Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2015, Vol. 34, No. 4, pp. 4–9

Reliability, Dimensionality, and Internal Consistency as Defined by Cronbach: Distinct Albeit Related Concepts

Ernest C. Davenport, Jr., and Mark L. Davison, *University of Minnesota,* Pey-Yan Liou, *National Central University,* and Quintin U. Love, *Pearson*

This article uses definitions provided by Cronbach in his seminal paper for coefficient α to show the concepts of reliability, dimensionality, and internal consistency are distinct but interrelated. The article begins with a critique of the definition of reliability and then explores mathematical properties of Cronbach's α. Internal consistency and dimensionality are then discussed as defined by Cronbach. Next, functional relationships are given that relate reliability, internal consistency, and dimensionality. The article ends with a demonstration of the utility of these concepts as defined. It is recommended that reliability, internal consistency, and dimensionality each be quantified with separate indices, but that their interrelatedness be recognized. High levels of unidimensionality and internal consistency are not necessary for reliability as measured by α nor, more importantly, for interpretability of test scores.

Keywords: Cronbach's alpha, dimensionality, internal consistency, reliability

There has been much critical comment on Cronbach's (1951) coefficient α (e.g., Cortina, 1993; Green, Lissitz, & Mulaik, 1977; Green & Yang, 2009; Rodriguez & Maeda, 2006; Sijtsma, 2009a, 2009b; Terwilliger & Lele, 1979), including criticism by Cronbach himself (Cronbach & Shavelson, 2004). A major point of contention seems to be the misconception that α serves also as indexes of internal consistency and unidimensionality. Green et al. (1977) believe confusion of these terms has led to misuse of α. They conclude that internal consistency implies interrelatedness, but not necessarily unidimensionality while homogeneity implies unidimensionality.

The goals of this article are (1) to show that reliability, internal consistency, and dimensionality as described by Cronbach (1951) are three distinct but interrelated concepts, and (2) to clarify the relationships among the concepts in ways that minimize misconceptions that can arise in test construction and the reporting of reliability data. We do this by offering mathematical, conceptual, and practical illustrations that distinguish these issues. We base most of our comments on Cronbach's 1951 paper as he was aware of these issues and the subsequent misconceptions were due to misunderstandings of concepts in Cronbach's seminal paper by others.

Ernest C. Davenport, Department of Educational Psychology, University of Minnesota, 56 East River Parkway, Minneapolis, MN 55455; lqr6576@umn.edu. Mark L. Davison, Department of Educational Psychology, University of Minnesota, 56 East River Parkway, Minneapolis, MN 55455. Pey-Yan Liou, Graduate Institute of Learning and Instruction and Center of Teacher Education, National Central University, 300 Zhongda Road, Zhongli District, Taoyuan City 32001, Taiwan. Quintin U. Love, Pearson Education, 19500 Bulverde Road, San Antonio, TX 78259.

Theoretical Bases

We begin with Cronbach's own words concerning reliability, internal consistency, and unidimensionality. Cronbach (1951) defines reliability as the accuracy or dependability of measurements. Homogeneity or internal consistency is the degree to which items measure the same thing. For a test to be interpretable, according to Cronbach, the items need to have a large first principal factor saturation, but the common factor structure accounting for the item covariances need not be unidimensional.

Given an observed score, X, is composed of two uncorrelated entities, true score T and measurement error E, the classical model of an observed score is written as: $X = T + E$. Reliability is then defined as the squared correlation between the test and its true score ρ^2_{TX} or as the ratio of true score variance σ^2_T to total observed score variance σ^2_X:

$$\textit{reliability} = \rho^2_{TX} = \frac{\sigma^2_T}{\sigma^2_X}. \quad (1)$$

α will approach 1 when differences in the observed scores are primarily due to the fact that the examinees differ on their true scores. One question of interest here is whether high levels of alpha require high levels of internal consistency or unidimensionality.

