Neuendorf

Internal Consistency Reliability: Can Cronbach's alpha be Too High?

As noted by Streiner (2003), "one of the central tenets of classical test theory is that scales should have a high degree of internal consistency, as evidenced by Cronbach's alpha" (p. 217). In the past, some sources have argued for attaining as high a Cronbach's alpha as possible. And, some of my own manuscripts have been criticized by journal reviewers for alphas that were deemed insufficient. However, some arguments mitigate against the "high alpha" perspective.

(NOTE: "Internal consistency" is also sometimes termed "homogeneity." And while Cronbach's alpha, based on a set of items' mean interitem correlation, is the most popular statistic testing internal consistency, other such statistics exist, such as the KR-20 for dichotomous variables. Further, some scholars (e.g., Clark & Watson, 1995) argue for the use of average interitem correlation (a "straightforward" test of internal consistency) rather than Cronbach's alpha (which can be artificially inflated by simply adding more redundant measures).)

Is Internal Consistency Reliability Even Relevant in All Cases of Multiple Indicators/Measures? (No.)

The argument may be made that some scales are composed of sets of measures that may not or even should not be correlated. Bollen and Lennox (1991) point out a key difference between (a) multiple indicators as effects of a latent construct, and (b) multiple indicators as causes of a latent construct. An example of (a) might be where the latent construct is self-esteem, and there are four measures of self-esteem. An example of (b) might be the use of possibly diverse measures such as education, occupational prestige, income, and neighborhood economic value as indicators of the latent construct SES (socioeconomic status). In case (a), we would expect the four indicators to correlate, because they all come from the same latent construct. In case (b), we would not necessarily expect the four indicators to correlate, as they may be tapping quite different aspects that actually *define* the latent construct rather than emerge from it. In case (a), we would be concerned with checking for internal consistency reliability; in case (b), we would not.

Similarly, and partly based on Bollen and Lennox's work, Streiner (2003) differentiates between "scales" and "indexes"—he thinks of a scale as a composite of measures that are manifestations of an underlying construct (such as anxiety); he thinks of an index as a type of inventory of various behaviors/attitudes/thoughts (e.g., a list of daily activities). While Streiner's use of the terms "scale" and "index" is not standard (see, e.g., Babbie, 2013), the differentiation he makes is sound.

For example, a student's thesis examined American students' exposure to "foreign" films (Ying, 2009). Her survey questionnaire included a checklist roster of many international movies, and respondents checked all they had seen. e.g.:
Please check the films that you've seen.

____A Very Long Engagement (France)

____About My Mother (Spain)
____Amelie (France)
____Amores Perros (Mexico)
____Antonia's Line (Holland)
____Asterix et Obelix: Mission Cleopatra (France)
____Au Revoir, Les Enfants (France)
____Autumn Tale (France)
____Avenue Montaigne (France)
____Babette's Feast (Denmark)
____Bad Education (Spain)
____Belle Epoque (Spain)
____Betty Blue (France)

The main scale created from these multiple indicators was a straight addition of the number of foreign films seen (i.e., each item was coded as 0=not checked, 1=checked). There was no reason to expect strong intercorrelations among the items. That is, just because a respondent had seen *Amelie*, we would not expect them to have seen *Bad Education* as well. The entire logic of this type of scale is one of a *counting* of activities, thoughts, or the like.

One more example (Measurement, 2001) might be “a sexism scale in which participants were supposed to report the extent to which they have experienced a number of different sexist situations. We would not necessarily expect the experience of one event to be related to experiencing another event. In a case such as this, the [internal consistency] reliability would be somewhat low, yet we may still want to sum the scores to give us an indication of how many events they experienced” (p. 56).

What’s the Problem with High Internal Consistency Reliability (i.e., a High Cronbach’s Alpha)? (Poor Content Validity, “Bad” Measurement)

On the one hand, it’s logical to think that multiple indicators of the same construct should be intercorrelated—if they are not, then how can they be measuring the same thing?

On the other hand, high intercorrelations among the measures might mean that the items are “overly redundant and the construct measured too specific” (Briggs & Cheek, 1986, p. 114).

Thus, high internal consistency may work against content validity, the extent to which a scale taps all aspects of a construct. High internal consistency might mean that only a portion of the construct has been measured—repeatedly! As noted by Clark and Watson (1995), “maximizing internal consistency almost invariably produces a scale that is quite narrow in content; if the scale is narrower than the target construct, its validity is compromised” (p. 316).

Additionally, highly redundant measures in a questionnaire or interview can frustrate the respondent (“Didn’t you just ask me that?”). This can clearly lead to poor measurement, as respondents tune out or get angry at the researchers.

Criteria/Rules of Thumb

As Clark and Watson (1995) note, the issue of internal consistency reliability assessment is complicated by the fact that "there are no longer any clear standards regarding what level. . . is considered acceptable" for Cronbach's alpha (p. 315); past criteria have ranged from .80 or .90 alpha coefficients, down to .60 or .70 alphas.

As noted above, some scholars find Cronbach's alpha to be too sensitive to number of measures/items, and prefer the use of the raw mean interitem correlation (MIC) as a statistical marker of internal consistency. For this, a rule of thumb is offered by Briggs and Cheek (1986): "The optimal level of homogeneity occurs when the mean interitem correlation is in the .2 to .4 range" (p. 114). Clark and Watson (1995) offer: "we recommend that the average interitem correlation fall in the range of .15-.50. . . if one is measuring a broad higher order construct such as extraversion, a mean correlation as low as .15-.20 probably is desirable; by contrast, for a valid measure of a narrower construct such as talkativeness, a much high mean intercorrelation (perhaps in the .40-.50 range) is needed" (p. 316).

References

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the evaluation of personality scales. *Journal of Personality*, *54*, 106-148.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309-319.

Measurement. (2001). *Journal of Consumer Psychology*, *10*(1&2), 55-69.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, *80*(3), 217-222.

Ying, L. (2009). *Relationship between foreign film exposure and ethnocentrism*. Unpublished masters thesis, School of Communication, Cleveland State University. [Cleveland State University campus-wide nominee for 2010 Midwestern Association of Graduate Schools Distinguished Masters Thesis Award]

10/18