Edgeworth Series in Quantitative
Behavioral Science (ESQBS), UNBC Paper #99-1

A Glance at Coefficient Alpha with an Eye Towards Robustness Studies: Some Mathematical Notes and a Simulation Model

Bruno D. Zumbo, Ph.D.
Professor of Psychology & of Mathematics
University of Northern British Columbia

Abstract

In many testing situations the use of parallel forms or test-retest reliability coefficients is impractical. Viable alternatives to these approaches are the internal consistency coefficients. One of the most commonly used internal consistency coefficients is Cronbach's coefficient alpha. The purpose of this paper is to discuss the robustness of coefficient alpha. To begin, I will provide a derivation of coefficient alpha. In the process of the derivation, I will try to provide insight into alpha's potential robustness. The paper will close with a simulation algorithm for studying alpha's robustness.

Paper 99-1 of the Edgeworth Series in Quantitative Behavioral Science.

When citing this paper please use the following format (for APA style):

Zumbo, B. D. (1999). *A glance at coefficient alpha with an eye towards robustness studies: Some mathematical notes and a simulation model* (Paper No. ESQBS-99-1). Prince George, B.C.: University of Northern British Columbia. Edgeworth Laboratory for Quantitative Behavioural Science.

Author note: This work was inspired by Ms. Sharon Shultz during the preparation of her thesis work. Her questions made me realize that a working paper may be needed explaining Zimmerman, Zumbo, and Lalonde's simulation algorithm. Send any correspondence to: Dr. Bruno D. Zumbo, University of Northern British Columbia, Prince George, B.C., e-mail: zumbob@unbc.ca (Internet).

1.0 Introduction

When resarchers require an estimate of reliability of their measurements (either test or scale reliability), parallel forms and test-retest approaches are often impractical. Coefficient alpha is often the reliability estimate of choice. Conceptually, coefficient alpha is the mean of all possible split-half correlations and estimates the lower bound[1] of the reliability. Alpha can be used for both dichotomous and ordered polytomous data and requires only one test administration. Without lack of generality, only the polytomous case will be considered in this paper.

2.0 Yet Another Derivation of Coefficient Alpha

Coefficient alpha can be derived in various ways, for example, derivations based on an ANOVA approach, a composite measures approach, or through linear operators and Hilbert space. For this study, the derivation based on composite measurements will be used. The derivation that follows is based on Novick and Lewis (1967).

The observed test score model in classical test theory is expressed as:

$$X = T + E \qquad (1)$$

where X is the observed score, T is the true score, and E is the measurement error score. Recall that the true score is the expected value of the propensity distribution. With this model it is assumed that the measurement error scores for an examinee are uncorrelated with that individual's true scores, the item error scores are uncorrelated, and the measurement error scores are expected to sum to zero over the population of examinees.

[1] A lower bound is a value that must be smaller than the reliability coefficient. That is, the reliability coefficient must be at least as big as the value for alpha.

The reliability can be expressed as

$$\frac{\sigma_{T_c}^{2}}{\sigma_{C}^{2}} \tag{2}$$

where $\sigma_{T_c}^{2}$ denotes the variance of the composite true score and σ_{C}^{2} denotes the variance of the composite observed test score or the variance of the sum of **k** parallel subtest scores (without restriction on the results the subtests can also represent items).

If all **k** parallel measures have equal true score variances and equal true score covariances where by definition the true score variance equals the sum of the elements of the **k** X **k** matrix of true score variances and covariances, then:

$$\sigma_{T_c}^{2} = k\sigma^{2}(T_i) + k(k-1)\sigma(T_i, T_j) \tag{3}$$

where $\sigma(T_i, T_j)$ denotes the covariance between T_i and T_j, and $\sigma^{2}(T_i)$ denotes the variance of T_i.