Cronbach's Alpha

Cronbach's (1951) α is the most commonly used index of reliability (Hogan, Benjamin, & Brezinksi, 2000; Thompson, 2003). For an observed score X, where X is an unweighted sum of individual items X_i, Cronbach's α can be written as

© 2015 by the National Council on Measurement in Education

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} S^2_{X_i}}{S^2_X}\right), \quad (2)$$

where m is the number of entities (items) in the composite, X, S^2_X is the variance of X, and $\sum S^2_{x_i}$ is the sum of item variances. If all items are measuring independent content (different dimensions), such that the X_i scores are uncorrelated, $S^2_X = \sum S^2_{X_i}$ and $\alpha = 0$ as one would want in a measure of reliability, a measure of internal consistency, and a measure of unidimensionality that ranges from 0 (no consistency) to 1 (total consistency).

For most of his paper, Cronbach assumed that all items have equal variances of 1.00, as we will do in most places to be consistent with Cronbach and to simplify some proofs. However, this assumption is not often met in practice. We will also assume that all item correlations are nonnegative, not an unreasonable assumption for items measuring the same construct and/or comprising a single test.

If X is the sum of items, then $S^2_X = \sum_{i=1}^{m}\sum_{j=1}^{m} S_{X_iX_j}$ the variance of the composite is the sum of all item variances and covariances, and this sum has m^2 terms. If the items are perfectly correlated and standardized to have variance 1.00, all terms in the sum equal 1.00 so that S^2_X is the sum of m^2 terms all equal to 1.00: $S^2_X = m^2$. The sum of the item variances is the sum of m terms all equal to 1.00: $\sum S^2_{X_i} = m$. Substituting m^2 for S^2_X and m for $\sum S^2_{X_i}$ in the α formula of Equation 2 yields $\alpha = 1$.

That α ranges from 0 for unrelated items to 1 for perfectly related items may give credence to the idea that α is also a measure of internal consistency and unidimensionality. When $\alpha = 0$, the test is unreliable, internally inconsistent, and multidimensional (in m dimensions). When $\alpha = 1$, the test is reliable, internally consistent, and unidimensional. In between 0 and 1, however, α may tell us something about the proportion of variation in the total score attributable to true scores but nothing definite about internal consistency or unidimensionality of the test.

α and Internal Consistency

A careful look at α (still assuming standardized items) shows α's mathematical characteristics. As stated above, the variance of a composite is $S^2_X = \sum_{i=1}^{m}\sum_{j=1}^{m} S_{X_iX_j}$. In our case the variance of the composite, $S^2_X = \sum S^2_{X_i} + m(m-1)\overline{S}_{X_iX_j}$ given that the sum of the $m(m–1)$ covariances equals $m(m–1)$ times the average of the covariances. Thus, the variance of the composite is the sum of the variances of the items plus the number of covariances times the average covariance. Moreover, given that the items are standardized $\sum S^2_{X_i} = m$ and the covariances are just correlations, thus, $S^2_X = m + m(m-1)\overline{r}_{X_iX_j}$. For ease of notation, we let $\overline{r}_{X_iX_j} = \overline{r}$ (the average of all unique pairwise correlations of X_i and X_j where $i \neq j$). From Equation 2, we get

$$\alpha = \frac{m}{m-1}\left(1 - \frac{m}{m + m(m-1)\overline{r}}\right)$$

$$\alpha = \frac{m}{m-1}\left(\frac{m + m(m-1)\overline{r}}{m + m(m-1)\overline{r}} - \frac{m}{m + m(m-1)\overline{r}}\right)$$

$$\alpha = \frac{m\overline{r}}{1 + (m-1)\overline{r}}. \quad (3)$$

Equation 3 explicitly shows that α is a function of both the number of items and the average interitem correlation, Cronbach's measure of internal consistency. Given Cronbach's preference for $\overline{r}$ as a measure of internal consistency, α is explicitly excluded from also being a measure of internal consistency because it is a function of internal consistency ($\overline{r}$) and something else (number of items). Moreover, one can see that reliability as indexed by α and internal consistency as indexed by $\overline{r}$ are functionally related. For α to be high, internal consistency need only meet the minimal condition that $\overline{r} > 0$. For if $\overline{r} > 0$, α can reach any desired level with enough items. The number of items needed to reach a desired level of alpha increases as $\overline{r}$ decreases. Moreover, internal consistency $\overline{r}$ is independent of test length while α is not.

