet the significance threshold with *p* values. We examined whether α values reported in the literature are inflated at the rule-of-thumb thresholds (α = .70, .80, .90), due to, for example, overfitting to in-sample data (α-hacking) or publication bias. We extracted reported α values from two very large literatures covering general psychology (>30,000 α values taken from >74,000 published articles in APA journals) and Industrial and Organizational psychology (>89,000 α values taken from >14,000 published articles in I/O journals). The distributions of these values show inflation at the rule-of-thumb thresholds. We discuss the scope, causes and consequences of α-hacking and how increased transparency, preregistration of measurement strategy, and standardized protocols could mitigate this problem. Code and data available at osf.io/pe3t7. Supplementary materials at osf.io/5xzy4.

A measure's reliability refers to the proportion of variance that is caused by the construct rather than noise (Allen & Yen, 2002, p.73). Reliability places a limit on the measure's validity (Murphy & Davidshofer, 2005) or, put another way, reliability attenuates observable associations between scores on any two measures: the less reliably the two variables are measured, the lower the observable correlation between the two variables. All else equal, higher reliability therefore increases statistical power to detect associations.

Reliability can be quantified in different ways but the most common metric by far is Cronbach's α (1951). It is based on inter-item correlations, i.e. their internal consistency. Under certain assumptions (e.g., tau-equivalent items, independent error) it converges with reliability (i.e., of tau-equivalence of items; Cortina, 1993). While in practice, α is often incorrectly treated as being synonymous with reliability (Cortina, 1993; Schmitt, 1996), it is the most commonly reported metric of reliability, and indeed is typically the only metric of structural validity reported (Flake et al., 2017). There is a well-established literature debating the use and misuse of α. Much of which focuses on the fact that many researchers inappropriately use it to test properties such as unidimensionality and homogeneity that are in fact assumed by α (i.e., it assumes rather than tests tau-equivalence; Cortina, 1993). Alternatives to α with relaxed assumptions have been suggested e.g., McDonald's ω (McDonald, 1999) along with repeated calls to use it over α, although apparently without much success (Flake et al., 2017). This paper adds to those concerns by examining whether α values reported in the psychological literature show signs of inflation, e.g. due to publication bias or hacking.

Rule-of-thumb thresholds

Cronbach's α is commonly interpreted using well known rules-of-thumb thresholds (e.g., α > .70). Nunnally & Bernstein (Nunnally & Bernstein, 1994) recommended an α value of at least .70 and their book and its earlier 1967 or 1978 editions are frequently cited, often omitting however the qualification that .70 is recommended for "early stages of research" (Lance et al., 2006). Nonetheless, their book remains a highly cited source for this threshold, with over 8000 citations at time of writing. Many, if not most, contemporary undergraduate introductory textbooks on research methods include rules of thumb, and regard α > .70 as something "researchers are looking for" (Morling, 2017, p. 131), "satisfactory" (Howitt & Cramer, 2020, p. 241),

[*] Shared last author

or "a good measure of internal consistency" (McQueen & Knussen, 2013, p. 389; see also Breakwell et al., 2012, p. 149; Howitt & Cramer, 2020, p. 241). Psychologists have used α > .70 as a binary decision rule for scale development for decades. Cortina observes that "[the] acceptance of α > .70 as adequate is implied by the fact that α > .70 usually goes uninterpreted. It is merely presented, and further scale modifications are seldom made." (1993, p. 101).

Publication bias and hacking

Similar to *p* values (Gigerenzer, 2018), when a rule-of-thumb becomes an important criterion for the publishability of findings, the pressure to meet the criterion mounts. This can be desirable, if the criterion itself is an indicator of quality, and the strategies scientists use to meet it increase the robustness of research, for example increasing sample sizes to improve precision of estimates. However, if the metric can also be inflated illegitimately or "hacked", some researchers will do so, wittingly or unwittingly. Provided such hacks are cost-efficient, they will spread at the expense of the qualities actually sought to improve (Bakker et al., 2012; Smaldino & McElreath, 2016). One way in which hacking of metrics can become apparent, is when the distribution of the metric in aggregate deviates from plausible statistical distributions. For instance, Masicampo and Lalande (2012) observed, as they called it, "a peculiar prevalence of *p* values just below .05" in published research articles in three leading psychology journals. Hartgerink et al. (2016) provided additional evidence for an over-abundance of barely-significant *p* values in some journals and biased reporting of *p* values across all journals, using a much larger sample of articles and journals (i.e., all articles published in APA journals from 1985-2013). While there has been debate about whether publication bias alone is a sufficient explanation of these over-abundances (Lakens, 2015a, 2015b), regardless of the specific cause or causes, distortions in the distributions of published estimates bias inferences.

The current research

Use of α shares several similarities with *p* values: we suspect it and its rules of thumb are used for decision making purposes, the incentive structures in scientific publishing reward reporting some results over others, and references to pressure on researchers to obtain estimates that meet the threshold of .70 without further consideration of the implications of reliability date back at least 25 years (Schmitt, 1996). Anecdotal reports of α-hacking, illegitimate tricks to inflate α, abound. We therefore sought to examine the empirical distribution of reported Cronbach's α coefficients, analogously to work on over-abundance of barely-significant *p* values (Hartgerink et al., 2016; Masicampo & Lalande, 2012). We hypothesized that there would be an excess of α values at commonly used rule-of-thumb thresholds values (α = .70, .80, .90) relative to other values.

