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 three very large datasets covering the general psychology literature (>30,000 α values taken from >74,000 published articles in APA journals), the Industrial and Organizational psychology literature (>89,000 α values taken from >14,000 published articles in I/O journals), and the APA's PsycTests database which aims to cover all psychological measures published since 1894 (>67,000 α values taken from >60,000 measures). The distributions of these values show robust evidence of excesses at the α = .70 rule-of-thumb threshold which cannot be explained by justifiable scale modifications. 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 are 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 a given variable is measured, the lower the observable correlation between the two variables. All else being 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 them 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 and Bernstein (1994) recommended an α value of at least .70, and their book and its earlier 1967 or 1978 editions are frequently cited for this. However, these citations often omit the qualification that .70 was recommended for "early stages of research" (Lance et al., 2006). Nonetheless, their book remains a highly cited source for this threshold, with over 8,000 citations at the 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,

[*] Joint last author

p. 131), “satisfactory” (Howitt & Cramer, 2020, p. 241), 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 $\alpha > .70$ as a binary decision rule for scale development for decades. Cortina observes that “[the] acceptance of $\alpha > .70$ as adequate is implied by the fact that $\alpha > .70$ usually goes uninterpreted. It is merely presented, and further scale modifications are seldom made.” (1993, p. 101). 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 $\alpha > .70$ 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.

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 aggregated metric 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). 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 use of α shares several similarities with *p* values: it and its rules of thumb appear to be 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. 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 $\alpha = .63$, 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.

The current research

In light of this, all of the same ingredients that allowed for *p*-hacking are apparently also present for α-hacking: measurement flexibility provides opportunity, and the combination of rules-of-thumb and structural incentives in publishing provide motive. It is therefore reasonable to expect that some degree of α-hacking is indeed taking place. However, this hypothesis has not yet been directly tested. Analogously to the work on over-abundances of barely significant *p* values (Hartgerink et al., 2016; Masicampo & Lalande, 2012), this work tested the hypothesis that evidence of α-hacking is present in the literature by examining distortions in the empirical distribution of reported Cronbach’s α coefficients. This approach has also more recently been applied to examine excesses of Area Under the Curve (AUC) values at rules-of-thumb cut-offs (White et al., 2023). We hypothesized that there would be an excess of α values at commonly used rule-of-thumb threshold 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 published literatures, one covering the psychology literature published in APA journals, and one covering the Industrial-Organizational (I/O) literature (i.e., applied psychology, management). These datasets were mostly non-overlapping (10.1% overlap in DOIs) and used very different extraction methods, providing us with convergent sources of evidence. Additionally, we extracted data from a third data source: the APA's proprietary PsycTests dataset, which is designed to be a comprehensive database of all psychological measures since 1896 and includes information about the source and use of each measure as well as empirical findings about its reliability and validity.

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 same analytic code was then used to analyze the PsycTests database. No formal preregistration was made, but the same preregistered hypotheses, tests, and code implementations of these tests were employed as in the I/O dataset. The substantive differences in the methods of extracting α estimates from the datasets represent an additional reason why the analyses may be better conceptualized as an assessment of generalizability.

Psychology dataset

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. The full dataset contains 74,470 articles published in 81 journals covering all major areas of psychology research including clinical, social, personality, cognitive, experimental, developmental, educational, and applied psychology. 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 (Hartgerink, 2016).

Industrial-Organizational (I/O) dataset

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.

APA PsycTests dataset

Distortions in the distributions of α-estimates across known psychological measures were assessed using the PsycTests database (version as of March 1st 2023). The full PsycTests dataset contains data from 60,491 scales reported in 71,692 articles published in 3,159 journals between 1896 and 2023. Full details of the dataset's curation and contents can be found on the APA's website (www.apa.org/pubs/databases/psyctests). Each entry in the database, which was provided in full by the APA, contains a text field that contains unstructured text information about reliability and validity information for the scale that was assembled and curated by APA staff members from the published literature. A variety of other meta-data is available in the database, including language, publication year, item availability and contents, construct classification, and whether a given measure is an original measure or a translation or modification of an existing measure.

