DOCUMENT RESUME

ED 438 307		TM 030 618
AUTHOR	Guthrie, Abbie C.	
TITLE	A Review of Coefficient Alpha and Some Basic Tenets of Classical Measurement Theory.	
PUB DATE	2000-01-27	
NOTE	24p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).	
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC01 Plus Postage.	
DESCRIPTORS	*Effect Size; *Measurement Techniques; *Reliability; *Scores	
IDENTIFIERS	*Alpha Coefficient	

ABSTRACT

Too many researchers speak of "the reliability of the test," thus indicating their basic misunderstanding of reliability. This paper explains classical reliability and the score features that influence coefficient alpha. It explains when coefficient alpha can be negative, even though it is conceptually a variance-accounted-for statistic. The recent recommendations of the American Psychological Association Task Force on Statistical Inference emphasize score reliability, because poor score reliability attenuates detected effect sizes. (Contains 24 references.) (Author/SLD)

A Review of Coefficient alpha

And Some Basic Tenets of Classical Measurement Theory

Abbie C. Guthrie

Texas A&M University 77849-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 27, 2000.

Abstract

Too many researchers speak of "the reliability of the test," thus belying their basic misunderstanding of reliability. The paper explains classical reliability, and the score features that influence coefficient alpha, including when it can be negative even though alpha is conceptually a variance-accounted-for statistic. The recent recommendations of the APA Task Force on Statistical Inference emphasize score reliability, because poor score reliability attenuates detected effect sizes.

It is common for students, practitioners and even scholars to speak of "the reliability of the test" or to say, "the test is reliable" when referring to an instrument of measurement. This unfortunate turn of phrase belies a basic confusion about the concept of reliability and further spreads the disease of misunderstanding. Pedhazur and Schmelkin (1991) wrote, "Statements about the reliability of a measure are . . . inappropriate and potentially misleading" (p. 82). Thompson (1992) explained how this seemingly innocuous and efficient way of speaking could be quite insidious:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. (p. 436)

For example, if a naïve researcher selects an instrument with a published reliability coefficient of .92, he or she may confidently believe that data collected will magically have the same reliability as was obtained in the normative sample. As a result, this researcher will probably not bother to evaluate the reliability of the data in hand and may thus grossly misinterpret their own results.

Score Reliability Impacts Effect Sizes

The APA Task Force on Statistical Inference recently emphasized, authors should "Always provide some effect-size estimate when reporting a *p* value" (p. 599, emphasis added). Later the Task Force also wrote,

> Always present effect sizes for primary outcomes.... It helps to add brief comments that place these effect sizes in a practical and theoretical context.... We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. (p. 599, emphasis added)

Kirk (1996) and Snyder and Lawson (1993) provided useful summaries of what various effect sizes can be computed in interpreting research results.

The Task Force also explained 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 (Feldt & Brennan, 1989). Thus, authors should provide reliability coefficients of the scores for the data being analyzed even when the focus of

> their research is not psychometric. (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 596)

The Task Force emphasized that, "Interpreting the size of observed effects requires an assessment of the reliability of the scores" (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 596), because score reliability attenuates detected study effects, and <u>these score reliability attenuations thus must be considered as part of result interpretation</u>.

<u>Tests are Not Reliable</u>

However, it is important to realize that reliability is a characteristic of <u>scores</u> not tests. Scholars in the fields of measurement and research methodology have been declaring this for years. Rowley (1976) wrote, "It needs to be established that an instrument itself is neither reliable nor unreliable... A single instrument can produce <u>scores</u> which are reliable, and other <u>scores</u> which are unreliable" (p. 53, emphasis added). Echoing this, Crocker and Algina (1986) noted, ". . . A test is not 'reliable' or 'unreliable'. Rather, reliability is a property of <u>scores</u> on a test for a <u>particular group of examinees</u>" (p. 144, emphasis added).