Method

Transparency statement

All code, processed data, and preregistration are available (osf.io/pe3t7) along with a Supplementary Materials document (osf.io/5xzy4).

Data sources

We examined α estimates in two different literatures and datasets, one covering the psychology literature and one covering the Industrial-Organizational (I/O) literature (i.e., applied psychology, management). These datasets were mostly non-overlapping (10.1% overlap) and used very different extraction methods. The analytic method was developed using the psychology dataset. R code for the analysis was preregistered prior to obtaining the I/O dataset. The analysis of the I/O dataset therefore represents a stronger, confirmatory assessment of the hypotheses. Despite this movement from exploratory to confirmatory analytic strategies, we consider it useful to define the analysis of the I/O literature as an assessment of the generalizability of the effect to what is arguably a different population rather than a replication (i.e., a second sample drawn from the same population). The substantive differences in the methods of extracting α estimates from the two datasets represent a second reason why the two analyses may be better conceptualized as an assessment of generalizability.

In order to assess distortions in the distribution of α values in the psychology literature, we made use of a dataset of the full text of all articles published in APA journals between 1985 and 2013. A list of all journals included in the dataset can be found in Table 1S in the Supplementary Materials (see osf.io/5xzy4). This dataset was previously used to assess reporting errors using the StatCheck program and further details of this dataset can be found in the original publication (Nuijten et al., 2015). The full dataset contains 74,470 articles covering all major areas of psychology research including clinical, social, personality, cognitive, experimental, developmental, educational, and applied psychology.

Distortions in the distributions of α-estimates in the I/O literature were assessed using the metaBUS database (version 2018.09.09). The full metaBUS dataset contains data from 14,038 articles published in 26 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 effect (i.e., correlation coefficient) extracted from a published correlation matrix. Many articles in this field report reliability estimates in the diagonal of correlation matrices. It is these values of Cronbach's α that were used in the present analyses. A variety of other meta-data is available in the database, including sample size, sample type, country of origin, publication year, construct classification, and the like. 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.

Data extraction

Psychology dataset

α estimates were extracted from the psychology dataset using regular expressions, which are sequences of characters that specify search patterns in text. These were implemented using the R package *stringr* (Wickham & RStudio, 2022). Our approach was therefore similar to that employed by Nuijten et al. (2015) in their original extraction of *p* values and test statistics from the dataset, although our exclusion criteria were necessarily more conservative because of the less standardized way in which α values are reported.

The general strategy was as follows: First, we defined multiple patterns of interest (e.g., variations of "Cronbach's α"). Variations included but were not limited to whether an apostrophe was used, the use of α/a/alpha, and reference to "Cronbach's α" vs. "Coefficient α". Second, we searched the full text of all articles in the dataset for occurrences of these patterns. Third, for each occurrence found, we extracted the text 50 characters prior to the occurrence and 50 characters after it. This provided a much smaller dataset of character strings, each of which may or may not contain an α estimate. 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. Only the first apparent α estimate was extracted from each instance of a character string to avoid duplication. Fifth, we applied a large number of exclusion criteria to each character string to exclude everything other than α estimates.

These exclusions prioritized specificity over sensitivity: that is, we prioritized excluding all non-α estimates and accepted that some valid α estimates would be excluded as a result of this. Some of the most important of these exclusions ensured that references to threshold values were excluded and not mistaken for occurrences of α estimates (e.g., "according to Nunnally (1967), a Cronbach's α of 0.70 is seen as acceptable for…"). Threshold criteria included but were not limited to references to any mention of variations of the phrase "cut-off criteria", comparisons (words such as "exceeded"), ranges ("between"), plurals (e.g., "αs for the subscales ranged from 0.5 to 0.8"), the presence of *p* values (which would suggest the α value was not Cronbach's α but the α value associated with a hypothesis test), and other metrics of reliability (ϰ, ω, etc.). These exclusion criteria were refined and added to through an iterative approach involving rounds of manual inspection of the extracted strings and α estimates. Two researchers inspected (a) every text string from which an α estimate at one of thresholds was extracted (.70, .80, .90) and (b) a random sample of 100 text strings from non-cutoff estimates to exclude non-valid or incorrectly extracted α estimates. If any non-valid extractions were found, the implementation of the exclusion criteria was updated to cover similar cases and a new round of manual inspections was conducted. All regular expressions for exclusions can be found in the R code (osf.io/pe3t7). 26,744 out of the original 60,153 instances were excluded. 33,409 α estimates were extracted that were deemed to be valid. 16.1% of articles in the dataset produced at least one α estimate.

