JOURNAL OF PERSONALITY ASSESSMENT, *80*(1), 99–103

STATISTICAL DEVELOPMENTS AND APPLICATIONS

Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency

David L. Streiner
Baycrest Centre for Geriatric Care
Department of Psychiatry
University of Toronto

Cronbach's α is the most widely used index of the reliability of a scale. However, its use and interpretation can be subject to a number of errors. This article discusses the historical development of α from other indexes of internal consistency (split-half reliability and Kuder–Richardson 20) and discusses four myths associated with α: (a) that it is a fixed property of the scale, (b) that it measures only the internal consistency of the scale, (c) that higher values are always preferred over lower ones, and (d) that it is restricted to the range of 0 to 1. It provides some recommendations for acceptable values of α in different situations.

Perhaps the most widely used measure of the reliability of a scale is Cronbach's α (1951). One reason for this is obvious; it is the only reliability index that does not require two administrations of the scale, or two or more raters, and so can be determined with much less effort than test–retest or interrater reliability. Unfortunately, the ubiquity of its use is matched only by the degree of misunderstanding regarding what α does and does not measure. This article is intended to be a basic primer about α. It will approach these issues from a conceptual and a statistical perspective and illustrate both the strengths and weaknesses of the index.

I begin by discussing what is meant by reliability in general and how α and other indexes of "internal consistency" determine this. In classical test theory, a person's total score (i.e., the score a person receives on a test or scale, which is sometimes referred to as the observed score) is composed of two parts: the true score plus some error associated with the measurement. That is:

$$Score_{Total} = Score_{True} + Score_{Error}, \quad (1)$$

It is assumed that the error is random with a mean of zero, so that it sometimes acts to increase the total score and sometimes decrease it, but does not bias it in a systematic way. Because every scale has some degree of measurement error, we can never determine the true score; it is the average of all scores a person would receive if he or she took the test an infinite number of times (Allen & Yen, 1979). A consequence of this is that a person's total score will vary around the true score to some degree. One way of thinking about reliability, then, is that it is the ratio of the variance of the true scores (σ^2_{True}) to the total scores (σ^2_{Total}):

$$Reliability = \frac{\sigma^2_{True}}{\sigma^2_{Total}}. \quad (2)$$

However, at any given time, a person's true score will be the same from one testing to another, so that an *individual's* σ^2_{True} will always be zero. Thus, Equation 2 pertains only to a *group* of people who differ with respect to the characteristic being measured.

Before continuing with issues of measuring reliability, however, it would be worthwhile to digress for a moment and expand on what is meant by the "true" score. In many respects, it's a poor choice of words and a potentially misleading term (although one we're stuck with), because "true," in psychometric theory, does not mean either "accurate" or "immutable." It simply means a score that is *consistent* at a given level of the underlying trait; that is, a score that's free from *random* error. However, it may not necessarily be free from *systematic* error, or bias, and will (I hope) change if the underlying trait does. Three examples may help to illustrate these points.

A person is administered the Wechsler Adult Intelligence Scale–III (WAIS–III) repeatedly and gets an average score of

80. However, if the test is given in English, which the person learned only 2 years ago, the "true" score of 80 is likely not an accurate reflection of her intelligence. Similarly, a person undergoing an assessment for purposes of child access and custody may deliberately understate the degree to which he uses corporal punishment. Repeat evaluations may yield similar scores on the test and the mean will be a good approximation of the true score (because of low random error), but the defensive response style, which produces a bias, means that the true score will not be an accurate one. Finally, a depressed person may have a *T* score around 75 on numerous administrations of a personality test. However, if she responds well to therapy, then both her depression and her true score should move closer to the average range.