In an effort to clarify the meaning of reliability, Gronlund and Linn (1976, p. 106, emphasis in original) made this point:

> Reliability refers to the results obtained with an evaluation instrument and not to the instrument itself. Any particular instrument may have a number of different reliabilities, depending on the group involved and the situation in which it is used. Thus it is more appropriate to speak of the reliability of 'the test scores,' or of 'the measurement,' than of 'the test,' or 'the instrument.'

The present paper illustrates why reliability is score-dependent by focusing on what reliability is conceptually and statistically. Following a brief review of classical test measurement theory as it relates to the concept of reliability, the paper primarily focuses on evaluation of internal consistency, specifically Cronbach's alpha, exploring what factors influence alpha. Finally, the paper addresses the importance of reporting reliability coefficients in published research.

The True Score Model and Reliability

Estimations of reliability seek to answer an important question as to how accurate, and therefore reproducible, are the scores on a measurement (Thorndike, Cunningham, Thorndike, & Hagan, 1991). A brief overview of the true score model of classical measurement theory will assist in the understanding of reliability. Dawson (1999) and Thompson and Vacha-Haase (in press) provide more complete reviews of these issues.

Ideally, a score obtained by an individual on a measurement would be exactly equal to the characteristic being measured, whether the characteristic is knowledge about math, physical strength, or attitudes about school. If this were true, we would be able to repeatedly test an individual with the same or similar instruments, obtaining identical scores each time. In reality, however, individuals' scores on instruments vary on repeated testing, because measurement is imperfect.

Classical measurement theory (cf. Dawson, 1999) accounts for this variation with the true-score model that partitions any observed score (X_i) on a measurement into two components--a true score (T_i) and an error score (E_i). Symbolically, the equation for this model is:

$$X_i = T_i + E_i \ ,$$

where X_i is the observed score, T_i is the true score, and E_i is the error score.

Theoretically, the true score represents the actual amount in the examinee of the characteristic being measured. Imagine that a test was administered an infinite number of times to an examinee (amazingly without causing any change in the characteristic or the examinee), and each time the examinee's score was placed in a distribution. The average of these infinite scores would be the examinee's true score. The true score (i.e., the characteristic being measured) is thought to be a constant, and any variation in the true score is attributed to error in measurement, hence, the error score (Crocker & Algina, 1986; Feldt & Brennan, 1993; Sax, 1974; Thorndike et al., 1991).

Error scores may be positive or negative in value. For example, guessing correctly on an exam would enhance an examinee's observed score, so that the observed score overestimates the true score. On the other hand, having a fender-bender on the way to the SAT may likely result in poorer performance, detracting from the examinee's observed score.

It is clear that taking an infinite number of measurements is impossible. So, how is one to know the true score? Well, there is good news and bad news. The bad news is that we don't get to know the true score. The good news is that reliability coefficients can help us estimate how much of the observed score is accounted for by the true score.

Accounting for Error

In classical measurement theory, there are three primary ways to estimate account for measurement error, each based on identifying a single source of error. The three sources of error are: (a) errors due to instability over time, (b) errors due to difference in test forms, and (c) errors due to inconsistency in a single instrument. Classical measurement theory attempts expresses reliability in terms of the ratio or percentage of true score variance that can be explained in the total observed variance. That is, what amount of variability in scores (i.e., observed score variance) is due to the variability among examinees of the characteristic being measured (i.e., true score variance) (Crocker & Algina, 1986; Dawson, 1999; Eason, 1991). Pedhazur and Schmelkin (1991) provide a thorough statistical explanation of this concept (pp.83-86).

Stability and Equivalence

The reliability coefficient of stability attributes error (i.e., variation) in scores to change in the examinee over a period of time. To obtain this reliability estimate, a measurement is given to a sample of examinees, and then the same instrument is administered again to the same examinees after a period of time. The correlation of the scores on the two administrations is the coefficient of stability. Correlation looks at how well two measures put the same people in the same order and whether score relationships are monotonic.