Industrial-Organizational (I/O) dataset

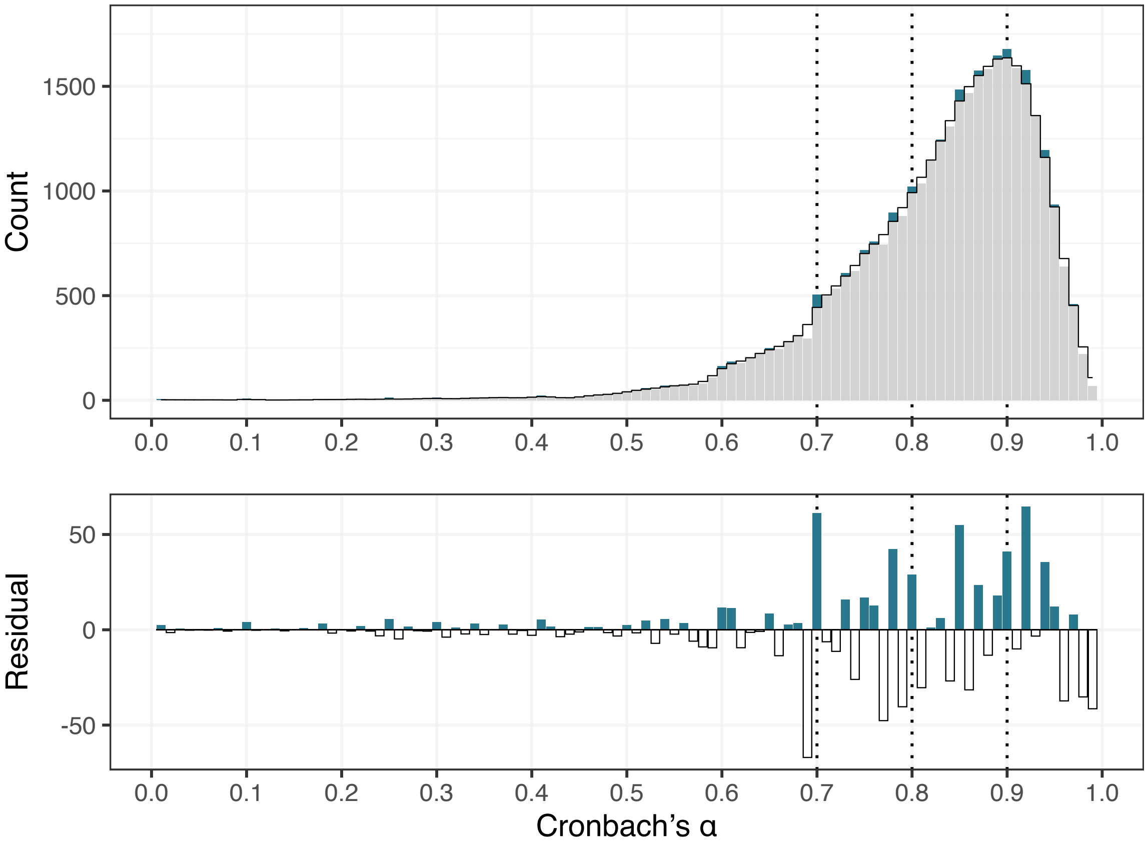
The metaBUS dataset already included the extraction of α values via a different method to that used in the psychology dataset: semi-automated extraction of estimates from correlation tables reported in manuscripts. In the I/O literature, reliability estimates are often reported in the diagonal of correlation tables, and it is these extracted estimates that we made use of in our analyses. All extracted estimates were inspected by trained graduate student raters, who also manually coded other details (e.g., whether the reliability estimates were Cronbach's α or another reliability metric, taxonomization of the constructs measured, etc.). Approximately ten percent of each coder's entries were checked on a weekly basis by a supervisor. See Bosco et al. (2017) for a full description of the curation of the metaBUS dataset.

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. Of these, 92,725 (44.5%) presented with reliability values and, of them, 89,926 (97.0%) were of the coefficient α type. α estimates from psychometric scales relating to psychological constructs, subjective reports, performance measures, behaviors, and attitudes were employed in the current analyses. This rate of data missingness was expected as many variables in metaBUS related to demographics variables rather than psychological constructs (e.g., chronological age) and thus did not contain reliability information. Finally, 282 rows were removed for which erroneous coding information was identified (i.e., the country in which the data was collected), resulting in an analyzable subset of 89,644 α values.

Analytic strategy

Although the distribution of single α values is known (van Zyl et al., 2000), the distribution of multiple α values that are derived from measures that differ in their sample sizes and number of items in unknown ways is not. As such, we employed a data-driven approach. The logic of our analysis was therefore that there is, at minimum, reason to believe that a large

Figure 1. Observed counts of α values with kernel smoothing (upper panel) and residuals (lower panel) in the psychology dataset.



sample of α values should follow a smooth (albeit unknown) distribution. Deflections from such a distribution, especially at a small number of a priori points (i.e., commonly used thresholds), would represent evidence that reported α values are being influenced by some other variable (e.g., α-hacking). On the basis that previous work examining the prevalence of barely-significant *p* values employed caliper tests, we also report them as robustness tests (e.g., Hartgerink et al., 2016). Only the kernel smoothing approach in the I/O dataset was preregistered.