The different effects of random and systematic error are captured in Judd, Smith, and Kidder's (1991) expansion of Equation 1:

$$Score_{Total} = Score_{CI} + Score_{SE} + Score_{RE}, \quad (3)$$

where CI is the construct of interest, SE the systematic error, and RE is the random error. In this formulation, $Score_{CI}$ + $Score_{SE}$ is the same as $Score_{True}$ in Equation 1. Two advantages of expressing the true score as the sum of the construct and the systematic error is that it illustrates the relationship between reliability and validity, and shows how the different types of error affect each of them:

$$Reliability = \frac{\sigma^2_{CI} + \sigma^2_{SE}}{\sigma^2_O}, \quad (4)$$

whereas

$$Validity = \frac{\sigma^2_{CI}}{\sigma^2_O}. \quad (5)$$

These last two equations show that random error affects both reliability and validity (because the larger it is, the smaller the ratio between the numerators and denominators), whereas systematic error affects only validity.

Returning to reliability, it can be defined on a conceptual level as the degree to which "measurements of individuals on different occasions, or by different observers, or by similar or parallel tests, produce the same or similar results" (Streiner & Norman, 1995, p. 6). That is, if a person's score on Scale D of the Minnesota Multiphasic Personality Inventory–2 is 73 at Time 1, then we would want his or her score to be very close to 73 one or 2 weeks later, assuming that the person has not changed during that interval. Similarly, if one rater gives a patient a score of 17 on the Hamilton Depression Scale, then an independent rater should also assign a score very near to 17. In addition to these two sources of error (time and observer), we can add a third source—that associated with the homogeneity of the items that comprise the scale.[1]

If a scale taps a single construct or domain, such as anxiety or mathematical ability, then to ensure content validity, we want the scale to (a) consist of items that sample the entire domain and (b) not include items that tap other abilities or constructs. For example, a test of mathematics should sample everything a child is expected to know at a given grade level, but not consist of long, written passages that may reflect the child's reading ability as much as his or her math skills. Similarly, an anxiety inventory should tap all of the components of anxiety (e.g., cognitive, behavioral, affective) but not include items from other realms, such as ego strength or social desirability. Because classical test theory assumes that the items on a scale are a random sample from the universe of all possible items drawn from the domain, then they should be correlated highly with one another. However, this may not always be true. For example, Person A may endorse two items on an anxiety inventory (e.g., "I feel tense most of the time"; "I am afraid to leave the house on my own"), whereas Person B may say True to the first but No to the second. This difference in the pattern of responding would affect the correlations among the items, and hence the *internal consistency* of the scale. A high degree of internal consistency is desirable, because it "speaks directly to the ability of the clinician or the researcher to interpret the composite score as a reflection of the test's items" (Henson, 2001, p. 178).

The original method of measuring internal consistency is called "split half" reliability. As the name implies, it is calculated by splitting the test in half (e.g., all of the odd numbered items in one half and the even numbered ones in the other) and correlating the two parts. If the scale as a whole is internally consistent, then any two randomly derived halves should contain similar items and thus yield comparable scores. A modification of this was proposed by Rulon (1939), which relies on calculating the variance of the difference score between the two half-tests (σ^2_d) and the variance of the total score (σ^2_{Total}) across people:

$$Reliability = 1 - \frac{\sigma^2_d}{\sigma^2_{Total}}. \quad (6)$$

The right most part of the equation (σ^2_d / σ^2_{Total}) is the proportion of error variance in the scores, which can be thought of as what the items do *not* have in common.

There are two problems with these indexes, however. The first is based on the fact that the reliability of a scale is pro-

[1]These—plus parallel forms—are the traditional sources of unreliability. Generalizability theory, the topic of a subsequent article, allows us to assess any other factors that may affect reliability, such as time of day or the interaction between rater and form.

portional to its length. Splitting a scale in half reduces its length by 50%, and hence underestimates the reliability. This difficulty can be solved relatively easily, though, by using the Spearman–Brown "prophecy" formula that compensates for the reduction in length. The second issue is that there are many ways to split a scale in half; in fact, a 12-item scale can be divided 462 ways and each one will result in a somewhat different estimate of the reliability.[2] This problem was dealt with for the case of dichotomous items by Kuder and Richardson (1937). Their famous equation, which is referred to as KR–20 because it was the 20th one in their article, reads:

$$KR-20 = \frac{k}{k-1}\left[1-\frac{\Sigma p_k q_k}{\sigma^2_{Total}}\right], \quad (7)$$

where k is the number of items, p_k the proportion of people who answered positively to item k, q_k is the proportion of people who answered negatively (i.e., $q_k = 1 - p_k$), and σ^2_{Total} is the variance of the total scores. KR–20 can be thought of as the mean of all possible split-half reliabilities.