One implication of Equation 3 is that for any given average correlation, alpha becomes solely a function of test length. For example, if $\overline{r} = .5$, Equation 3 simplifies to

$$\alpha = \left(\frac{m}{1+m}\right). \quad (4)$$

Thus, one can get any reliability desired by solving for the number of items that would lead to the desired reliability value. For example, given an average interitem correlation of .5, one would need 19 items to achieve a reliability of .95. Note that alpha is solely a function of number of items conditional on any average correlation. The only difference is that as the average correlation changes the function would also change.

In closing this section, it should be noted that when item variances are unequal, α is still a function of internal consistency and test length, but not Cronbach's measure of internal consistency. Then we have

$$\alpha = \frac{m\frac{\bar{c}}{\bar{s}^2}}{1 + (m-1)\frac{\bar{c}}{\bar{s}^2}}, \quad (5)$$

where $\bar{c}$ is the average covariance among item pairs ($i \neq i'$) and $\bar{s}^2$ is the average item variance. The ratio $\bar{c}/\bar{s}^2$ is the proportion of average item variance that is shared variance. It is a measure of internal consistency more appropriate when item variances are unequal, and it equals $\bar{r}$ when all item variances are equal.

Alpha and Dimensionality

Cronbach (1951) distinguished internal consistency from reliability. He also addressed dimensionality. Cronbach says "for a test to be interpretable, however, it is not essential that all items be factorially similar. What is required is that a large proportion of the test variance be attributable to the first principal factor running through the test" (p. 320). Thus, the first eigenvalue of the interitem correlation matrix, expressed as a proportion of total item variance, is an indicator of dimensionality. Cronbach stated that this value could be high even with a multidimensional test and that "Items with quite low inter-correlations can yield an interpretable scale"

© 2015 by the National Council on Measurement in Education

(Cronbach, 1951, p. 332). Hattie (1985) concurs with Cronbach's definition of dimensionality. He says there are over 30 indices of unidimensionality with the proportion of variance accounted for by the first principal factor being one. Cronbach's definition of dimensionality differs from that suggested by a factor analytic approach where dimensionality is the number of common factors needed to reproduce the item correlations or covariances (McDonald, 1999). Cronbach's idea of a major direction through the data is consistent with a general factor that can represent a higher order construct. Zopluoglu (2013) found that the first principal component was large regardless of the number of common factors, if items were complex and/or the primary factors were correlated. In both instances, the items have something in common across multiple factorial dimensions, and the major direction through the data is related to that higher order construct.

We use the proportion of variance accounted for by the first principal component as our index of unidimensionality for several reasons. Cronbach uses it. It has a history of use (Carmines & Zeller, 1979; Hattie, 1985; Reckase, 1979). It is consistent with multidimensional constructs (e.g., classroom tests, fluid intelligence, personality constructs) where the construct is a higher order factor. Finally, as discussed below, it is a definition of unidimensionality closely related to internal consistency and alpha.