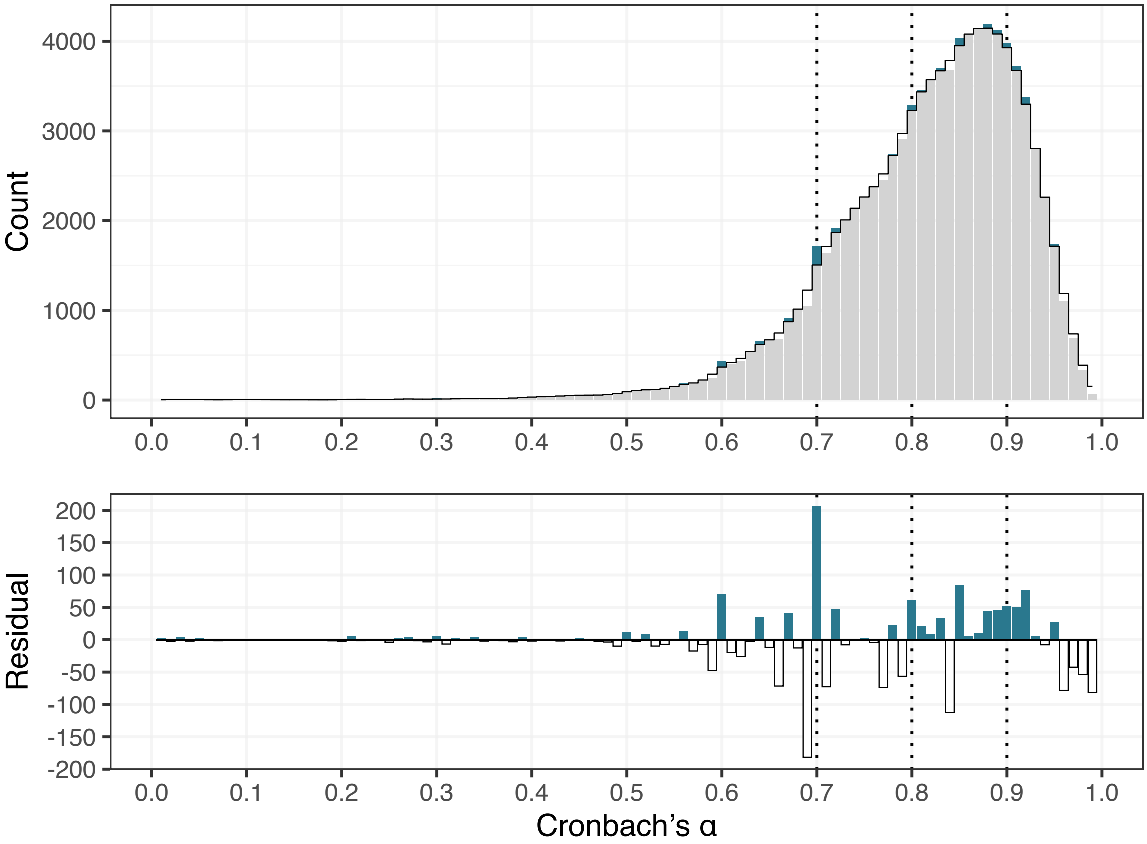
Results

Kernel smoothing

We applied kernel smoothing to the extracted *α* estimates in order to estimate their distribution and quantify the excess of α values at the thresholds. Kernel smoothing was selected over other modeling approaches because it involves relatively fewer assumptions and demonstrated better fit to the observed αs than alternatives. Results of an exploratory Beta regression model that was fit to the psychology dataset can be found in the Supplementary Materials (Note 1S and Figure 1S).

The extracted α estimates were rounded to two decimal places (using the half-up method and the R package janitor: Firke et al., 2021). The rounded α estimates were then converted to counts for each value of α. Density was estimated at 99 equally spaced bins in the interval (i.e., from 0.01 to 0.99). We opted for the default options in R's "density" function: gaussian kernels with a smoothing bandwidth set using Silverman's rule of thumb (Silverman, 1986; i.e., the settings kernel = "gaussian" and bw = "nrd0"). All other options for kernels that are available within R's density function were explored within the psychology dataset. However, as expected with large sample sizes (Sheather, 2004), the choice of kernel did not have a noticeable impact on the resulting density distribution in the psychology dataset. The bandwidth was chosen based on Silverman's rule-of-thumb, which seemed to provide the best fit to the data as it yielded a relatively narrow bandwidth, which is appropriate for large sample sizes (Trosset, 2009, p.172). These analytic choices and the code implementing them were explored in the psychology dataset and then preregistered for the analysis of the I/O dataset. In the case of the bandwidth, we preregistered the actual bandwidth returned in the psychology dataset for use in the I/O dataset (i.e., bw = 0.01). That is, we preregistered the bandwidth to be used rather than the bandwidth determination method. The observed counts and fitted smoothed curves for the psychology and I/O datasets can be found in Figures 1 and 2 (upper panels) respectively.

Figure 2. Observed counts of α values with kernel smoothing (upper panel) and residuals (lower panel) in the I/O dataset.



Bins corresponding to the rule-of-thumb thresholds where we hypothesized that excesses would be found are colored in blue and non-thresholds are in gray.

