**Scale reliability: A primer**

Gavin T L Brown, PhD, FAPS

Director Quantitative Data Analysis and Research Unit,

The Faculty of Education and Social Work, *The University of Auckland*

Email: gt.brown@auckland.ac.nz

Scale reliability is an estimate of how strongly a set of items covary with each other. If the value is close to 1.00 then the items are highly related. Generally, values over .70 are needed to support the notion that the scale is a consistent or reliable representation of the construct.

It is noteworthy that reliability estimates can be inflated by using items that are highly homogeneous (i.e., items use repetitions of the same language). Cattell and Tsujioka (1964) call that method a ‘bloated specific’ and scales should minimise items that are repetitious. This requires writing items that reflect the construct but do not duplicate each other. The implication here is that high estimates of reliability may reflect homogeneity rather than factor-trueness, so a lower threshold of reliability may be a good thing.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, *24*(1), 3-30. <https://doi.org/10.1177/001316446402400101>

**Coefficient alpha**

The most commonly used method in psychology and education is alpha (Cronbach, 1951). This measure is an *essentially tau-equivalent* model that assumes constant item variances for the true scores and that true score means and the error variances of the items vary.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

However, assuming that true score variance (equal sensitivity) is constant across all items is improbable and unrealistic. Hence, alpha is an inappropriate measure of internal consistency reliability and will provide a lower estimate of reliability than the population or true level of reliability. Alpha can be inflated when the errors for each item are correlated, a violation of assumptions concerning residuals. Nonetheless, many reviewers and examiners want to see alpha because they are familiar with it even though it is probably wrong. See the following critiques.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399-412. <https://doi.org/10.1111/bjop.12046>

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educational and Psychological Measurement*, *37*(4), 827-838. <https://doi.org/10.1177/001316447703700403>

Green, S. B., & Yang, Y. (2009). Commentary on Coefficient Alpha: A cautionary tale. *Psychometrika*, *74*(1), 121–135. <https://doi.org/10.1007/S11336-008-9098-4>

McNeish, D. (2018). Thanks coefficient alpha, we’ll take it from here. *Psychological Methods*, *23*(3), 412-433. <https://doi.org/10.1037/met0000144>

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154. <https://doi.org/10.1007/s11336-008-9102-z>

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>

Teo, T., & Fan, X. (2013). Coefficient Alpha and beyond: Issues and alternatives for educational research. *Asia-Pacific Education Researcher*, *22*(2), 209–213. <https://doi.org/10.1007/s40299-013-0075-z>

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach’s α, Revelle’s β, and Mcdonald’s ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133. <https://doi.org/10.1007/s11336-003-0974-7>

**McDonald’s omega**

Alternative estimators of scale reliability exist but not all are easily obtained in modern software. The best alternative seems to be McDonald’s omega (1999) statistic. This assumes that the means and variances of the true scores and the error variances vary and no assumption about constant means and variances is made. Under violations of *tau-equivalence*, which are likely to be the norm in psychology, omega outperforms alpha. McDonald’s omega is available alongside alpha in Jamovi and JASP.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

**Coefficient H**

Confirmatory factor analysis is a more powerful way of determining if items belong together as it forces all off-paths to be zero and provides multiple indices of fit to the data. If you have used CFA to create a scale, then it makes more sense to use construct replicability to evaluate how well a set of items represent the latent variable. This approach informs whether an SEM measurement model is suitable and replicable across studies. H values represent the correlation between a factor and an optimally-weighted item composite, with high values (H >.80) indicating a well-defined latent variable, which is likely to be stable across studies. Low values (H < .80) will generate unit-weighted item scores (i.e., not optimally-weighted scores) for the underlying latent variable. Hence, Coefficient H is ideal for a CFA/SEM framework to judge the feasibility of a measurement model given a particular set of items

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and Future - A Festschrift in Honor of Karl Jöreskog* (pp. 195-216). Scientific Software International Inc.

There is an online calculator.

Hammer, J. H. (2016, October). *Construct Replicability Calculator: A Microsoft Excel-based tool to calculate the Hancock and Mueller (2001) H index*.In [calculator]. <http://drjosephhammer.com/wp-content/uploads/2016/10/Construct-Replicability-H-index-Calculator.xlsx>

But the protocol is straightforward and can be done easily in Excel. Create a column for each of the 6 steps and insert the correct formula to automatically calculate H.

* Enter the standardised loading value for each item (this is a beta regression weight)
* Square that value to get proportion variance (*R*2) explained by that loading (closer to 1 is best)
* Calculate the ratio of variance divided by 1 minus that variance (*R*2/1-*R*2); bigger than 1 is good
* Sum the ratios you just calculated (Σ*R*2)
* Add 1 to the sum of ratios (Σ*R*2+1) (Note that if the sum is large then adding 1 doesn’t change it much)
* Divide the sum by the sum plus 1 (Σ*R*2/(Σ*R*2+1) (closer to 1 is best)

Conceptually this is equivalent to Cohen’s *f*2 statistic which is how he converted variance explained (*R*2) into an effect size (*f*2)

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. [https://doi.org/10.1037//0033-2909.112.1.155](https://doi.org/10.1037/0033-2909.112.1.155)

What kind of combinations give .80 or higher? Clearly the fewer items, the stronger loading each item has to have. They don’t all need to be the same; just the combination has to produce the relevant average. But standardised loadings less than .50 are unlikely to generate a robust Coefficient H.

|  |  |  |
| --- | --- | --- |
| # items in factor (k) | Average Loading (λ) | Coefficient H |
| 3 | .75 | .80 |
| 4 | .71 | .80 |
| 5 | .66 | .80 |
| 6 | .63 | .80 |
| 7 | .60 | .80 |
| 8 | .58 | .80 |
| 9 | .57 | .81 |
| 10 | .55 | .81 |
| 11 | .53 | .81 |
| 12 | .50 | .80 |