Friedman and Weisberg (1981) showed a result that is key to understanding the relationship between alpha, internal consistency, and unidimensionality as we have defined them. They showed that when all interitem correlations are positive, as should be true for items comprising a single test, the first principal component eigenvalue is an approximate function of the average correlation of the items:

$$\lambda_1 \approx 1 + (m-1)\,\bar{r} \tag{6}$$

with equality holding when items have equal variances and all interitem correlations are equal. They also state that Equation 6 is a lower bound for the first eigenvalue that never overestimates λ_1, and the first eigenvalue "hugs the lower bound closely" (Friedman & Weisberg, 1981, p. 15). The discrepancy in the estimate "deteriorates slightly as the variance of the correlations increases" (Friedman & Weisberg, 1981, p. 14). Thus we can rewrite the formula for Cronbach's alpha (Equation 3) as

$$\alpha \approx \frac{m\bar{r}}{\lambda_1} = \frac{\bar{r}}{V}, \tag{7}$$

$$V\,\alpha \approx \bar{r}. \tag{8}$$

Equation 8 explicitly shows the approximate relationship between unidimensionality as measured by the proportion of variance accounted for by the first principal component *V,* reliability as defined by α, and internal consistency defined as $\bar{r}$. Another implication of Equation 8 is that, as the number of items increases, α goes to 1 and V approaches $\bar{r}$ (undimensionality equals internal consistency).

Examples

Here we use examples to illustrate the relationships and distinctions between reliability, internal consistency, and dimensionality. Structure 1 shows that one can have internal consistency when measuring more than one dimension. Structures 1 and 2 show that, if dimensionality is defined as the number of factors needed to reproduce item covariances, items need not be unidimensional to have at least moderately high levels of internal consistency and alpha. Example 3 shows that a single underlying dimension does not assure high levels of internal consistency or reliability.

Table 1 shows three factor structures (The corresponding correlational matrices are given in the appendix). The first is complex as items load on three uncorrelated factors. While such a structure is not commonly posited in education, it would be appropriate for items whose content crosses disciplines. This structure could represent a STEM (science, technology, engineering, and mathematics) ability test with an expository reading section (Dimension 1) all of whose passages cover math or science content, and hence the items also load on a math (Dimension 2) or science (Dimension 3) dimension. Also, some science items require math, and some math items apply to problems in science. Rather than accounting for cross domain item correlations by allowing dimensions to be correlated, as is the usual practice, this structure posits that items load on more than one factor. All items are related due to its own content dimension or one of the other dimensions. Given this structure, these items should have substantial internal consistency (average intercorrelation), while not being unidimensional (not measuring the same thing). The complexities allow for the items to have something in common, even if what they have in common is different for different item pairs (Thompson's, [1916] overlapping factors).

The second structure has three well-defined, simple structure factors (no item complexities); simulating a test with three distinct content strands. Here we have posited that the dimensions are uncorrelated, not because that is representative of real tests, but to show that even uncorrelated, simple structure dimensions can give rise to moderately internally consistent items and a moderately high level of alpha. This structure is clearly multidimensional, with four items being internally consistent within each of the three factors while having no consistency for items between factors. The third structure consists of one very weak factor. This structure illustrates how items can have very low internal consistency and give rise to a very low reliability as measured by alpha even though they are unidimensional in that only one common factor accounts for the item correlations. This structure does not correspond to a good test, but it has a clean unidimensional structure of the type often sought in test construction.

Table 2 contains our measures of internal consistency $\bar{r}$, reliability α, and unidimensionality *V,* for each of the three factor structures. Note that these values were obtained analytically from the factor structures utilizing equations given above (simulations verified the accuracy of these values). The structures for the examples vary in the degree to which the resulting tests are reliable, unidimensional, and/or internally consistent. The first structure is multidimensional, with overlapping factors. It has a strong dimension going through the test leading to a high first principal component despite the multiple common factors. The items load high on the factors as indicated by their low uniquenesses and high commonalities (.72), thus error variance is low indicating high reliability. The first structure also allows for a host of positive correlations. Thus, suggesting a fair amount of internal consistency even though the relationships of the items come from measuring three different things.