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, and are commonly used for searching unstructured text data. These were implemented using the R package *stringr* (Wickham & RStudio, 2022). Our approach was therefore similar to that employed by Nuijten et al. (2015) and Hartgerink (2016) 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. 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. 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. 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 the 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. In its entirety, the metaBUS dataset also includes the correlations reported in correlation tables. In the original creation of the metaBUS dataset, 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, 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.

APA PsycTests dataset

α estimates were extracted from the PsycTests dataset in an almost identical manner to the psychology dataset, given that both involved searching unstructured text for α estimates using Regular Expressions. The field for reliability and validity information for each measure in the PsycTests database reliability was shorter than the full article texts searched in the psychology dataset. In addition to this, the field often reported multiple α estimates within a given sentence. We therefore made one modification to the search strategy compared to the psychology dataset: the code was modified to be able to extract multiple α estimates from a single candidate string of text. Like those from the psychology dataset, these extractions prioritized specificity over sensitivity. The implementation of the exclusion criteria was again refined and supplemented through an iterative approach involving rounds of manual inspection of the extracted strings and α estimates in order to maintain their specificity (i.e., we observed and had to exclude new patterns that represent non-α estimates). Rather than exhaustively searching the extracted estimates, we employed a random sampling method. Two researchers inspected 1005 randomly sampled reliability estimates and their surrounding text to ensure that they represented α estimates. 615 of these values were sampled from threshold values (α = .70, .80, or .90), and 390 from non-threshold values. Where non-α estimates were observed, additional exclusion rules were implemented to exclude all similar patterns. That is, we drew random samples of varying sizes, manually searched for incorrectly extracted non-α values in these samples and, if found, updated the exclusion rules to cover these cases before drawing another sample. The sizes of the samples we drew were based on our expectation of the prevalence of non-α estimates based on the previous round of inspections. Using a sample size calculator implemented in the R package epiR (Stevenson et al., 2023), we started with smaller samples (<100), reflecting our uncertainty regarding the expected false positive rates. As we modified the Regex code to exclude more sources of false positives, we entered increasingly precise expected prevalences when calculating sample sizes. Our final 2 samples for example were 243 (α estimates at the thresholds) and 188 (non-threshold estimates) large, reflecting expected prevalences of 0-3% and 0-4%, respectively. The false positive rate in the last sample we inspected at the thresholds in the non-threshold sample was 1.6%, 95% CI = [0.5%, 4.3]. The false positive rate in the analytic sample was likely lower than this again thanks to the final round of Regex modifications. 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.

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 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.