In each dataset, residuals were calculated for each of the bins. That is, we calculated the excess or deficit of the observed count of each bin (in gray or blue in Figures 1 and 2, upper panels) relative to its predicted value according to the smoothed curve (in black). These residuals are plotted by themselves in Figures 1 and 2 (lower panels).

Two nested hypotheses were tested regarding excesses at .70 (Hypothesis 1) and excesses at .70, .80, or .90 (Hypothesis 2), on the basis that .70 is the single most commonly employed rule-of-thumb thresholds, but .70, .80, and .90 are all common. These hypotheses were tested using independence permutation tests implemented using the R package *coin* (Hothorn et al., 2021). The magnitude of the excesses or deficits (i.e., the residuals) was quantified by converting the observed and predicted counts to proportions.

Tests of the first hypothesis in each dataset compared the .70 bin against all other bins. In the psychology dataset, a 14% excess of α values of .70 relative to other values was found, $Z = 3.15$, $p = .01042$, $p < .00001$. Tests of the second hypothesis in each dataset compared the .70, .80 and .90 bins against all other bins. The hypothesized excesses were found across the three bins, $Z = 3.94$, $p = .00035$, with an excess at .80 = 3%, and at .90 = 3%.

Both of these effects were found to generalize to the I/O dataset, for which we preregistered verbal hypotheses and the code implementations of their inference tests. The test of the first hypothesis found a 14% excess of α values of .70, $Z = 5.01$, $p < .00001$. The test of the second hypothesis found excesses across the three bins, $Z = 4.53$, $p = .00016$, with an excess at .80 = 2%, excess at .90 = 1%. We therefore rejected the null hypothesis that there was no evidence of excesses of α values at common rule-of-thumb thresholds.

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. Conclusions of the preregistered analyses in the I/O dataset were not affected by the removal of these articles (proportion of excess α values differed by ≤ 1% between analyses). See Note 2S and Figure 2S in the Supplementary Materials.

Influence of construct frequency

We considered it plausible that α-hacking might be more common with newly created ad hoc measures than frequently used ones that may have more well-established items, scoring strategies, etc. It was possible to explore this in the I/O dataset as the metaBUS extraction process included manual labeling the construct that each α estimate came from using a taxonomy (see Bosco et al., 2020). We performed exploratory subgroup analyses in measures of constructs (a) that occurred only once in the dataset (i.e., ad hoc measures that were not reused in future studies), (b) that appeared more than once (i.e., non-ad hoc measures reused in future studies), and (c) that appeared more than 100 times (i.e., frequently studied measures, whose cut-off was chosen based on the distribution of frequencies: see Figure 3S in the Supplementary Materials). In each subgroup, we applied kernel smoothing using the same method as previously and calculated the residuals at α = .70. Statistically significant excesses of similar magnitudes were found in each subset: all excesses 12-14%, all *p*s < .013. This suggested that *α*-hacking was not more prevalent in ad hoc measures. Unfortunately, the sample size did not allow for any meaningful analysis of changes in excesses over time.

Caliper tests

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 either side of a cut-off (Hartgerink et al., 2016). We judged these tests to be less suitable for our current purposes 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 and 2). Still, we implemented caliper tests as a secondary test for the sake of robustness, see Note 3S and Figures 4-7S 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). 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; .89 vs. .90 caliper ratios: psychology = 1.02, I/O = 0.96).

Discussion

Across two very large databases, we observed excesses in the proportions of *α* values at a commonly-used threshold criterion (α = .70). These excesses were observed in both the psychology and I/O literatures. When estimated using kernel density smoothing, the magnitudes of the excess of α values of .70 was 14% in both datasets. When using caliper ratios, the method used in previous work on the excess of significant *p* values (Hartgerink et al., 2016), the magnitude of the counts of *α* = .70 versus .69 was also large (psychology = 1.71, I/O = 1.64). Excesses at other thresholds (α = .80 and .90) were smaller and less robust to the choice of analytic method.

Do these excesses show that something is amiss, or are psychologists just exceptionally good at precisely calibrating their study design and data collection efforts to meet this criterion? We believe calibration is extremely implausible because at typical sample sizes (e.g., 50 to 500) and number of items in a scale (e.g., 3 to 50), the standard error of Cronbach's α from around .02 to .08 (see Table 3S and Note 4S in the Supplementary Materials; van Zyl et al., 2000). This precludes effective calibration as an explanation for the combination of a dearth of α values at .69 and excess at .70, because estimates in typical studies are not estimated precisely enough to reliably make this distinction. In this sense, the distribution of αs is suspicious, much like a player in Blackjack who gets exactly 21 too often.