Structure 2 has no correlation between factors and no overlap of items from one factor to the next. Thus, it will

Table 1. Factor Structures

	Three Factors Complex				Three Factors Strong / Simple				One Factor Weak	
Items	Factor1	Factor2	Factor3	H^2	Factor1	Factor2	Factor3	H^2	Factor1	H^2
Item 01	.6	.6	.0	.72	.95	.00	.00	.90	.1	.01
Item 02	.0	.6	.6	.72	.95	.00	.00	.90	.1	.01
Item 03	.6	.0	.6	.72	.95	.00	.00	.90	.1	.01
Item 04	.6	.6	.0	.72	.95	.00	.00	.90	.1	.01
Item 05	.0	.6	.6	.72	.00	.95	.00	.90	.1	.01
Item 06	.6	.0	.6	.72	.00	.95	.00	.90	.1	.01
Item 07	.6	.6	.0	.72	.00	.95	.00	.90	.1	.01
Item 08	.0	.6	.6	.72	.00	.95	.00	.90	.1	.01
Item 09	.6	.0	.6	.72	.00	.00	.95	.90	.1	.01
Item 10	.6	.6	.0	.72	.00	.00	.95	.90	.1	.01
Item 11	.0	.6	.6	.72	.00	.00	.95	.90	.1	.01
Item 12	.6	.0	.6	.72	.00	.00	.95	.90	.1	.01

Table 2. Statistics for the Factor Structures

	Three Factor Complex	Three Factor Strong Simple	One Factor Weak
Average item intercorrelation	.46	.25	.01
Coefficient Alpha	.91	.80	.11
Proportion of Variance for the first Eigenvalue	.50	.31	.09

have a smaller first principal component and be assessed to be less unidimensional. The fact that the structure is well defined where all items have high commonalities and low uniquenesses should lead to consistent results over repeated samples implying high reliability. A main feature of Structure 2 is that items on a factor are highly intercorrelated (internally consistent), whereas there is no consistency between items loading on different factors. Thus, internal consistency should not be high here. Note, too, that the simple structure nature of the data will allow for many zero intercorrelations driving down the average interitem correlation $\bar{r}$.

The final structure is unidimensional with low loadings for each item. Due to the weakness of this structure, it should have a weak first principal component. Error for each item will suggest that the test is not reliable leading to a low value for α. The resulting interitem correlations will be small suggesting little internal consistency.

In Table 2, the proportion of variance accounted for by the first principal component eigenvalue for the Complex structure is .50 suggesting by most thresholds that it is unidimensional (Hattie, 1985). This is consistent with Structure 1 having complex items (overlapping factors) indicating the existence of a strong first principal component. The corresponding value for the Strong Simple structure is .31, also appearing unidimensional by the lower threshold of 20% suggested by Reckase (1979) to indicate a unidimensional test as cited in Hattie (1985). It is ironic that our index of unidimensionality is the lowest for the only unidimensional structure (Structure 3), .09. However, it may also be misleading to characterize that structure as unidimensional, as it has only one common factor but possibly 12, strong specific factors and/or error that, collectively, account for far more variance than the common factor. To be considered as having a strong dominant factor for measurement purposes, a test needs a factor that dominates other common factors, specific factors, and error.

The average correlation for the first structure is .46, indicating a fairly decent degree of average relationship for the items. The corresponding value is .25 for the Strong Simple Structure and only .01 for the Weak Structure. Our use of the average item correlation for internal consistency is monotonically related to internal consistency as modeled in our examples by strongly related overlapping items, moderately related nonoverlapping items, and weakly related unidimensional items; however, there are no current rules of thumb for how large the average interitem correlations needs to be for a test to be deemed internally consistent. What do values of.46,.25, and/or.01 actually mean for this index? It may be worthwhile to examine the magnitude of $\bar{r}$ relative to qualitative statements of the internal consistency of a test.

Coefficient α as the reliability estimate suggests the first two structures are fairly reliable, with α values of .91 and .80, respectively, but not the third structure, $\alpha = .11$. Note, the first two structures have multiple factors and the second is noninternally consistent as well. Even so, most estimates of reliability for these two structures would be high given that the uniquenesses of both structures are low indicating a low level of error variance which is consistent with the classical definition of reliability as a ratio of the true score variance to total score variance. The final structure would lead to unreliable tests in that 99% of the variance for each item comes from unspecified specific factors and/or random error, making the true score variance small relative to the total variance.

