The replicability of item dropping based on Cronbach’s alpha

Ian Hussey

Cronbach’s α is the most commonly reported metric of reliability. Popular statistics software packages used to calculate , such as SPSS and the R package psych, present the user not only with the observed estimate in the sample but also a table of what the value would be if a given item was excluded.

# Outline

## Introduction

Replicability crisis in social science.

Most often focuses on replicability of results of inference tests in new data, i.e., congruence of significance of *p* values. Replicability is reduced by questionable research practices.

Calls for increased focus on measurement – flake and fried. Metrics of validity are underreported, and the less likely a test is to be reported the more likely it is to be failed.

Many practices can be labelled as questionable or deleterious. But how bad, in either absolute or relative terms? It is surprising that it was not until very recently that the relative impact of different practices on the false positive rate was explored (felix paper).

Simulations show that common *p* hacking practices such as X or Y increase the false positive rate from 0.05 (implied by an alpha of 0.05 to determine the significance of a *p* value) to XX.

Measurement practices such as item dropping based on cronbach’s α have also been shown to increase the false positive rate of subsequent hypothesis tests, and therefore decrease replicability.

However, less attention has been paid to the replicability of measurement decisions themselves.

Reliability of measurement, influence on observable associations between variables.

Α is the most common metric of reliability.

Item dropping. Essential part of scale development.

A common form of item dropping is dropping the item with the maximum value of cronbach’s α if removed.

However, also often used in substantive studies. Often used and not reported – can’t help develop scale if its not reported, only serves to boost false positive rate.

Most forms of p hacking represent instances of overfitting ones analyses on the data at hand. The same could be said for some measurement decisions such as item dropping: it may increase reliability in the current sample, which would seem to be a good thing, but to what degree are we overfitting on the data at hand? To what degree do item dropping decisions hold in new samples?

# Method

I use a large sample of data (N > 125,000) on 25 individual differences scales that are commonly used in psychology research to test this. I use this real data to simulate pairs of original and replication studies that employed an identical number of participants.

For example, I sampled 100 of the XXX participants who completed the BRS and labelled them as the “original” study. I then sampled the same number of participants (without replacement) and assigned them to be the “replication” study. In both samples, I calculated Cronbach’s-α-if-item-dropped for all items, and then selected the item with the maximum α value as the drop recommendation. I then assessed the congruence of the item drop recommendation between the original and replication study samples.

In order to estimate the long run replicability of item drop recommendation based on max-Cronbach’s-α-if-item-dropped, I performed the above resampling method a large number of times (XXX) for each of the 25 different scales, and using different sample sizes (i.e., 25, 50, 100, 250, 500 in each sample).

Lastly, given that researchers’ decisions to drop items based on Cronbach’s α may or may not be contingent on numerical improvements in Cronbach’s α compared to the full-scale α, I examined both cases. Two variations of the item dropping strategy were implemented: In strategy 1, a single item was always dropped in each sample (original and replication) based on the max-Cronbach’s-α-if-item-dropped. In strategy 2, a single item was dropped based on max-Cronbach’s-α-if-item-dropped only if the max value was higher than the full-scale Cronbach’s α value.

Each resample therefore produced a binary (true/false) data point on whether the item drop recommendation replicated or not. This data was then analyzed in a multi-level logistic model. Replication (true/false) was entered as the dependent variable, sample size (25, 50, 100, 250, 500), strategy (1 vs 2), and their interaction were entered as categorical fixed effects, and scale was entered as a random intercept. Wilkinson notation for the model was:

replicated ~ 1 + sample\_size \* strategy + (1 | scale)

No hypotheses were tested; I sought to estimate the estimate the replication rate of item dropping decisions in each condition. This was done by calculating the marginal means for each condition using the marginalmeans R package (REF). Results can be found in Table 1 and Figure 1.

# Discussion

The replication rate is low, even for large samples. Researchers who drop items based on cronbach’s α if removed,

Limitations. Uses real data, but a limited number of scales. Scales’ properties may not be representative of other fields, as they were not randomly selected. Scales in development may demonstrate a different replication rate.

Resampling could be used by other researchers to examine the robustness of their item drop decisions.

## Outline

Item dropping is an essential part of scale development, and can be legitimately used as part of ongoing scale refinement.