Publication bias is more plausible, especially in the form that authors, editors and reviewers may be less inclined to publish studies if included scales do not meet the .70 criterion. Publication bias is less obviously problematic for α values than for *p*-values. We do not want the scientific literature to be filtered by statistical significance, but might desire a literature filtered for measures with high reliability. However, the estimation precision of α noted above precludes a purely benign review process that selects for high population reliability. Instead, publication bias would also act on stochastic variation of the in-sample estimates. We observed an excess at .70 also for well-established measures (i.e., those used more than 100 times). In such cases, it is clear that publication bias would inflate our impression of the reliability of these scales. Other aspects of the data are also at odds with this benign explanation of striving for highly reliable measures. It follows, if publication bias for minimum α exists, it will exert pressure on researchers to increase their α values in accordance with Goodhart's law (Strathern, 1997, p. 268). If they can take shortcuts to doing so, i.e. α-hacking, some are likely to do so (Flake et al., 2022). One piece of evidence for hacking was the noticeable deficit of values at .69 (see Figures 1 and 2), as it seems implausible that editors and reviewers would discriminate against .69 more than .68.

We believe α-hacking is a likely but potentially only partial explanation for the observed distribution of α values. As with *p*-hacking, field norms may be partially unclear on which practices are problematic. Clearly, rounding up α values is inappropriate, but, for example, some researchers may incorrectly believe that dropping or reversing items ad-hoc is benign or even helpful, even without then validating these changes in independent data.[1] The willingness to carry out such modifications is likely to be influenced by existing incentives (e.g., to report reliability exceeding common thresholds).[2]

Of course, α-hacking and publication bias are not mutually exclusive, and we suspect both play a role. There have been comparable debates about the causes of excesses of barely-significant *p* values (see Hartgerink et al., 2016). However, the cure may often be the same: increased transparency about which analyses were planned (e.g., through preregistration) and which were data-dependent.

For measures, this may often mean that the hard work of scale development is more explicitly separated out from primary research. This way, hard questions such as the tradeoffs between internal consistency (which, when high, can represent a form of redundancy), participant time, and construct breadth can be explicitly investigated, and the resulting scales be validated in independent data. For such work to become more commonplace, field norms may have to change.

Potentially, measurement-related Questionable Research Practices (aka Questionable Measurement Practices: Flake & Fried, 2020), which we call α-hacking, such as ad-hoc modifications to scales are currently perceived to be as permissible as some *p*-hacking practices were before the publication of Simmons, Nelson, & Simonsohn (2011). However, ad-hoc measures and ad-hoc modifications to standardized measures may have more pernicious and further-ranging consequences than expected. α-hacking does not just inflate the perceived reliability of our measures but also reduces the replicability of any effects based on those measures, particularly so when techniques to increase α (e.g., dropping one or more items) remain unreported. Relatedly, statistical power is a function of reliability (Heo et al., 2015; Parsons, 2018) and will be overestimated when α is inflated. Additionally, given that psychometric meta-analyses adjust for reliability, α-hacking would bias their results (Schmidt & Hunter, 2015; Wiernik & Dahlke, 2020).

Improvements of the in-sample estimates of α are not worth the cost of decreased comparability to existing work, unlikely generalization of the α boost to replication studies, and less accurate estimates of the population value of α. For example, if a scale originally consisted of 7-items and each study dropped two different items in their analysis, then only three items would overlap between studies. This may exacerbate issues of measurement invariance, and may strengthen the appearance of homogeneity of findings while actually inflating their heterogeneity as differently modified measures will decreasingly overlap in their content validity (see Elson, 2019 for further discussion of this general problem caused by flexible measures).

It is worth considering how our results can speak to the potential scale and severity of any underlying distortions in reported α values more generally. Our analyses can only detect distortions in α values that produce an α at a given value (i.e., .70, .80, .90), but not those that overshoot those thresholds. For example, if a researcher were to inappropriately drop an item in order to increase in-sample α from .66 to .72, this distortion would not be detected as an excess by our analyses. Equally, if many reported α values were upwardly biased from the .60s to the .70s, for example due to item dropping or other forms of overfitting on in-sample data (see Cortina et al., 2020), our analyses would not detect this as no comparisons are made against the (unknown) true distribution of α values. As such, our results may well represent just the tip of the iceberg of biases in reported α values. There are again useful links with the literature on biases in reported *p* values here: significant methodological investment has produced multiple methods by which we can detect and even partially correct for *p*-hacking and publication bias under certain assumptions, but our understanding of whether those assumptions hold up in real life is at yet limited (Carter et al., 2019; Renkewitz & Keiner, 2019).

Further research is needed to estimate the full extent of bias in published α values. For example, comparisons could be made between α values reported in the published literature and in bias-resistant methods such as Registered Reports (as has been done with *p* values: Scheel et al., 2021). Similarly, 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).

Limitations

This study examines biases in reported αs at specific thresholds. These analyses cannot speak to any other, possibly broader forms of bias in reported α values. The current results represent a first study which attempts to provide one form of evidence that α-hacking occurs. Future research is needed to consider and examine other forms of α-hacking, and to estimate its prevalence and severity.

The validity of the analysis of the psychology dataset is bounded by the validity of our extraction of α estimates and exclusion of all non-α estimates. Our

[1] Item dropping is certainly facilitated by statistical software: when calculating α, both SPSS and the popular R package *psych* (Revelle, 2018) both suggest alternative values for α if that item was dropped.