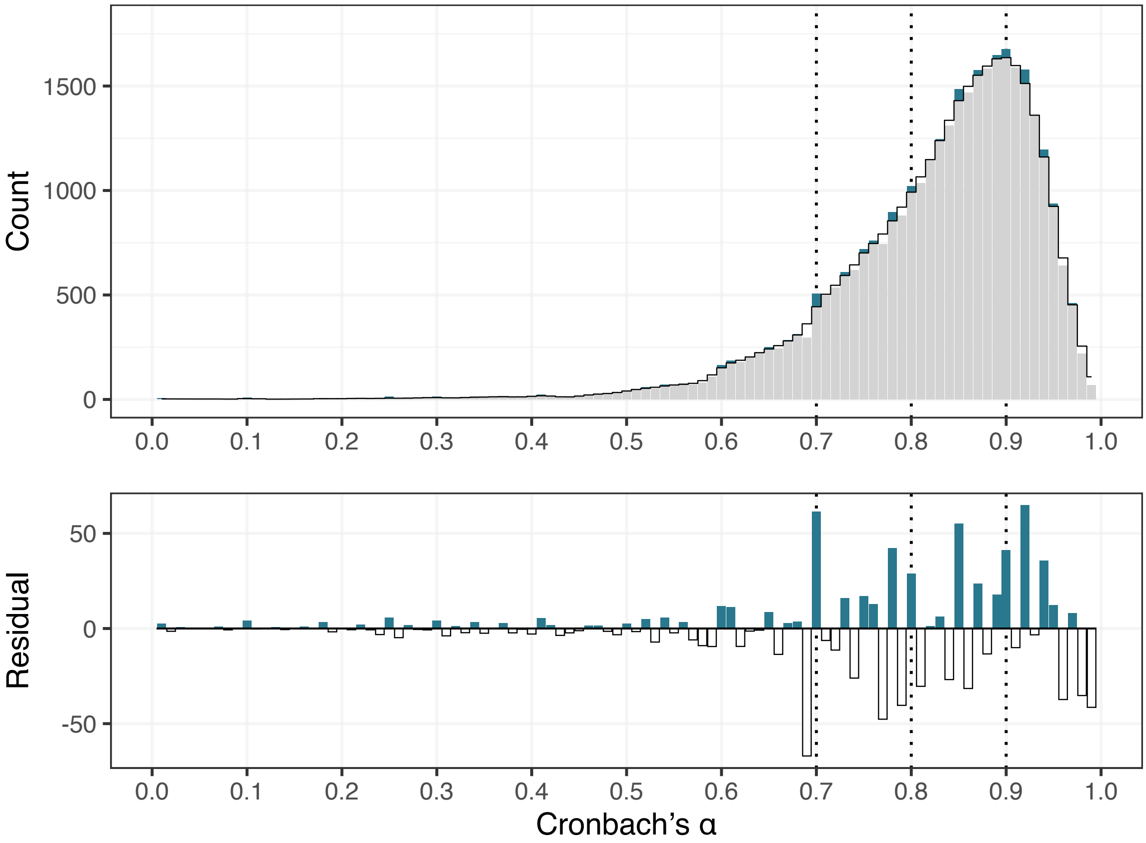
Results

Kernel smoothing and residuals

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

Figure 1. Observed counts of α values with kernel smoothing (upper panel) and residuals (lower panel) in the psychology dataset (33,409 α values).



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 each dataset can be found in Figures 1 to 3 (upper panels) respectively. 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 to 3, upper panels) relative to its predicted value according to the smoothed curve (in black). These residuals are plotted by themselves in Figures 1 to 3 (lower panels).

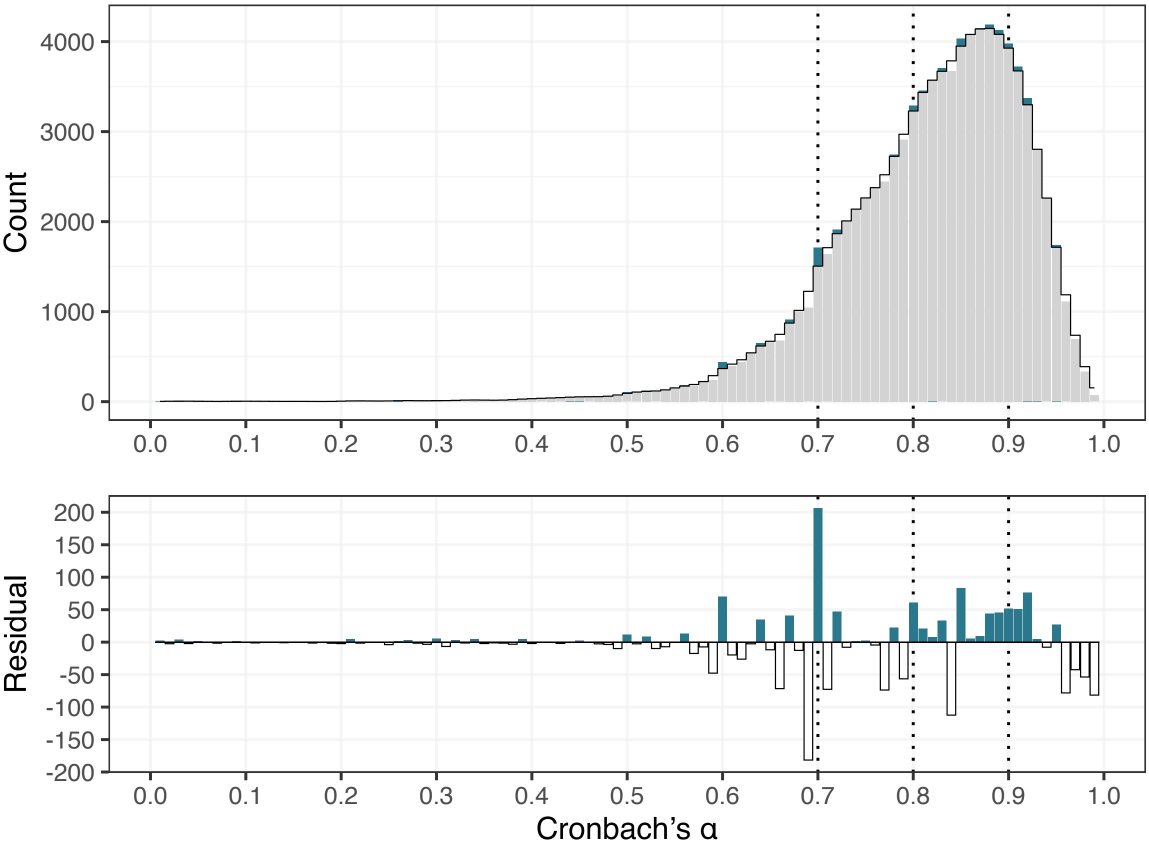
Two nested hypotheses were tested regarding excesses at .70 (Hypothesis 1) and excesses at .70, .80, or .90 (Hypothesis 2), on the basis that .70 is the single most commonly employed rule-of-thumb thresholds, but .70, .80, and .90 are all common. These hypotheses were tested using independence permutation 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, $Z = 3.15$, $p = .01042$.[2] 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

