Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 5 (2019) 91–92

www.elsevier.com/locate/hpe

Editorial

How we Underestimate Reliability and Overestimate Resources Needed: Revisiting Our Psychometric Practice

Available online 31 May 2019

Keywords: Cronbach's alpha; Reliability; Internal consistency; Assessment; Resources

Introduction

Across education research, Cronbach's alpha (α) [1] has been our default estimator of 'internal consistency' and 'reliability' (henceforth: ρ), despite longstanding critique including from Cronbach himself ([2], p. 397): "*It is an embarrassment to me that the formula became conventionally known as Cronbach's* α." Basically, α is a function of the number of items or assessors (k) and the *intraclass correlation* (ICC):

$$\alpha = (k * ICC)/[1 + ((k-1) * ICC)]$$

The idea behind *ICC* here is that differences between respondents create a constant standard deviation (*SD*) of scores across k and a constant correlation (r) across pairs of k proportional to these differences. The benefits of this *compound symmetry* (CS) structure – being simpler and less sample-size demanding than more flexible alternatives – are often outweighed by inadequate *ICC*- and ρ-estimates.

Underestimated ρ, overestimated k

Although r and α are widely interpreted as measures of *consistency*, $r = 1$ results in $\alpha = 1$ only if all k have the same *SD*; apart from this unrealistic case, we find horrors. For two items A and B with $r = 1$, we find $\alpha = 0.960$ if SD_A is 1.5 times SD_B and $\alpha = 0.889$ if SD_A is 2 times SD_B. For a measure of *consistency*, which should equal 1 if $r = 1$, this tendency is deeply unsettling. Furthermore, when $k > 2$ (as usual), heterogeneity in r across pairs of k constitutes a second source of distortion in the estimation of ρ by α. In practice, both types of heterogeneity often result in ρ being underestimated and, consequently, k needed to achieve a desired ρ being overestimated (i.e., more items or assessors than needed). Several alternatives to α have been proposed (for an overview, see Ref. [3]). Here, I focus on two alternatives that are easy to compute in software packages that also report α. One of these is McDonald's omega (ω) [4], and another is a variant of α that allows *SD* to vary across k (henceforth: α_{SDA}). This structure is also known as *CS heterogenous* (CSH).

Adjusted α and McDonald's ω as alternatives to α

The aforementioned problem of two items with $r = 1$ resulting in $\alpha < 1$ due to varying *SD* can be easily circumvented by using α_{SDA}: using CSH instead of CS, we find $\alpha = 1$ if $r = 1$ regardless of the difference in *SD* between items. If $k > 2$ and differences in r across item pairs are small, α_{SDA} provides an easy alternative to α; they will yield very similar results whenever *SD* differences are small but can differ substantially when *SD* differences are moderate or large. For instance, for two items with $r = 0.7$, we find $\alpha = 0.824$ with equal *SD*,

Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region.

https://doi.org/10.1016/j.hpe.2019.05.003

$\alpha = 0.786$ if SD_A is 1.5 times SD_B, and $\alpha = 0.719$ if SD_A is 2 times SD_B. Using α_{SDA}, we find 0.824 in all cases. Using α, to achieve $\rho = 0.9$, we would recommend: $k = 4$ with equal SD, $k = 5$ if SD_A is 1.5 times SD_B, and $k = 8$ if SD_A is 2 times SD_B. Using α_{SDA}, our (correct) advice would be $k = 4$ in all cases, half the number of items or assessors compared to α if SD_A is 2 times SD_B.

We can compute α_{SDA} through a mixed-effects linear model using CSH or by computing α on standardized item scores. In the mixed-effects approach, the estimated *ICC* can be used as input in the aforementioned α-formula. In the standardization approach, we simply transform observed item scores into z-scores, after which all items have the same SD (i.e., 1). These two methods should yield the same result when $k = 2$, while the mixed-effects approach tends to be slightly more accurate when $k > 2$.

Although α_{SDA} accounts for varying SD, it does not account for heterogeneity in r across item pairs. If the latter is moderate or large, α_{SDA} may also result in inappropriate estimates of ρ, and ω provides a better alternative that is available in various statistical packages including the zero cost Open Source Jamovi [5] and JASP [6]. However, greater flexibility (i.e., accounting for heterogeneity in both SD and r) comes at the cost of a higher demand on sample size. Although there is some discussion on recommended sample sizes, and sample size demands tend to increase with the number of items, it is probably better not to use ω when the sample size is $N < 100$, even if only three or four items are concerned. However, where sample sizes are much smaller than that, ρ estimation may altogether become tricky business. Finally, one vital assumption underlying α and its alternatives is that the items over which it is computed can be conceived as indicators of the same variable of interest, and this assumption is to be tested with psychometric methods such as item response theory or factor analytic models. Computing α, α_{SDA} or ω over items that measure different things (e.g., knowledge vs. skill, or a series of different OSCE stations for that matter) is generally not recommended. For a more detailed technical discussion of this matter, see for instance Chapters 3 and 13–17 in my new book [7].

Recommendation

All statistics have in common that they are useful under a set of assumptions but can become useless in the face of substantial departures from at least one of these assumptions. Although α, α_{SDA}, and ω may yield very similar ρ-estimates (i.e., differences in the third or fourth decimal) whenever departures from CS are small, they tend to diverge when departures from CS are more substantial. While α_{SDA} provides an easy alternative to α in the face of substantial SD differences, ω is probably the best alternative when there is substantial heterogeneity in r and the sample size is large enough. When in doubt, report all three coefficients, along with item SDs and a correlation matrix, and let the reader decide which coefficient to trust.

Ethical approval

Not applicable.

Funding

None.

Conflict of interest

No conflicts of interest.

References

1. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334. https://doi.org/10.1007/BF02310555.
2. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. *Educ Psychol Meas*. 2004;64:391–418. https://doi.org/10.1177/0013164404266386.
3. Revelle W, Zinbarg RE. Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. *Psychometrika*. 2009;74:145–154. https://doi.org/10.1007/s11336-008-9102-z.
4. McDonald RP. *Test theory: a unified treatment*. Mahwah, NJ: Erlbaum; 1999.
5. Jamovi project. Jamovi version 0.9.6.9. Retrieved from: https://www.jamovi.org/ (Accessed: 23 April 2019).
6. Love J, Selker R, Marsman M, et al. JASP version 0.9.2.0. Retrieved from: https://jasp-stats.org/ (Accessed: 23 April 2019).
7. Leppink J. *Statistical methods for experimental research in education and psychology*. Springer; 1999. https://doi.org/10.1007/978-3-030-21241-4 (Accessed: 3 June 2019).

Jimmie Leppink*
Hull York Medical School, University of York, United Kingdom

**Corresponding author.* Hull York Medical School, University of York, Heslington, York, YO10 5DD, United Kingdom.
E-mail address: hyjl17@hyms.ac.uk

23 April 2019