However, it can be used in ways that cannot aid refinement, for example when it is not reported (REF). This can serve to improve α values in a given sample, for example to meet a cut-off value for acceptable reliability (i.e., α hacking) or as an experimenter degree of freedom that may influence the results of a subsequent hypothesis test using scores on the measure (i.e., p hacking).

Both of these represent instances of overfitting or conditioning analyses on their results, which reduces replicability and validity of claims.

Suggestions for “α if item removed” are provided by common statistical software.

Change in in-sample α does not necessarily imply replicable changes in out-of-sample α.

Surprisingly, no work to date has assessed the replicability of these suggestions. That is, when a recommendation for dropping is made on the basis of one sample (the in-sample recommendation), what proportion of time does this agree with the recommendation made in a second sample (the out of sample recommendation)?

3 analyses:

1. if you were set on dropping an item, would the choice of which item to drop replicate?

2. if you dropped items based on α improving, would the decision to drop an item or not replicate between samples? choice of (no drop or same item) vs differ decision

3. if α improves in sample given a specific item drop, does dropping the same item also improve reliability out of sample? This is a slightly less severe test than the previous ones, as there could be improvements but smaller ones than the best item. Akin to 1 vs 2, this one could either consider all cases or just cases where an item was recommended to be dropped relative to the full scale (i.e., dropping optional)

## Cronbach’s α is the most commonly reported metric of reliability. An α value of .70 is often used as a cut-off for acceptability, although there is good reason to be skeptical of both cut-offs generally and this number specifically (REFs). Previous research has argued that this has resulted in pressure on scale developers and users to construct scales that are not merely reliable but which meet this specific α cut-off criterion, sometimes even at the expense of validity (REF). Users of scales are likewise under pressure to report that α in their sample is sufficient that the results of the tests of their substantive hypotheses are not brought into question (REF). This has resulted in a situation for α estimates that is akin to the pressure to produce statistically significant findings. Just as there is a wealth of research demonstrating an excess of barely significant *p* values (REF), recent research has also demonstrated evidence of an excess of α values at exactly the .70 cut-off (REF). α values are therefore not just a metric of reliability, but are also a goal to be achieved within the research process. As with *p* values, once a metric becomes a goal, its validity is undermined (REF).

Distortions in the α values reported in articles serve to undermine the replicability and credibility of the broader literature. From a meta-science perspective, it therefore seems important to understand what gives rise to these distortions. Obviously, much could be said about the incentive structures that provide the pressure to provide pristine results (REF). However, it is also useful to consider more proximal causes, such as feedback from the environment that serves to establish and maintain research practices.

In this paper, I consider one such source: the default output of popular statistical software packages used to calculate α, including SPSS and the R package psych. Both present the user not only with the α estimate observed in the sample but also a table of what α would be if a given item was excluded. This output plausibly serves as an implicit recommendation for how to improve α, or at least highlights one possibly legitimate action the researcher could take in order to do so.

Selection of items from a pool is a key step in the development of a scale (REF), and can be an important part of ongoing scale refinement. However, item dropping also commonly occurs in more ad hoc contexts;

Item dropping often goes unreported in articles (REF), and simulation studies have demonstrated that it can be used to p-hack the results of substantive hypotheses (REF). We can therefore make a useful distinction between item dropping in a way that contributes towards scale development and refinement versus item dropping in a way that cannot make such a contribution (e.g., because it is not even reported) and seems to serve the purpose of boosting the in-sample estimate of α and/or changes the results of inference tests using that scale. Put simply, there are at least some situations in which item dropping can be described as a form of α-hacking (REF).

As with *p*-hacking, α-hacking does not have to be intentional. It is all too easy to self-deceive or condition analyses on results in some way. As such, while its useful to draw a conceptual distinction between legitimate and non-legitimate use of item dropping in order to be able to describe the presence of misuse of this practice, this does not mean that it is possible or useful to try to categorize all instances of item dropping as either legitimate or illegitimate. Instead, we can focus on describing the ….

Flake & Fried emphasis the need for transparency in order to assess questionable measurement practices. Questionable vs answerable, …

# Author notes

## Funding

IH was supported by Ghent University grant 01P05517 and the META-REP Priority Program of the German Research Foundation (#464488178).

## Conflict of interest

None.

# References