International Journal of Behavioral Medicine (2019) 26:455–456
https://doi.org/10.1007/s12529-019-09803-5

Statistical Guideline #2: Report Appropriate Reliability for your Sample, Measure, and Design

Suzanne C. Segerstrom[1]

Published online: 9 July 2019

Abstract

From the Editors: This is the second column from the Statistics Guru. The Statistics Guru will appear in every issue. In these columns, we briefly discuss appropriate ways to analyze and present data in the journal. As such, the Statistics Guru can be seen both as an editorial *amuse bouche* and a set of guidelines for reporting data in the *International Journal of Behavioral Medicine.* If you have ideas for a column, please email the Statistical Editor, Suzanne Segerstrom at segerstrom@uky.edu.

Keywords Statistical guidelines · Reliability · Generalizability

Introduction

A frequent request from editors and reviewers is for authors to report the reliability of a scale in their sample. The second statistical guideline for *IJBM* is for authors to provide this report for every measure. “Reliability is not a property of a test *per se*, but rather a property of a scale applied in a given context to a particular population” ([1], p. 401). Even within a population, reliability is subject to sampling variability. That is, scale reliability in a sample of adolescents may not be reproduced in older adults or even in a different sample of adolescents. Informed interpretation of statistical analyses relies on scale reliability obtained in the sample under consideration. In addition, authors should consider reliabilities other than Cronbach’s alpha, reliability of change in a scale over time, and reliability ensuing from the consolidation of repeated measurements and report them if applicable.

The general definition of scale reliability from classical test theory is the ratio of true score variance to scale score variance; thus, higher reliability suggests a higher proportion of the obtained score due to the (unobserved) true score. Most authors report Cronbach’s alpha for scales with continuous responses, which seem to comprise the majority of psychological scales (scales with binary responses can be characterized with Kuder-Richardson formula 20 reliability). Alpha is a function of item interrelatedness and the number of items in the scale. Note that alpha is not a measure of internal consistency, which can refer to the dimensionality of the scale. A multidimensional scale can yield a higher alpha than a unidimensional scale [2].

Psychometricians note that alpha makes unreasonable assumptions, particularly that all items have equal true-score variances (the “essentially tau-equivalent” assumption). Alpha also usually underestimates reliability. The greatest lower bound (glb; [3]) and omega [1] have been proposed as alternative measures whose more relaxed assumptions yield a more accurate estimate. Table 1 shows glb, omega, and alpha (using the Factor freeware program [6]; see [1] for how to estimate omega using R) for three scales administered to almost 1,000 undergraduates [7]. Alpha resulted in a downwardly biased estimate of reliability in all cases. Omega was less biased, particularly in the case of the less homogeneous Rumination Scale, which also yielded the largest difference between the alpha and glb. Given the hegemony of alpha, one suggestion is to report alpha along with another, less biased estimate, with confidence intervals if possible [1, 3].

This is one in a series of statistical guidelines designed to highlight common statistical considerations in behavioral medicine research. The goal is to briefly discuss appropriate ways to analyze and present data in the *International Journal of Behavioral Medicine* (IJBM). Collectively, the series will culminate in a set of basic statistical guidelines to be adopted by IJBM and integrated into the journal’s official Instructions for Authors, and also to serve as an independent resource. If you have ideas for a future topic, please email the Statistical Editor, Suzanne Segerstrom at segerstrom@uky.edu.

✉ Suzanne C. Segerstrom
segerstrom@uky.edu

[1] Department of Psychology, University of Kentucky, 125 Kastle Hall, Lexington, KY 40506-0044, USA

Table 1 Alpha, omega, and GLB reliabilities for three repetitive thought scales

Scale	Cronbach's alpha	McDonald's omega	Greatest lower bound
RRQ Rumination [4]	0.910	0.910	0.936
RRQ Reflection [4]	0.907	0.908	0.937
Rumination Scale [5]	0.712	0.717	0.792

Why is underestimation undesirable? Surely this is a case of being conservative, which is usually not a problem, right? In this case, it can be. Reliability is sometimes used mathematically to "correct for attenuation," and a relationship could be overcorrected if reliability is underestimated. More often, reliability is used heuristically to interpret a relationship in the context of its upper bound, which is scale reliability. If the Rumination Scale were (theoretically) correlated $r = .50$ with a measure of depression, the context of an upper bound of .70 (alpha) versus .80 (glb) could be important to interpretation.