[2] 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.

Figure 2. Observed counts of α values with kernel smoothing (upper panel) and residuals (lower panel) in the I/O dataset (89,644 α values).



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.

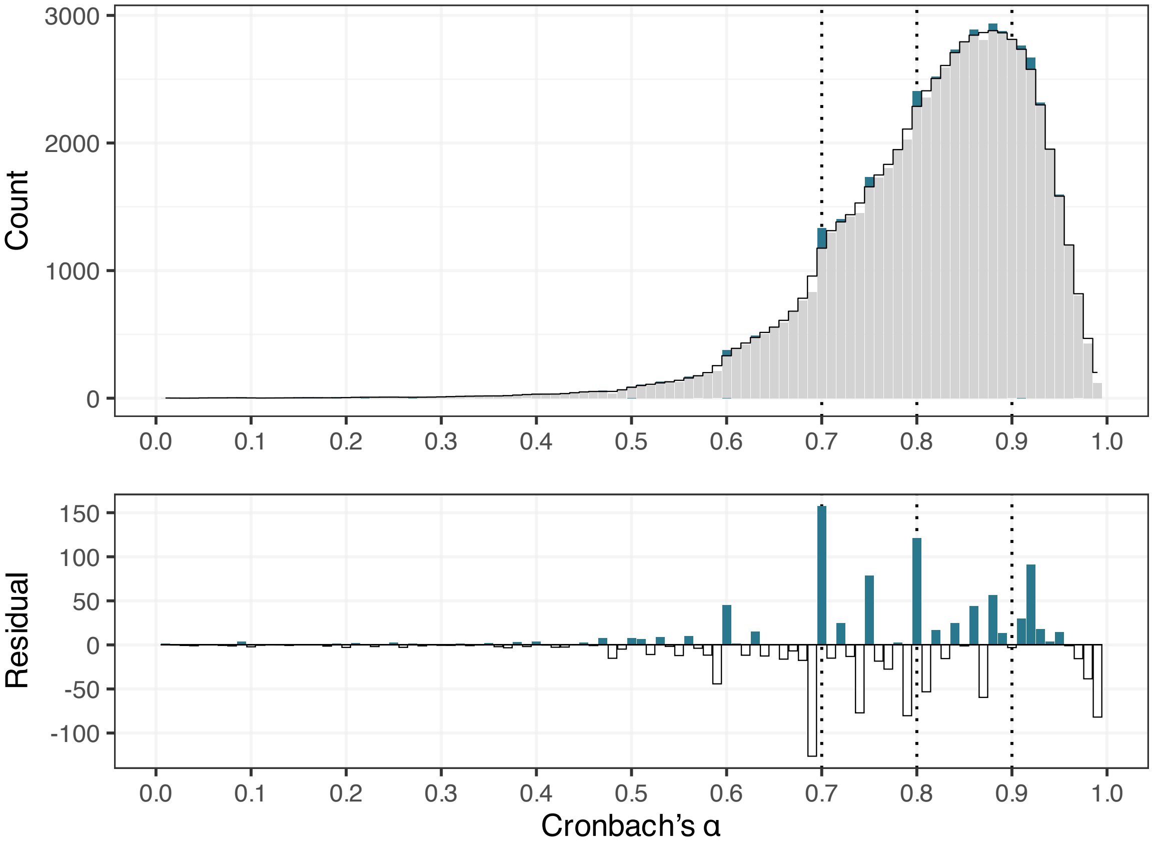
At the time of preregistration, the two datasets were understood to be non-overlapping. Upon obtaining the I/O dataset we discovered that two APA psychology journals were included in both datasets (*Journal of Applied Psych*ology and *Journal of Occupational Health Psychology*; see Tables S1 and S2 in 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.