Based on the fact that $\sigma(T_i, T_j) = \sigma^{2}(T_i)$,

$$\sigma_{T_c}^{2} = k\sigma(T_i, T_j) + k(k-1)\sigma(T_i, T_j) \tag{4}$$

or

$$\sigma_{T_c}^{2} = k^{2}\sigma(T_i, T_j). \tag{5}$$

However, when the **k** subtests are not strictly parallel then

$$\sigma_{T_g}^{2} \geq \sigma(T_i, T_g). \tag{6}$$

For nonparallel subtests at least one subtest (in this case subtest **g**) has a true score variance which is greater than or equal to its covariance with any other subtest. But for any two subtests that are not strictly parallel the sum of their true score variances is greater than or equal to twice their covariances,

$$\sigma_{T_i}^{\ 2} + \sigma_{T_j}^{\ 2} \geq 2\sigma(T_i, T_j). \qquad (7)$$

In addition, with nonparallel subtests, the sum of **k** true score variances will be greater than or equal to the sum of **k(k-1)** covariances divided by **(k-1).** This can be denoted as

$$\sum \sigma_{T_i}^{\ 2} \geq \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j)}{k-1}, (i \neq j). \qquad (8)$$

When the sum of the covariances is added to each side of the equation then,

$$\sum \sigma_{T_i}^{\ 2} + \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j), (i \neq j) \geq \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j)}{k-1} + \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j), (i \neq j). \qquad (9)$$

As seen in equation 3, the variance of the true score composite equals the sum of the variances of the true scores plus the sum of the covariances of the true scores. Therefore, equation (9) can be expressed as:

$$\sigma_{T_C}^{\ 2} \geq \frac{k \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j)}{k-1}, (i \neq j) \qquad (10)$$

or

$$\sigma_{T_C}^{\ 2} \geq \frac{k}{k-1} \sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j), (i, \neq j) \qquad (11)$$

where $\sum_{i=1}^{k} \sum_{j=1}^{k} \sigma(T_i, T_j)$, $i \neq j$ is the sum of **k(k-1)** covariances of non parallel subtests.

If both sides of equation (11) are divided by the variance of the composite then

$$\frac{\sigma_{T_C}^{2}}{\sigma_C^{2}} \geq \frac{k}{k-1}\left[\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma(T_i, T_j)}{\sigma_C^{2}}\right], (i \neq j). \qquad (12)$$

This is reliability since, by definition, reliability is the ratio of true score variance to observed score variance.

Since the true score covariance equals the observed score covariance,

$$\frac{\sigma_{T_C}^{2}}{\sigma_C^{2}} \geq \frac{k}{k-1}\left[\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\sigma(X_i, X_j)}{\sigma_C^{2}}\right], (i \neq j). \qquad (13)$$

In addition, it can be shown that the observed score variance equals the sum of the covariances and the item variances. Therefore,

$$\frac{\sigma_{T_C}^{2}}{\sigma_C^{2}} \geq \frac{k}{k-1}\left[\sigma_C^{2} - \sum\frac{\sigma_{X_i}^{2}}{\sigma_C^{2}}\right], (i \neq j). \qquad (14)$$

Hence, coefficient alpha is the lower bound estimate of reliability.

3.0 Assumptions Underlying Coefficient Alpha

For the purposes of this study, the underlying assumptions of coefficient alpha can be categorized as assumptions underlying the derivation of alpha and assumptions underlying the estimation of alpha. A discussion of these categories follows.

3.1 Assumptions Underlying the Derivation

Derivations of coefficient alpha, such as the derivation of Lord & Novick (1967), used a classical test theory approach to the derivation. This method does not assume a normal observed score distribution. However, an ANOVA approach to the derivation of coefficient alpha, such as the derivation presented by Feldt (1965), involves the assumptions of ANOVA which include normality. The assumptions underlying the derivation also include additivity (Novick & Lewis, 1967; Lord & Novick, 1968; Zimmerman, 1969; Zimmerman, Zumbo, & Lalonde, 1993). Additivity means that the matrix of true scores must be additive in nature.

Zimmerman, Zumbo, and Lalonde (1993) found that violation of the assumptions of uncorrelated errors between subtests and additivity resulted in a greater variability of the estimator. When additivity was violated, alpha underestimated the reliability. However, when the uncorrelated errors assumption was violated, alpha overestimated the reliability.

3.2 Assumptions Underlying the Estimation

Although normality is not an assumption of the derivation of coefficient alpha presented above, it is an assumption of the estimation of coefficient alpha. The assumptions of estimation stem from least squares estimation. As seen above, coefficient alpha is a function of sample variances and covariance. Therefore, the assumptions of least-squares estimation, which include normality, apply to coefficient alpha. Future research should investigate how nonnormal error score distributions affect the estimation of the population reliability using coefficient alpha.