[2] Although we refer to such practices as α-hacking based on the popularity of α, the same principles would apply to any other reliability metrics (e.g., McDonald's ω, ICC, etc.), in the same way that *p*-hacking is an umbrella term for inference tests. *p*-hacking, as traditionally understood, is a form of overfitting statistical models on the data at hand (Yarkoni & Westfall, 2017).

extraction method therefore prioritized specificity over sensitivity at the level of individual estimates. Although, separately, it should be noted that our approach cannot distinguish between multiple estimates taken from the sample (e.g., α calculated using the full scale and then after dropping an item). On the one hand, this could result in unmodeled dependencies among the data. On the other hand, if items were dropped (or other post hoc modifications were made to the scale) in order to increase α to meet the rule-of-thumb thresholds, this would be appropriately captured by our analyses (e.g., excesses at the thresholds due to α-hacking).This approach was additionally limited by the lack of standardized reporting practices for α in comparison to *p* values. While we have high confidence that only valid estimates of α were included in the final dataset, this was at the sacrifice of sensitivity. Many potentially valid but unclear or difficult to extract α values were excluded. It is possible that this extraction method was biased in some way. Inferences about the true distribution of *α* values in the psychology literature should therefore be made with caution. However, the I/O dataset does not suffer from this issue due to its very different extraction method and the more standardized nature of *α* reporting in those journals (i.e., in the diagonals of correlation tables). The fact that evidence of α-hacking was found in both databases, using very different extraction methods, increases our confidence in the results.

It is important to acknowledge that we studied reported α values, which may not represent the full sample of reliability estimates from the measures employed in the component studies. The reported values may be distorted in ways other than α-hacking around the thresholds, for example: (1) not calculating reliability estimates, which is more common in stimulus-response laboratory tasks than self-report scales and which can hide very low reliability (Lilienfeld & Strother, 2020); (2) under-reporting of α values (Flake et al., 2017); or (3) opportunistically switching to other metrics of reliability (e.g., McDonald's ω, ICC, or split-half reliability).

Our analyses are also limited to distortions at the thresholds. We can say little about the distribution of reported *α* estimates or its correspondence with the true distribution of the reliability of measures in these literatures. The distribution of individual *α* values based on sample size and number of items is known (van Zyl et al., 2000), but not the population of scales which differ in their number of items. Perhaps some features of the observed distribution are due to the legitimate selection and refinement of scales with high *α* values (causing its left-skew) or shortening of scales with very high α due to perceived item redundancy (causing few values above .95). Under these circumstances, it is important to note that the estimates of inflation should not be interpreted as the prevalence of α-hacking, which remains unknown. Simulations could help us understand the severity of the problem under realistic conditions.

Finally, we use the term α-hacking, which should not be misunderstood as connoting intentional deception. Comparable discussions about researchers' intentions in specific cases have been had in the *p*-hacking literature are generally an unproductive distraction (Nelson et al., 2018). We use the term hacking to make clear that plausible explanations for the effect we observed here attribute them to researchers' behaviors which serve to modify an index rather than some passive effect of the system (as with publication bias).

Conclusion

The distributions of Cronbach's α values in large samples from the psychology and I/O literatures show excesses of α values at commonly used thresholds. Features of the distribution suggest that these excesses are not solely driven by a benign selection for high true reliability, but may be biased by publication bias and/or α-hacking. These excesses at the thresholds may only be the tip of the iceberg of biases in reported α values. Just like *p*-hacking, α-hacking occurs when researchers overfit to in-sample data by exploiting researcher degrees of freedom, wittingly or not. Just like *p*-hacking, α-hacking could be reduced through more transparent research practices, tailored to target the specific forms of overfitting, flexibility, and underreporting that give rise to it. Where *p*-hacking has played an important role in the replication crisis in psychology, α-hacking may contribute to a growing measurement crisis (Flake & Fried, 2020; Lilienfeld & Strother, 2020).

Previous research has discussed at length the misuse of α and the issues of using thresholds for decision making, all with very limited impact on the continued (mis)use of α (e.g., Cortina, 1993; Schmitt, 1996; Sijtsma, 2009). We are agnostic to whether *α* and indeed cut-offs should or should not be used. *α*-hacking, in the sense of overfitting to in-sample data, is a different and potentially more pressing problem. However, expediting increased transparency in scale development could, at the same time, lead to more informed choices of reliability coefficients and a less problematic impact of thresholds.

Therefore, future research should more precisely preregister and fully report not only their analytic strategy but their measurement strategy. We echo similar calls for greater transparency made by Flake and Fried (2020). This includes the content and implementation of measures (Heycke & Spitzer, 2019), their scoring, any changes made to them relative to previous studies (e.g., item dropping, rewording, scoring), the methods of quantifying reliability (and other measurement properties), all decision making rules, and any ad-hoc modifications. Of course, such full reporting is much easier if a standardized protocol can simply be cited. Indeed, we believe increased requirements for measurement transparency will also entail increased measurement standardization and thus help psychology mature to become a more integrated science.