The limitation of handling only dichotomous items was solved by Cronbach (1951), in his generalization of KR–20 into coefficient α, which can be written as:

$$\alpha = \frac{k}{k-1}\left[1-\frac{\Sigma\sigma^2_k}{\sigma^2_{Total}}\right], \quad (8)$$

where $\Sigma\sigma^2_k$ is the sum of the variances of all of the items. Coefficient α has the same property as KR–20, in terms of being the average of all possible splits.[3]

That pretty much describes what α is and can do. In the next section, I look at the other side of the equation and discuss what α is not and cannot, or does not, do.

MYTHS ABOUT ALPHA

Myth 1: Alpha Is a Fixed Property of a Scale

The primary myth surrounding α (and all other indexes of reliability, for that matter) is that once it is determined in one study, then you know the reliability of the scale under all circumstances. As a number of authors have pointed out, however, reliability is a characteristic of the test *scores*, not of the test itself (e.g., Caruso, 2000; Pedhazur & Schmelkin, 1991; Yin & Fan, 2000). That is, reliability depends as much on the sample being tested as on the test. This has been reinforced in the recent guidelines for publishing the results of studies (Wilkinson & The Task Force on Statistical Inference, 1999), which stated that, "It is important to remember that a test is not reliable or unreliable. Reliability is a property of the scores on a test *for a particular population of examinees*" (p. 596; emphasis added). The reasons for this flow out of Equations 1, 2, and 8. Equation 2 tells us that reliability is the ratio of the true and total score variances. However, Equation 1 shows that you can never obtain the true score. Consequently, any measured value of the reliability is an estimate and, as with all estimates of parameters, subject to some degree of error. Finally, Equation 8 reflects the fact that the reliability depends on the total score variance, and this is going to differ from one sample of people to another. The more heterogeneous the sample, then the larger the variance of the total scores and the higher the reliability. Caruso (2000) did a meta-analysis of reliability studies done with the NEO and found, for example, that the mean reliability of the Agreeableness subscale was .79 when it was used in studies with the general population, but only .62 in clinical samples. Similarly, Henson, Kogan, and Vacha-Haase's (2001) meta-analysis of teacher efficacy scales found that the reliability estimates for the Internal Failure scale ranged from .51 to .82, and from .55 to .82 for the General Teaching Efficacy scale. The reliabilities were affected by a number of attributes of the samples including, not surprisingly, the heterogeneity of the teachers. Consequently, a scale that may have excellent reliability with one group may have only marginal reliability in another. One implication of this is that it is not sufficient to rely on published reports of reliability if the scale is to be used with another group of people; it may be necessary to determine it anew if the group is sufficiently different, especially with regard to its homogeneity.

Myth 2: Alpha Measures Only the Internal Consistency of the Scale

It is true that the higher the correlations among the items of a scale, the higher will be the value of α. But, the converse of this—that a high value of α implies a high degree of internal consistency—is not always true. The reason is that α is also strongly affected by the length of the scale. For example, Cortina (1993) demonstrated that a six-item scale with an average item correlation of .30 has a value of α of .72. Keeping the average correlation the same, but increasing the number of items to 12 and 18 increased α to .84 and .88, respectively. This was not surprising, and has been known for many years for unidimensional scales (e.g., Lord & Novick, 1968). But, Cortina then showed that when a scale with two uncorrelated dimensions was analyzed, keeping the item correlations the same within each "subscale," α was .45 with six items (i.e.,

[2] The number of possible splits is one half of the combination of k items taken $k - 2$ at a time.