Commonly used approaches to most problems in classical test theory are based on statistics that are not robust or resistant to nonnormality. The terms "robust" or "resistant" are defined as insensitivity to changes in the underlying distribution

(Huber, 1981; Mosteller & Tukey, 1977). For example, the variance is not robust or resistant to nonnormality since small deviations from normality can greatly affect its value (Lind & Zumbo, 1993; Shoemaker & Hettmansperger, 1982). Consequently, Lind and Zumbo (1993) called for the investigation of the robustness or resistance of measures from classical test theory, such as coefficient alpha, to nonnormal data. A recent study by Wilcox (1992), developed a new measure of reliability which is a robust analogue of coefficient alpha. Wilcox (1992) derived this analogue based on the assumption that coefficient alpha is not robust to even slightly nonnormal observed score distributions. However, the robustness of coefficient alpha has not been studied in depth.

To the author's knowledge very little empirical or analytical research has been conducted to investigate the violation of the normality assumption underlying the estimation of coefficient alpha. As indicated in the review by Feldt, Woodruff, and Salih (1987), a sampling theory of coefficient alpha would allow researchers to obtain an unbiased estimate of the population value, to establish confidence intervals for coefficient alpha, and to determine if coefficient alpha has a specific value for a given population.

The exact sampling distribution for coefficient alpha has not yet been determined. However, a transformation of coefficient alpha has been derived independently by Feldt (1965) and Kristof (1963). This transformation was based on an ANOVA derivation and has been proven to be distributed as an F distribution. Therefore, assumptions of least squares estimation, including normality, apply. In other words, if a normal distribution of observed scores is used, a transformation of coefficient alpha can be used to determine confidence intervals or to test hypotheses. The transformation of coefficient alpha is:

$$\frac{1 - \text{population reliability}}{1 - \text{coefficient } \alpha}.$$

Based on this transformation, Feldt et al. (1987) stated that coefficient alpha would tend to underestimate the population reliability and would be a biased estimator when the number of examinees is small (e.g. n = 50).

Feldt et al. (1987) presented a formula that would give an unbiased estimate of the population reliability:

$$[(N - 3)\alpha / (N - 1)] + 2 / N - 1.$$

Bay (1973) also derived coefficient alpha using a mixed model ANOVA approach. He also derived the same transformation equation that was presented by Feldt (1965). Based on these derivations, Bay (1973) suggested that the sampling distribution of the reliability estimate would be robust against the violation of the normality assumption if (a) large numbers of examinees are used, (b) the reliability is close to zero, or (c) a large number of subtests is used and the true score or test score kurtosis is close to zero.

Bay (1973) also performed a Monte Carlo computer simulation. The simulation involved 30 examinees and eight subtests with 2000 replications. Six different true score distributions were used: normal, uniform, exponential, the sum of two independent uniform distributions, the sum of three independent uniform distributions and the sum of six independent uniform distributions. In addition, three different error score distributions were used, the normal, exponential, and uniform. The means, variances, and the mean squared errors of the alphas were obtained. The results of the computer simulation allowed Bay to conclude that a leptokurtic true score distribution could cause coefficient alpha to seriously underestimate the population reliability, and the effect of nonnormality of error score distributions is negligible when a large number of subtests is used.

A small component of Zimmerman, Zumbo, and Lalonde (1993) involved the estimation of coefficient alpha under a more general test score model than Bay. Forty examinees, ten subtests, and 2000 replications were used. The error score distributions used were the normal, uniform, exponential, and the mixed-normal. Reliability values of 0.65, 0.75, and 0.90 were used. Based on these parameters, Zimmerman et al. (1993) found that coefficient alpha was unbiased and its efficiency did not change over the distributions.

One difference between Bay (1973) and Zimmerman et al. (1993) should be noted. Bay (1973) varied the true score and the error score distributions while Zimmerman et al. only varied the error score distributions.

To date it has been shown that a transformation of coefficient alpha is distributed as a F distribution. This transformation is based on the assumption of normally distributed observed scores. Therefore, this transformation can only be used for hypothesis testing and to determine confidence intervals when the underlying observed score distribution is normal. In addition, the sampling distribution of coefficient alpha is unknown. If the sampling distribution of coefficient alpha were known, confidence intervals could be calculated from the data and hypotheses could be tested directly. This is similar to the commonly found discussion of computing confidence intervals of the mean based on the z-distribution.