These effects were also found to generalize to the PsychTests dataset. The test of the first hypothesis found a 12% excess of α values of .70, $Z = 3.94$, $p = .00072$. The test of the second hypothesis found excesses across the three bins, $Z = 3.77$, $p < .00001$, with an excess at .80 = 5%, but no excess at .90 (0%). We therefore rejected the null hypothesis that there was no evidence of excesses of α values at common rule-of-thumb thresholds. The excesses at the thresholds across all three datasets are illustrated in Figure 4.

Influence of construct frequency

We considered it plausible that α-hacking might be especially 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; $N = 34{,}778$ α values), (b) that appeared more than once (i.e., non-ad hoc measures reused in future studies; $N = 43{,}662$ α values), and (c) that appeared more than 100

Figure 3. Observed counts of α values with kernel smoothing (upper panel) and residuals (lower panel) in the PsycTests dataset (67,512 α values).



times (i.e., frequently employed measures, whose cut-off was chosen based on the distribution of frequencies; $N =$ 11,204 α values; see Figure 3S in the Supplementary Materials for the distribution of the frequency of use of measures). 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 *p*s < .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. Unfortunately, the sample size also did not allow for any meaningful analysis of changes in excesses over time.

Influence of measure revision and translation

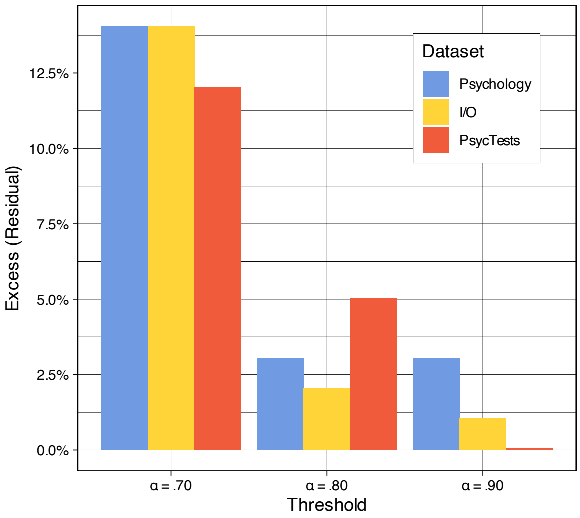
We considered it plausible that the frequency of α-hacking might differ between original measures relative to revisions and translations of existing measures. It was possible to explore this in the PsycTests dataset, which included information about whether a given measure was original vs. a revision or translation of an existing one. We performed exploratory subgroup analyses in (a) original measures (N = 38,774 α values) vs. (b) revisions and translations (N = 19,633 α values) using the same analytic strategy as above (see Figure 4S and 5S in the Supplementary Materials). Statistically significant excesses were found in both subgroups at α = .70 (all *p*s < .0001), indicating that α-hacking was present in both original measures and revised and translated measures. Descriptively, excesses were of similar magnitudes in both subsets (i.e., 12-15%, respectively). As with the previous section, this was a descriptive comparison that must be interpreted with caution, and no meaningful analysis of changes in excesses over time was possible.

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 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).