At first glance, our results seemingly call into question the proportion of variance accounted for by the first principal component as a valid measure of unidimensionality. However, we are not concerned with the number of primary factors needed to account for item covariances. Rather *V* measures the extent to which there is a prevailing direction through the data which we and Cronbach (Cronbach, 1951, p. 320) have defined as the only dimensionality requirement to make a composite score meaningful.

Alpha and Multidimensional Tests

With sufficient test length, α can be high for tests composed of multidimensional items, but is α an appropriate measure of reliability for such a test? Novick and Lewis (1967)

© 2015 by the National Council on Measurement in Education

showed that, given uncorrelated errors, α is less than or equal to reliability with equality holding when items have equal true scores. When items are multidimensional, they might not be true score equivalent, and α, computed with items as the unit of analysis, might be an underestimate of reliability.

Cronbach (1951) also discussed the computation of α by dividing items into subtests, here called item parcels, such that the total test score is the sum of the item parcel scores. With parcels, α can be computed using Equation 2 substituting the number of parcels P for the number of items, and substituting the parcel score variances for the item variances. Davison and Davenport (2015) describe conditions under which parcels, but not items, will be true score–equivalent, in which case α computed from parcels, but not items, will equal the test reliability.[1] In practice, approximately true score equivalent parcels might be assembled by stratifying items based on content and format, and then assigning items to parcels so that each parcel contains approximately the same mix. For instance, a test of receptive language might contain 33 items including reading and listening items in multiple choice and constructed response formats. The items fall into four strata: multiple-choice reading (12 items), multiple-choice listening (15 items), constructed- response reading (3 items), and constructed-response listening (3 items). The items could be divided into three parcels with 11 items each by assigning an equal number from each stratum to each parcel. For instance, each parcel might contain four multiple-choice reading, five multiple-choice listening, one constructed response reading, and one constructed response listening. Alpha could then be computed from the parcel scores, rather than item scores, to obtain a more precise estimate of the test's reliability.

Using the examples given in Table 1, one may speculate on the reliability for parcels relative to Factor Structure 2, if the parcels were defined as follows: Parcel 1: Items 1, 5, and 9, Parcel 2: Items 2, 6, and 10, Parcel 3: Items 3, 7, and 11, and Parcel 4: Items 4, 8, and 12. We can see that each parcel is made up of one item from each of the dimensions and that the parcels are essentially equivalent; for each factor: the sum of item loadings is the same for each parcel on each factor, .95. Note that the composite score using the items or parcels is the same. Alpha estimated for items will be a lower bound for reliability, since items are not true score–equivalent, but alpha estimated from parcels will be an estimate of reliability, since the parcels are true score–equivalent.

Discussion

The above proofs and examples do not invalidate α as an index of reliability. However, alpha is not a measure of internal consistency, because α also depends on test length (Equation 3). α is also not a measure of unidimensionality. The relationship between α and first component saturation is complex, but α is an approximate function of both test length and first component saturation (Equation 7). α is a function of both internal consistency and first component saturation, but it is a pure measure of neither.

In the limit, as the number of items approaches infinity, the first component saturation will approach internal consistency. Internal consistency is a fundamental property of the items that does not depend on the number of items. Internal consistency affects α in two ways. First, internal consistency must be positive; otherwise, even a long test will not be reliable. Second, if positive, the level of internal consistency determines the number of items needed for any desired level of α.

The relationships among the three concepts have implications for psychometric research and test development. First, in research reports, internal consistency describes items, not the test, because internal consistency does not reflect a very fundamental property of the test—its length. Second, α should not be described as a measure of internal consistency or unidimensionality, nor should high levels of α be interpreted as a sign that the items are internally consistent or unidimensional. Given enough items, α can be high even if the items are multidimensional and have low internal consistency. High levels of α may indicate little more than that the test is long. Alpha can be described as an internal consistency measure of reliability, as it is a function of internal consistency. α can also be interpreted as an approximate estimate of alternate forms reliability or equivalence (Cronbach, 1951).