Author notes

Ian Hussey (corresponding author: ian.hussey@rub.de), Ruhr University Bochum, Germany. ORCID 0000-0001-8906-7559; Taym Alsalti, University of Leipzig, Germany. ORCID 0000-0002-1767-1367; Frank Bosco, Virginia Commonwealth University, USA. ORCID 0000-0002-3497-4335; Malte Elson, University of Bern, Switzerland. ORCID 0000-0001-7806-9583; Ruben Arslan, University of Leipzig, Germany. ORCID 0000-0002-6670-5658. This research was supported by the META-REP Priority Program of the German Research Foundation (#464488178). Parts of this project related to the metaBUS database were funded by the National Science Foundation (NSF) (1424231), Social Sciences and Humanities Research Council (SSHRC) (869-2013-003), the SHRM Foundation (168), and the VCU Presidential Research Quest Fund. ME and RA are joint last author.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. https://doi.org/10.1037/a0038047

Bosco, F. A., Field, J. G., Larsen, K. R., Chang, Y., & Uggerslev, K. L. (2020). Advancing meta-analysis with knowledge-management platforms: Using metaBUS in psychology. *Advances in Methods and Practices in Psychological Science*, *3*, 124–137. https://doi.org/10.1177/2515245919882693

Bosco, F. A., Steel, P., Oswald, F., Uggerslev, K., & Field, J. (2015). Cloud-based Meta-analysis to Bridge Science and Practice: Welcome to metaBUS. *Personnel Assessment and Decisions*, *1*(1). https://doi.org/10.25035/pad.2015.002

Bosco, F. A., Uggerslev, K. L., & Steel, P. (2017). MetaBUS as a vehicle for facilitating meta-analysis. *Human Resource Management Review*, *27*, 237–254. https://doi.org/10.1016/j.hrmr.2016.09.013

Breakwell, G. M., Smith, J. A., & Wright, D. B. (2012). *Research methods in psychology: Approaches and methods* (4th ed.). Sage.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., Tonidandel, S., Summerville, K. M., Heggestad, E. D., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology*, *105*, 1351–1381. https://doi.org/10.1037/apl0000815

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Elson, M. (2019). Examining Psychological Science Through Systematic Meta-Method Analysis: A Call for Research. *Advances in Methods and Practices in Psychological Science*, *2*(4), 350–363. https://doi.org/10.1177/2515245919863296

Firke, S., Denney, B., Haid, C., Knight, R., Grosser, M., & Zadra, J. (2021). *janitor: Simple Tools for Examining and Cleaning Dirty Data* (2.1.0). https://CRAN.R-project.org/package=janitor

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist*, *77*, 576–588. https://doi.org/10.1037/amp0001006

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. https://doi.org/10.7717/peerj.1935

Heo, M., Kim, N., & Faith, M. S. (2015). Statistical power as a function of Cronbach alpha of instrument questionnaire items. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0070-6

Hothorn, T., Winell, H., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2021). *coin: Conditional Inference Procedures in a Permutation Test Framework* (1.4-2). https://CRAN.R-project.org/package=coin

Howitt, D., & Cramer, D. (2020). *Research methods in psychology* (6th ed.). Pearson.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology / Psychologie Canadienne*, *61*, 281–288. https://doi.org/10.1037/cap0000236

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

McDonald, R. P. (1999). Test homogeneity, reliability, and generalizability. In *Test theory: A unified approach* (pp. 76–120). Lawrence Erlbaum Associates.

McQueen, R. A., & Knussen, C. (2013). *Introduction to research methods and statistics in psychology: A practical guide to the undergraduate researcher* (2nd ed.). Sage.

Morling, B. (2017). *Research methods in psychology: Evaluating a world of information* (3rd ed.). WW Norton & Company.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological Testing: Principles and Applications*. Pearson/Prentice Hall.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. https://doi.org/10.3758/s13428-015-0664-2

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd edition). McGraw-Hill.

Parsons, S. (2018). Visualising two approaches to explore reliability-power relationships. *Preprint*. https://doi.org/10.31234/osf.io/qh5mf

Renkewitz, F., & Keiner, M. (2019). How to Detect Publication Bias in Psychological Research. *Zeitschrift Für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 25152459211007468. https://doi.org/10.1177/25152459211007467

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. SAGE Publications, Ltd. https://doi.org/10.4135/9781483398105

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350

Sheather, S. J. (2004). Density Estimation. *Statistical Science*, *19*(4), 588–597.

Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Routledge. https://doi.org/10.1201/9781315140919

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Stefan, A., & Schönbrodt, F. (2022). *Big Little Lies: A Compendium and Simulation of p-Hacking Strategies*. PsyArXiv. https://doi.org/10.31234/osf.io/xy2dk

Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4

Trosset, M. W. (2009). *An Introduction to Statistical Inference and Its Applications with R* (UK ed. edition). Routledge.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*(3), 271–280. https://doi.org/10.1007/BF02296146

Wickham, H., & RStudio. (2022). *stringr: Simple, Consistent Wrappers for Common String Operations* (1.5.0). https://CRAN.R-project.org/package=stringr

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94–123. https://doi.org/10.1177/2515245919885611

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on*

Psychological Science, *12*(6), 1100–1122.
https://doi.org/10.1177/1745691617693393