The reliability coefficient of equivalence looks at the variation in scores on two different forms of the same test. Some test developers want to develop parallel forms of a measurement in order to take repeated measures of the same construct, without item memory being a factor in score variability. To evaluate the equivalence of forms, a sample of examinees is divided in half with one half taking Form A and the other half taking Form B. Then, the examinees take the measurement again, with each half using the opposite form. The time period between the measurements is kept to a minimum, to diminish variability due to change in the examinees over time. Again, the scores are correlated, pairing the scores on the different forms.

A high correlation indicates parallel or equivalent forms that can be used interchangeably. It is important to remember, however, that a high coefficient of equivalence only shows that the forms tend to order the examinees in the same order and at the same intervals. It does not tell you that the scores themselves are equivalent. Suppose three students' paired scores on two forms of a math achievement test are (a) 95 and 60, (b) 85 and 50, and (c) 55 and 20. The two forms will have perfect equivalence, but it is clear they do not measure the construct equally well.

Internal Consistency: Split-Half Method

Due to the obvious impracticality of developing parallel forms and performing multiple administration of measurements, the most commonly used estimation of reliability is that of internal consistency. Estimates of internal consistency address the question, "To what degree is the variability in observed scores due to common factors?" (Thorndike et al., 1991) All measures of internal consistency are somewhat analogous to the equivalent forms method.

The simplest way to estimate internal consistency is the split-half method. In this method, the items of an instrument are divided in half, usually randomly or by odd and even numbers, and a score is computed for each half. These scores are correlated, yielding a split-half coefficient. However, this coefficient is based on a hypothetical instrument that is only half the length of the original instrument. The Spearman-Brown formula is used to correct for this:

$$r_{xx} = 2r_{1/2\ 1/2} \ / \ 1 + r_{1/2\ 1/2} \ ,$$

where r_{xx} = the reliability of a measurement, and $r_{1/2\ 1/2}$ = the correlation between its two halves. For example, if the correlation between two halves of an instrument were .72, the Spearman-Brown corrected estimate of reliability would be

$$r_{xx} = 2(.72) \ / \ 1 + .72 = .84 \ .$$

The Spearman Brown correction shows that reliability is expected to increase with the addition of items of the same quality. This makes sense given that reliability is a ratio of true variance to total variance, and increasing a sample of items usually (but not always) increases the variability of scores on those items.

The ironic problem with the split-half method of reliability estimation is that it may be inconsistent, due to the fact that there are many ways to split a test. One formula for calculating the number of possible split halves of an instrument with k items is:

$$[.5(k!)]/[(.5(k))!]^2 .$$

Using this formula, it is possible to figure out that there are three possible splits for a test with k = 4 items, 10 possible splits for a test with k = 6 items, and 126 possible splits for a test with k = 10 items. Since each different split could yield a different reliability coefficient, it is clear that as the number of items increases, calculating the split-half reliability coefficient is like pulling the coefficient out of a hat.

Internal Consistency: Coefficient alpha

In 1951, Cronbach presented a formula for coefficient alpha that yields the theoretical mean of all possible split-half coefficients. Crocker and Algina (1991) described Cronbach's alpha as a "lower bound" estimate of the reliability coefficient, meaning that the actual reliability cannot be any lower than alpha, but can be and usually is higher (by how much is impossible to tell).

Alpha is not a simple correlation of scores, although it is analogous to the parallel form of reliability estimation. In a sense, coefficient alpha treats all items in an instrument as if they were each a parallel form of the same measurement, each measuring the same construct. The formula for coefficient alpha is:

$$\alpha = (k/k-1)\ [1-(\Sigma\sigma_k^2/\sigma_T^2)],$$

where k = number of items, σ_k^2 = item variance, and σ_T^2 = total test variance.

For a more thorough understanding of alpha, one might consider an alternate formula for total test variance (σ_T^2) presented in Pedhazur and Schmelkin (1991, p. 93):

$$\sigma_T^2 = \Sigma\sigma_k^2 + [\Sigma COV_{ij}\ \ x\ 2],$$

where COV_{ij} = the covariance of items i and j ($i \neq j$). Thompson (1999) provides an excellent treatment of the implications of this alternate formula.