Not only is the sampling distribution of coefficient alpha unknown, but also, the affects of nonnormality on the estimation of coefficient alpha have only been preliminarily tested by Bay (1973) and Zimmerman et al. (1993). Both Bay (1973) and Zimmerman et al. suggest that under certain conditions the estimation of reliability is robust. These conclusions are restricted, however, to the very limited parameters investigated.

The effects of nonnormality with various numbers of examinees, numbers of subtests, and population reliabilities have not been examined.

4.0 A Simulation Model

In the simulation model, the error scores of a given person over replications were distributed according to the various distributions. This implied that for any given replication, the error scores over examinees are distributed in the same way. The distributional shape, as in all simulation studies, needs to chosen to both represent a wide variety of conditions and parameters of interest (e.g., skewness and kurtosis, or outlier contamination) but also with a consideration of limitations based on what could be reasonably found in the practise of psychometrics.

The following algorithm would help us investigate the robustness:

Step 1: The population reliability, the number of subtests, and the number of examinees, as well as the error score distribution are input into the algorithm as constants.

Step 2: A true score matrix was created so that the assumption of additivity was met. The true score matrix was created by constructing a matrix with test items on one axis and examinees on the other axis and adding the respective terms. In this way the true score distribution was uniform. Table 3 is an example of an additive true score matrix.

Table 3

Example of An Additive True Score Matrix

Examinee Number	Item Number			
	1	**2**	**3**	**4**
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

Step 3: The mean true score and its variance are calculated. Then using an error distribution with a mean of zero, independent error scores from the specified distribution are selected and added to the true score items. The variance of the error distribution is initially set to one and is then modified so that a specified reliability could be achieved.

Step 4: Coefficient alpha is calculated.

Step 5: The selection of an independent error score and the calculation of coefficient alpha are repeated to achieve 1000 replications.

Step 6: The mean, variance, mean squared error, skewness, and kurtosis of coefficient alpha over the 1000 replications are calculated. The 0.95 confidence interval of the mean is also calculated. The variance calculated is the variance from the sample mean of coefficient alpha while the MSE is the averaged squared deviation from the population value of alpha.

The simulation model (algorithm) is an adaptation of the methodology used in Zimmerman et al. (1993). It is important to note that given the model in Zimmerman et al. (1993) the distribution of observed scores over replications has the same shape as the distribution of error scores over replications. Finally, in these simulations, the additivity and uncorrelated error assumptions were satisfied.

The methodology discussed herein is different than the methodology used by Bay (1973). Bay (1973) varied both the true score and the error score distribution. The present simulation model only varies the error score distribution while the true score distribution remains uniformly distributed. It can be clearly seen by examining Table 3 that the total score across the four items is uniformly distributed for that sample.

REFERENCES

Bay, K.S. (1973). The effect of non-normality on the sampling distribution and standard error or reliability coefficient estimates under an analysis of variance model. British Journal of Mathematical and Statistical Psychology, 26, 45-57.

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

Feldt, L.S. (1965). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34(3), 363-373.

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. Psychometrika, 45(1), 99-105.

Feldt, L.S., Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient alpha. Applied Psychological Measurement, 93-103.

Huber, P.J. (1981). Robust Statistics. New York: Wiley.

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. Psychometrika, 28, 221-238.

Lind, J.C. & Zumbo, B.D. (1993). The continuity principle in psychological research: An introduction to robust statistics. Canadian Psychology, 34, 407-414.

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

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

Wilcox, R.R. (1992). Robust generalizations of classical test reliability and Cronbach's alpha. British Journal of Mathematical and Statistical Psychology, 45, 239-254.

Woodruff, D.J. & Feldt, L.S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. Psychometrika, 51(3), 393-413.

Zimmerman, D.W. (1969). An item sampling model for the reliability of composite test. Educational and Psychological Measurement, 29, 49-59.

Zimmerman, D.W. & Zumbo, B.D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), A Handbook for data analysis in the behavioural sciences (pp. 481-517). Hillsdale, N.J.: Lawrence Erlbaum.

Zimmerman, D.W., Zumbo, B.D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. Educational and Psychological Measurement, 53, 33-49.

Zumbo, B.D. (1995). The lurking assumptions in using generalizability theory to monitor individual's progress. In D. Laveault, B.D. Zumbo, M. Gessaroli, and M. Boss (Eds.), Modern theories of measurement: Problems and issues. Ottawa, Ontario: University of Ottawa.