Structural equation modeling and/or factor analysis of items are useful for clarifying the nature of a construct measured by a test. Furthermore, structural equation modeling can be used to estimate the reliability of a test when items are not true score–equivalent (Kamata, Turhan, & Darandari, 2003; McDonald, 1999; Raykov, 1997, 2001). If more than one dimension is required to account for item covariances, however, the analysis does not invalidate the test. Rather, it suggests that the test measures a higher order dimension, not a first order dimension. To borrow vocabulary from chemistry, the construct measured by the test is a compound, not an element. Tests composed of multidimensional items may have lower internal consistency, and therefore may need to be longer in order to achieve a desired level of reliability as measured by α.

A test needs to measure a well defined construct, but the construct need not be at the first order. Many useful test scores in education and psychology are at a higher order. The composite score on the ACT Test reflects student performance in English, mathematics, reading, and science. The general intelligence score of the Wechsler Intelligence Scale for Children reflects performance in verbal comprehension, working memory, perceptual reasoning, and processing speed. In the NEO-PI personality measure, each domain (for example, extraversion) is composed of multiple facets (warmth, gregariousness, assertiveness, excitement seeking, and positive emotions). Alpha, computed using multidimensional items as the units of analysis, will likely underestimate score reliability, because items are not true score–equivalent. Alpha computed using true score–equivalent parcels as the unit of analysis or using coefficients based on structural equation modeling can provide a more accurate estimate of score reliability.

An estimate of item internal consistency can be useful in planning test length. If the test is a revision of an existing test, the average correlation of items (or $\bar{c}/\bar{s}^2$if the items have unequal variances) for the existing test can provide an estimate of internal consistency. If there is a pilot version of a new test, the internal consistency of items in the pilot version may provide an index of internal consistency. Once an estimate of internal consistence is available, Equation 3 can be solved to estimate the number of items needed to reach any desired level of α.

In conclusion, reliability is a necessary but insufficient attribute for a test. Reliability, internal consistency, and unidimensionality may appear to be desirable attributes for a test. However, overemphasis on the latter two features may be misplaced (Cattell & Tsujioka, 1964; Hattie, 1985). Internal consistency describes items, not the test, because internal consistency does not reflect the test's length. Internal consistency reliability only requires positively correlated items, although low item correlations will necessitate more items. Even multidimensional items can have moderately high levels of internal consistency, and tests composed of multidimensional items can be both reliable and interpretable, given enough items and/or the test assesses a higher order construct.

Note

[1]Davison and Davenport (2015) define true score variance as the proportion of variance attributable to common factors k. For two parcels (p, p') with items $i(p)$ and $i(p')$, respectively, and item factor loadings $\lambda_{k,i(p)}$ and $\lambda_{k,i(p')}$,respectively, the parcels will be true score–equivalent if $\sum_{i(p)} \lambda_{k,i(p)} = \sum_{i(p')} \lambda_{k,i(p')}$for all common factors k. For each factor, the sum of item loadings must be the same for both parcels.

Appendix

Expected Correlation Matrices

Complex

1.000	.360	.360	.720	.360	.360	.720	.360	.360	.720	.360	.360
.360	1.000	.360	.360	.720	.360	.360	.720	.360	.360	.720	.360
.360	.360	1.000	.360	.360	.720	.360	.360	.720	.360	.360	.720
.720	.360	.360	1.000	.360	.360	.720	.360	.360	.720	.360	.360
.360	.720	.360	.360	1.000	.360	.360	.720	.360	.360	.720	.360
.360	.360	.720	.360	.360	1.000	.360	.360	.720	.360	.360	.720
.720	.360	.360	.720	.360	.360	1.000	.360	.360	.720	.360	.360
.360	.720	.360	.360	.720	.360	.360	1.000	.360	.360	.720	.360
.360	.360	.720	.360	.360	.720	.360	.360	1.000	.360	.360	.720
.720	.360	.360	.720	.360	.360	.720	.360	.360	1.000	.360	.360
.360	.720	.360	.360	.720	.360	.360	.720	.360	.360	1.000	.360
.360	.360	.720	.360	.360	.720	.360	.360	.720	.360	.360	1.000