Conceptually, "alpha measures how internally consistent scores are based on the degree to which item scores measure the same construct" (Thompson, 1999, p. 13). To do this, the formula incorporates how performance on items correlates with overall test scores and how items correlate with one another. What you get is the proportion of total test variance that can be explained by common factors (Reinhardt, 1996). That is, factors that are consistent throughout the measurement and would, theoretically, remain consistent across measurements.

For example, if (and only if) items have no correlation with one another, hence, no covariance, the sum of the item variances will be equal to the total test variance (Sax, 1974). When applied to the formula for alpha, you will see that $\alpha = 0$ in this case. This means there is no consistency in the construct being measured by the items. Interestingly, if items correlate negatively with one another (i.e., one item goes up and the other goes down), the covariance of those two items will also be negative. If enough item covariances are negative, then the sum of the covariances will also be negative, making the sum of item variances larger that the total test variance. The repercussion of this would be that alpha would be negative! Both Reinhardt (1996) and Thompson (1999) illustrated in detail how this is possible, even though alpha is a variance-accounted-for statistic.

The most influential factor in coefficient alpha is the total test variance. Reinhardt's (1996) excellent exploration of factors that affect coefficient alpha using a mini Monte Carlo method showed that the magnitude of total test variance accounted for 60% of the variance in alpha. This finding underscores the concept that reliability is dependent on the sample, and thus the scores from that sample. Thompson (1994a) writes, "Reliability is driven by variance--typically, greater score variance leads to greater score reliability . . . more heterogeneous samples often lead to more variable scores, and thus to higher reliability" (p. 3, emphasis in original).

As a measure of replicability, reliability indicates how much or how little people would stay in the same order on repeated measures (Gronlund & Linn, 1976). Intuitively, it makes sense that the more spread out scores are from one another (i.e., increased variance), the less likely they are to shift around if measured again.

Another important factor in reliability is the length of the test. Recall the Spearman-Brown correction formula used the calculation of the split-half coefficient. That formula is a special form of the Spearman-Brown Prophecy formula:

$$r_{kk}' = kr_{tt} / 1 + (k-1)r_{tt} ,$$

where r_{kk}' = the predicted reliability when length of a test is adjusted by a factor of k, and r_{tt} = the reliability of an instrument. Using this formula, one can see the affect of test length on reliability.

For example, if a test of 20 items has a reliability of .70, the predicted reliability for the same measurement at 30 items would be:

$$r_{kk}' = 1.5(.70)/1 + (1.5-1).70 = .77 .$$

For 50 items, the new reliability would be .83, and for 80 items the predicted reliability would be .90. The reason for the increase in estimated reliability with increase sampling of items is explained by Thorndike (1991, p. 107):

> As the length of the test is increased, the chance of errors of measurement [being random] more or less cancel out, the score comes to depend more and more completely on the characteristic of the person being measured, and a more accurate appraisal of the individual is obtained.

Of course, this assumes that the quality of the items would be equal to or greater than the quality of the existing items.

Classical reliability estimates are each limited to the type of error they are designed to detect. Thus, coefficient alpha tells one nothing about stability of the measurement over time. Further, alpha is not an indicator of unidimensionality (i.e., performance explained by one underlying factor) of the measurement (Reinhardt, 1996). As a result, some scholars see them as simplistic. As Eason (1991) stated, "The inability to analyze more than source of error variance at a time severely limits classical test theory as a psychometric approach" (p. 83). Newer, more statistically complex methods have been developed that are able to more clearly define sources of error variance (cf. Eason, 1991; Lawson, 1991). Generalizability theory is one such method (Eason, 1991). Generalizability theory uses analysis of variance (ANOVA) to "[consider] the multiple sources of error that may influence scores, as well as the interaction effects of error influences" (Eason, 1991, p. 84).