Figure 4. Excesses (residuals) of α values at the thresholds across the three datasets.



Discussion

The current study provides a test of the hypothesis that published α values are hacked or biased in some way. Results clearly suggest that they are. Across three very large databases covering tens of thousands of measures and publications in psychology and I/O, 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 as well as the measures covered by the APA's PsycTests database. When estimated using kernel density smoothing, the magnitudes of the excesses of α values of .70 were found to be in a consistent range between the datasets and subset (all 12-15%). When using caliper ratios, the method used in previous work on the excess of significant p values (e.g., Hartgerink et al., 2016), the magnitude of the counts of *α* = .70 versus .69 was also large (caliper ratios from 1.60 to 1.71). Excesses at other thresholds (α = .80 and .90) were smaller and less robust to the choice of analytic method.

Possible explanations

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.

One other explanation that we believe should be dismissed is that psychologists are just exceptionally good at precisely calibrating their study design and data collection efforts to exactly meet this reliability threshold. We believe precise calibration to this value is extremely implausible because, at typical sample sizes (i.e., 50 to 500) and numbers scale items (i.e., 3 to 50), the standard error of Cronbach's α ranges from .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 distributions of αs are suspicious, much like a player in Blackjack who gets exactly 21 too often – or indeed a series of studies with p values mostly between .025 and .049 (Simonsohn et al., 2016).

Publication bias is a more plausible explanation. For example, 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 also observed an excess at .70 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. If publication bias for minimum α exists, it would 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 do so, i.e. by α-hacking, some are likely to engage in such Questionable Measurement Practices (Flake et al., 2022). One piece of indirect evidence for hacking was the noticeable deficit of values at .69 in each dataset (see Figures 1 to 3), 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 to one decimal place is inappropriate, but, for example, some researchers may (incorrectly) believe that dropping or reversing items ad-hoc is benign or even helpful.[3] This is a common misconception: in fact, common methods of item dropping generally produce unreliable results (i.e, the item recommended to be dropped has poor replicability in new samples) and artificially increase observed α values by a relatively large degree (e.g., often by 0.10 and by as much as .30: Kopalle & Lehmann, 1997). The willingness to carry out such modifications is likely influenced by existing incentives (e.g., to report high-reliability coefficients, especially those exceeding common thresholds).[4] 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). This debate does not need to be entirely settled for our results to be important. Just as with *p* values, knowing there are meaningful distortions in the literature is enough to motivate us as a field to examine and, more importantly, take steps to prevent both potential sources in the future, so that we produce a less biased or distorted cumulative science.

Prevalence of α-hacking

Our results present strong evidence that α-hacking is occurring in psychological research. It is worth considering whether our results can also speak to the prevalence of such α-hacking. It is important to bear in mind that our analytic strategy can only detect distortions in α values that produce an α at the thresholds (i.e., .70, .80, or .90), but not other values. For example, if a researcher dropped, without good justification, one or more item from a scale with a true α value of .63 in order to create an apparent α of .73 – a plausible degree of bias in α according to Kopalle & Lehmann's (1997) results – this would represent a problematic distortion of the measures' reliability, and yet it would not be detectable by this analytic approach. Our results therefore represent an extreme lower bound of the actual occurrence of α-hacking, likely just the tip of the iceberg. Future research is needed to estimate the prevalence of α-hacking. 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). 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 still limited (Carter et al., 2019; Renkewitz & Keiner, 2019).

Magnitude and consequences of α-hacking

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, according to Nunnally and Bernstein's (1994) guidelines, suitable only for "early stages of research" to being suitable "when making important decisions" (e.g., α > .90). 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.

[3] 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.

[4] 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).

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., α_x and α_y), following the formula for correlation attenuation (e.g., Revelle, 2009):

$$r_{xy} = \rho_{xy} \times \sqrt{\alpha_x \alpha_y}$$

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 their 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. 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 anymore? 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 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).