[3] As Cortina (1993) pointed out, this is strictly true only if all of the items have the same standard deviation; to the degree that they differ, α will be smaller than the average split-half reliability. It should be mentioned in passing that some computer programs calculate both α and "standardized α," in which all of the items have been converted to have a mean of 0 and a standard deviation of 1. Standardized α is higher than α, but should not be used unless it is intended that all of the items will be standardized in actual use of the scale.

three from each subscale), .65 with 12 items, and .75 with 18 items. A scale composed of three orthogonal (i.e., uncorrelated) subscales had an α of .64 with 18 items. He concluded that

> if a scale has more than 14 items, then it will have an α of .70 or better even if it consists of two orthogonal dimensions with modest (i.e., .30) item intercorrelations. If the dimensions are correlated with each other, as they usually are, then α is even greater. (p. 102)

In other words, even though a scale may consist of two or more independent constructs, α could be substantial as long as the scale contains enough items. The bottom line is that a high value of α is a prerequisite for internal consistency, but does not guarantee it; long, multidimensional scales will also have high values of α.

Myth 3: Bigger Is Always Better

For most indexes of reliability, the higher the value the better. We would like high levels of agreement between independent raters and good stability of scores over time in the absence of change. This is true, too, about α, but only up to a point. As I just noted, α measures not only the homogeneity of the items, but also the homogeneity of what is being assessed. In many cases, even seemingly unidimensional constructs can be conceptualized having a number of different aspects. Lang (1971), for example, stated that anxiety can be broken down into three components—cognitive, physiological, and behavioral—whereas Koksal and Power (1990) added a fourth, affective, dimension. Moreover, these do not always respond in concert and the correlations among them may be quite modest (Antony, 2001). Consequently, any scale that is designed to measure anxiety as a whole must by necessity have some degree of heterogeneity among the items. If the anxiety scale has three or four subscales, they should each be more homogeneous than the scale as a whole, but even here, α should not be too high (over .90 or so). Higher values may reflect unnecessary duplication of content across items and point more to redundancy than to homogeneity; or, as McClelland (1980) put it, "asking the same question many different ways" (p. 30). In the final section, I will expand on this a bit more.

Myth 4: Alpha Ranges Between 0 and 1

Because reliability is a ratio of two variances, it would seem at first glance that it should always be a number between 0 and 1. However, there are times when α is negative. This happens mainly when some of the items are negatively correlated with others in the scale. The primary cause of this, fortunately, is an oversight by the test developers, and easily remedied. Many texts on scale construction recommend that the scoring for roughly half of the items be reversed (e.g., having to endorse *strongly agree* to indicate the presence of a trait for some items, and *strongly disagree* for other items) to minimize Yea-saying bias (e.g., Streiner & Norman, 1995). Needless to say, the scoring for the reversed items should also be reversed. If this isn't done, the items will be negatively correlated, leading to a value of α that is below zero. Of course, if the items *are* scored correctly and some correlations are still negative, then it points to serious problems in the original construction of the scale.

A less frequent cause of a negative value of α is when the variability of the individual items exceeds their shared variance, which may occur when the items are tapping a variety of different constructs (Henson, 2001). Because negative values of α are theoretically impossible, Henson recommended reporting them as zero, but negative or zero, the conclusions are the same—the items are most likely not measuring what they purport to.

USING ALPHA

Not all indexes of reliability can be used in all situations. For example, it is impossible to assess interrater reliability for self-administered scales and difficult to determine test–retest reliability for conditions that change over brief periods of time (which is not to say that some of our students haven't tried). Similarly, there are certain types of scales for which α is inappropriate. It should not be used for "power" tests that measure how many items are completed in a fixed period of time (such as the Digit Symbol Coding subtest of the WAIS–III). The issue here is that it is assumed that people will differ only in terms of the number of items completed, and that everyone will be correct on most or all of the completed ones. So, for any given person, the correlations between items will depend on how many items were finished, and not the pattern of responding.

Closely related to this are many of the other subtests of the Wechsler scales and similar types of indexes, where the items are presented in order of difficulty. Again, the expected pattern of answers is that they will all be correct until the difficulty level exceeds the person's ability and the remaining items would be wrong; or there should be a number of two-point responses, followed by some one-point answers, and then zeros. If α is computed for these types of tests, it will result in a very high value, one that is only marginally below 1.0.