Importance of Reporting Reliability

Because reliability is a function of the scores obtained on a particular administration of an instrument to a particular group of people, it is common sense that an estimate of reliability should be calculated for any measurement. Thompson (1992) writes, "One important implication of the realization that reliability inures to data (rather than tests) is that reliability should . . . be explored whenever data are collected (p. 436). Furthermore, the results of that analysis should be included in any report of substantive research. Vacha-Haase (1998) noted, "Given the diversity of participants across studies, simple logic would dictate that authors of every study should provide reliability coefficients of scores for the data being analyzed, even in nonmeasurement substantive inquiries" (p. 8).

Reviews of reliability reporting practices in journals have not born out that these convictions are widely held. Willson (1980) reviewed the reliability reporting practices in the American Education Research Journal, finding that only 37% reported reliability coefficients for the data being analyzed, condemning it a "inexcusable at this late date" (p. 9). Vacha-Haase (1998) researched 628 articles of substantive research using the BEM Sex Role Inventory for a meta-analytic study, finding only 13% provided reliability information for the data in hand, and 65% did not make any mention of reliability at all. While others have found more promising results (cf. Thompson & Snyder, 1998), it is clear that progress needs to be made. Perhaps change is coming, given the recent report by the APA Task Force on Statistical Inference (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 596).

Why should researchers care about the reliability of their data? Because "score reliability inherently attenuates effect sizes" (Thompson, 1994b, p. 840). Just as coefficient alpha is a ratio of item variance to total test variance, the r^2 effect size in a ratio of $SOS_{EXPLAINED}/SOS_{TOTAL}$. Thorndike (1991) explained the impact of the reliability of data on the correlation of two measures in the following equation:

$$r_{12} \leq [(r_{11})(r_{22})]^{.5},$$

where r_{12} is the possible correlation of two measures, r_{11} and r_{22} are the respective reliability coefficients for the data obtained by each measurement. Therefore, effect sizes such as r^2 cannot exceed the product of the reliability of the scores of two measures. For example, if you are studying two measurements, with reliability coefficients of .75 and .82 respectively, the detected effect size will only be .62 even if the two variables are perfectly correlated. It is important to know and report the impact of the reliability coefficients on possible results prior to doing research and retrospectively in interpreting data. Thompson and Snyder (1998) asserted:

> The concern for score reliability on substantive inquiry is not just some vague statistician's nit-picking. Score reliability directly a) affects our ability to achieve statistically significance and b) attenuates the effect sizes for the studies we conduct. (p. 76)

Vacha-Haase, Ness, Nilsson, and Reetz (1999) agreed, stating, "Score reliability is critically important, even in substantive (i.e., nonmeasurement) studies, because the score reliability of the data being analyzed directly affects substantive results and their interpretation" (p. 336).

Progress is slow; and while generalizability theory may be promising in its ability to define sources of error and identify interaction effects of error, it seems unlikely that it will be widely used and reported at this time. Generalizability theory is a statistical back handspring compared to the cartwheel of coefficient alpha. At this time, most researchers still do not analyze or report even the classical estimates of reliability of data in substantive studies. It is doubtful that researchers will perform back handsprings when they are not yet even consistently practicing cartwheels.

References

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 197-334.

Dawson, T. E. (1999). Relating variance partitioning in measurement analyses to the exact same process in substantive analyses. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 101-110). Stamford, CT: JAI Press.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 105-146).

Gronlund, N.E., & Linn, R. L. (1976). Measurement and evaluation in teaching (3rd ed.). New York: Macmillan.

Kirk, R. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), Advances in educational research: substantive findings, methodological developments (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

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

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Sax, G. (1974). Principles of educational measurement and evaluation. Belmont, CA: Wadsworth.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1994a, April). Common methodology mistakes in dissertations, revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 368 771)

Thompson, B. (1994b). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1999, February). Understanding coefficient alpha, really. Paper presented at the annual meeting of the Education Research Exchange, College Station, TX.

Thompson, B. & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling and Development research articles. Journal of Counseling and Development, 76, 436-441.

Thompson, B., & Vacha-Haase, T. (in press). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagan, E. P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: Macmillan.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. Journal of Experimental Education, 67(4), 335-341.

Wilkinson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: **http://www.apa.org/journals/amp/amp548594.html**]

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com