Limitations

Because our analyses are limited to distortions at the thresholds, we can say little about the true distribution of the reliability of measures in these literatures. While the distribution of individual α values is known to be a function of sample size and the number of items (van Zyl et al., 2000), the distribution of multiple α values from the population of scales is not known due to unknown and likely heterogeneous methods of selection of these scales and their resulting α values. Perhaps some features of the observed distributions are due to the legitimate selection and refinement of scales with high α values (causing its left-skew) or the shortening of scales with very high α due to perceived item redundancy (causing few values above .95).

The validity of the analysis of the psychology and PsychTests datasets is bounded by the validity of our extraction of α estimates and exclusion of all non-α estimates. Our 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 same 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 nonetheless 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 datasets, this was at the sacrifice of sensitivity. Many potentially valid but unclear or difficult-to-extract α values were excluded from the psychology and PsycTests datasets. It is possible that this extraction method was biased in some way. Inferences about the true distribution of α values in the psychology and PsycTests datasets 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).

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 and 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).

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).

Conclusion

The distributions of Cronbach's α values in large samples from three different datasets examining the published psychology literature show excesses of α values at commonly used thresholds. Features of the distributions suggest that these excesses cannot be explained solely by benign selection for high true reliability, but are more likely to be biased by publication bias and α-hacking (i.e., Questionable Measurement Practices). 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. Also 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 as 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 also 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@unibe.ch), University of Bern, Switzerland. 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.

Acknowledgements

We thank the American Psychological Association for their support and for providing access to the APA PsycTests database. We thank Prof. Michèle Nuijten for helping us understand why the results of our permuted Z tests do not pass statcheck despite being correct (i.e., because the Z values of permuted tests are approximated rather than exact, causing statcheck to return false positives).

References

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

American Psychological Association. (2023). *APA PsycTests.* https://www.apa.org/pubs/databases/psyctests

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

Elson, M., Hussey, I., Alsalti, T., & Arslan, R. C. (2023). Psychological measures aren't toothbrushes. *Communications Psychology, 1*(1), 1-4. https://doi.org/10.1038/s44271-023-00026-9

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. (2016). 688,112 Statistical Results: Content Mining Psychology Articles for Statistical Test Results. *Data, 1*(14). https://doi.org/10.3390/data1030014

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

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale Adaptation in Organizational Science Research: A Review and Best-Practice Recommendations. *Journal of Management, 45*(6), 2596–2627. https://doi.org/10.1177/0149206319850280

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.

Lakens, D. (2015a). On the challenges of drawing conclusions from p-values just below 0.05. PeerJ, 3, e1142. https://doi.org/10.7717/peerj.1142

Lakens, D. (2015b). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). The Quarterly Journal of Experimental Psychology, 68(4), 829–832. https://doi.org/10.1080/17470218.2014.982664

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. Advances in Methods and Practices in Psychological Science, 1(2), 259–269. https://doi.org/10.1177/2515245918770

Lakens, D. (2022). *Improving Your Statistical Inferences.* Version 1.0.0. https://lakens.github.io/statistical_inferences/ https://doi.org/10.5281/ZENODO.6409077

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.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1(1), 0021. https://doi.org/10.1038/s41562-016-0021

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. (2009). Chapter 7: Classical Test Theory and the Measurement of Reliability. In Revelle (2009) *An introduction to psychometric theory with applications in R.* https://personality-project.org/r/book/Chapter7.pdf

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. https://doi.org/10.1037/a0033242

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

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science, 16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schimmack, U. (2021). The Validation Crisis in Psychology. *Meta-Psychology, 5.* https://doi.org/10.15626/MP.2019.1645

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. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of *p*-hacking strategies. *Royal Society Open Science, 10*(2), 220346. https://doi.org/10.1098/rsos.220346

Stevenson, M., Sergeant, E., Heuer, C., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., Sebastiani, P., Solymos, P., Yoshida, K., Jones, G., Pirikahu, S., Firestone, S., Kyle, R., Popp, J., Jay, M., ... Rabiee, A. (2023). epiR: Tools for the Analysis of Epidemiological Data (2.0.65) [Computer software]. https://cran.r-project.org/package=epiR

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

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

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

White, N., Parsons, R., Collins, G., & Barnett, A. (2023). Evidence of questionable research practices in clinical prediction models. *BMC Medicine, 21*(1), 339. https://doi.org/10.1186/s12916-023-03048-6

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