Strong–Simple

1.000	.903	.903	.903	.000	.000	.000	.000	.000	.000	.000	.000
.903	1.000	.903	.903	.000	.000	.000	.000	.000	.000	.000	.000
.903	.903	1.000	.903	.000	.000	.000	.000	.000	.000	.000	.000
.903	.903	.903	1.000	.000	.000	.000	.000	.000	.000	.000	.000
.000	.000	.000	.000	1.000	.903	.903	.903	.000	.000	.000	.000
.000	.000	.000	.000	.903	1.000	.903	.903	.000	.000	.000	.000
.000	.000	.000	.000	.903	.903	1.000	.903	.000	.000	.000	.000
.000	.000	.000	.000	.903	.903	.903	1.000	.000	.000	.000	.000
.000	.000	.000	.000	.000	.000	.000	.000	1.000	.903	.903	.903
.000	.000	.000	.000	.000	.000	.000	.000	.903	1.000	.903	.903
.000	.000	.000	.000	.000	.000	.000	.000	.903	.903	1.000	.903
.000	.000	.000	.000	.000	.000	.000	.000	.903	.903	.903	1.000

Weak

1.000	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010
.010	1.000	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010
.010	.010	1.000	.010	.010	.010	.010	.010	.010	.010	.010	.010
.010	.010	.010	1.000	.010	.010	.010	.010	.010	.010	.010	.010
.010	.010	.010	.010	1.000	.010	.010	.010	.010	.010	.010	.010
.010	.010	.010	.010	.010	1.000	.010	.010	.010	.010	.010	.010
.010	.010	.010	.010	.010	.010	1.000	.010	.010	.010	.010	.010
.010	.010	.010	.010	.010	.010	.010	1.000	.010	.010	.010	.010
.010	.010	.010	.010	.010	.010	.010	.010	1.000	.010	.010	.010
.010	.010	.010	.010	.010	.010	.010	.010	.010	1.000	.010	.010
.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	1.000	.010
.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	.010	1.000

References

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage.

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*, 3–30.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418.

Davison, M. L. & Davenport, E. C., Jr. (2015, April). *Coefficient alpha and dimensionality*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago IL.

Friedman, S., & Weisberg, H. F. (1981). Interpreting the first eigenvalue of a correlation matrix. *Educational and Psychological Measurement*, *41*, 11–21.

Green, S. B., Lissitz, R. W., & Mulaik, S. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827–838.

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121–135.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164.

Hogan, T. P., Benjamin, A., & Brezinksi, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523–532.

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidinesional composite scale scores*. Paper presented at the meeting of the American Eduational Research Association, Chicago, IL, July 16, 2005. Available at http://www.coe.fus.edu/AERA/Kamata2.pdf

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

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measures. *Psychometrika*, *32*, 1–13.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173–184.

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor test: Results and implications. *Journal of Educational Statistics*, *4*, 207–230.

Rodriguez, M. C. and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322.

Sijtsma, K. (2009a). On the use, misuse, and very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, *74*, 169–173.

Terwilliger, J. S., & Lele, K. (1979). Some relationships among internal consistency, reproducibility, and homogeneity. *Journal of Educational Measurement*, *16*, 101–108.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Thompson, G. H. (1916). A hierarchy without a general factor. *Journal of Psychology*, *8*, 271–281.

Zopluoglu, C. (2013). *Assessing dimensionality of latent structures underlying dichotomous item response data with imperfect models* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

© 2015 by the National Council on Measurement in Education