Another consideration is study design. For single (e.g., cross-sectional) assessments, scale reliability as described above is adequate. However, sometimes we are interested in scale *change*, not its level at a single time point. In that case, we are not interested in how much the items covary with each other across different people (if one person has a high score on one item, does he or she also have high scores on other items?) but across time (if one item goes up, do the others go up with it?). In longitudinal studies with repeated administrations, it is common to report a range of Cronbach's alphas across the administrations, but this approach does not answer the second question. Generalizability theory, a logical extension of classical test theory, can be easily used to estimate the reliability for change [8]. Multilevel designs also require the consideration of multilevel reliability [9].

Scale reliability, although the most commonly reported type of reliability, is not the only kind that should be considered. Often, an average of repeated measurements of a biomarker, such as blood pressure or salivary cortisol, is taken as a more useful estimate, given the many sources of error variance in a single measurement. However, the "physiometrics," such as reliability, resulting from this approach are rarely reported [10, 11]. Another example relevant to repeated measurement arises from investigation of intraindividual variability; for example, the individual standard deviation has different reliability from the individual mean [12].

In short, it is inadequate to report how a mean or standard deviation (over items or observations) performs in a validation sample. Instead, appropriate measures of reliability should be reported (1) for the analytic sample, (2) for all measures including biomarkers, and (3) taking into consideration the design of the study (cross-sectional, multilevel, or longitudinal).

Compliance with Ethical Standards

Conflict of Interest The author declares that there are no conflicts of interest.

Human and Animal Studies This article does not contain any studies with human participants or animals performed by the author.

Informed Consent There was no requirement for informed consent.

References

1. Dun TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol. 2014;105:399–412.
2. Davenport EC, Davison ML, Liou PY, Love QU. Reliability, dimensionality, and internal consistency as defined by Cronbach: distinct albeit related concepts. Educ Meas Issues Pract. 2015;34:4–9.
3. Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika. 2009;74:107–20.
4. Trapnell PD, Campbell JD. Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. J Pers Soc Psychol. 1999;76:284–304.
5. Martin LL, Tesser A, McIntosh WD. Wanting but not having: the effects of unattained goals on thoughts and feelings. In: Wegner DM, Pennebaker JW, editors. Handbook of mental control. Englewood Cliffs, NJ: Prentice-Hall; 1993. p. 552–72.
6. Baglin J. Improving your exploratory factor analysis for ordinal data: a demonstration using FACTOR. Pract Assess Res Eval. 2014;19:14. Available online: pareonline.net/getvn.asp?v=19&n=5. Accessed 21 May 2019.
7. Segerstrom SC, Stanton AL, Alden L, Shortridge BE. A multidimensional structure for repetitive thought: what's on your mind, and how, and how much? J Pers Soc Psychol. 2003;85:909–21.
8. Cranford JA, Shrout PE, Iida M, Rafaeli E, Yip T, Bolger N. A procedure for evaluating sensitivity to within-person change: can mood measures in diary studies detect change reliably? Personal Soc Psychol Bull. 2006;32:917–29.
9. Geldhof GJ, Preacher KJ, Zyphur MJ. Reliability estimation in a multilevel confirmatory factor analysis framework. Psychol Methods. 2014;19:72–91.
10. Segerstrom SC, Boggero IA, Smith GT, Sephton SE. Variability and reliability of diurnal cortisol in younger and older adults: implications for design decisions. Psychoneuroendocrinol. 2014;49:299–309.
11. Segerstrom SC, Smith G. Methods, variance, and error in psychoneuroimmunology research: the good, the bad, and the ugly. In: Segerstrom SC, editor. The Oxford handbook of psychoneuroimmunology. New York, NY: Oxford; 2012. p. 421–32.
12. Wang L, Grimm KJ. Investigating reliabilities of intraindividual variability indicators. Multivar Behav Res. 2012;47:771–802.