Third, α is inappropriate if the answer to one item depends on the response to a previous one, or when more than one item deals with a single problem. This would arise, for example, if the person has to read a passage and respond to a series of questions about it. The reason is that if the person does not understand or miscomprehends the paragraph, then this will affect a number of items. When these items end up in the different halves, it will spuriously inflate the correlation between them.

Finally, as I have discussed earlier, α should not be used if it is suspected that the scale is multifaceted. If there are more

than 20 or so items, α can be quite respectable, giving the misleading impression that the scale is homogeneous.

So, how high should α be? In the first version of his book, Nunnally (1967) recommended .50 to .60 for the early stages of research, .80 for basic research tools, and .90 as the "minimally tolerable estimate" for clinical purposes, with an ideal of .95. He increased the starting level to .70 in later versions of his book (Nunnally, 1978; Nunnally & Bernstein, 1994). In my opinion (and note that this *is* an opinion, as are all other values suggested by various authors), he got it right for research tools, but went too far for clinical scales. As outlined in Myth 3, except for extremely narrowly defined traits (and I can't think of any), αs over .90 most likely indicate unnecessary redundancy rather than a desirable level of internal consistency.

CONCLUSIONS

Internal consistency is necessary in scales that measure various aspects of personality (a subsequent article will examine situations where it is not important). However, Cronbach's α must be used and interpreted with some degree of caution.

1. You cannot trust that published estimates of α apply in all situations. If the group for which the scale will be used is more or less homogeneous than the one in the published report, then α will most likely be different (higher in the first case, lower in the second).
2. Because α is affected by the length of the scale, high values do not guarantee internal consistency or unidimensionality. Scales over 20 items or so will have acceptable values of α, even though they may consist of two or three orthogonal dimensions. It is necessary to also examine the matrix of correlations of the individual items and to look at the item-total correlations. In this vein, Clark and Watson (1995) recommended a mean interitem correlation within the range of .15 to .20 for scales that measure broad characteristics and between .40 to .50 for those tapping narrower ones.
3. Values of α can be too high, and point to redundancy among the items. I recommend a maximum value of .90.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Antony, M. M. (2001). Assessment of anxiety and the anxiety disorders: An overview. In M. M. Antony, S. M. Orsillo, & L. Roemer (Eds.), *Practitioner's guide to empirically based measures of anxiety* (pp. 7–17). New York: Kluwer Academic/Plenum.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the administration and scoring of the MMPI–2*. Minneapolis: University of Minnesota Press.

Caruso, J. C. (2000). Reliability generalization of the NEO personality scales. *Educational and Psychological Measurement, 60,* 236–254.

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

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

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

Hamilton, M. A. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6,* 278–296.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177–189.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404–420.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). New York: Harcourt Brace Jovanovich.

Koksal, F., & Power, K. G. (1990). Four Systems Anxiety Questionnaire (FSAQ): A self-report measure of somatic, cognitive, behavioral, and feeling components. *Journal of Personality Assessment, 54*, 534–545.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lang, P. J. (1971). The application of psychophysiological methods. In S. Garfield & A. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (pp. 75–125). New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McClelland, D. C. (1980). Motive dispositions: The merits of operant and respondent measures. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1; pp. 10–41). Beverly Hills, CA: Sage.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test of split-halves. *Harvard Educational Review, 9*, 99–103.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.) Oxford, England: Oxford University Press.

Wechsler, D. (1997). *WAIS–III administtation and scoring manual* (3rd ed.). San Antonio: TX: Psychological Corporation.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60,* 201–223.

David L. Streiner
Kunin-Lunenfeld Applied Research Unit
Baycrest Centre for Geriatric Care
3560 Bathurst Street
Toronto, Ontario, Canada M6A 2E1
E-mail: dstreiner@klaru-baycrest.on.ca

Received May 28, 2002
Revised June 29, 2002