How replicable are the recommendations of

“Cronbach’s alpha if item removed”?

Ian Hussey

Cronbach’s 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

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 alpha values in a given sample, for example to meet a cut-off value for acceptable reliability (i.e., alpha 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 “alpha if item removed” are provided by common statistical software.

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

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 alpha 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 alpha 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 ….

# 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