AN EXAMINATION OF
TWO METHODS OF FORMING CONFIDENCE INTERVALS
FOR COEFFICIENT ALPHA

by

KIMBERLY ANNE BARCHARD

B.A., Simon Fraser University, 1993

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Department of Psychology)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1995

© Kimberly Anne Barchard, 1995

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representative. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Psychology

The University of British Columbia
Vancouver, Canada

September 10, 1995

ABSTRACT

Coefficient alpha (Cronbach, 1951) is a commonly used measure of reliability. Feldt (1965) and Hakstian and Whalen (1976) developed methods of forming confidence intervals for coefficient alpha. In this thesis, three issues related to these confidence intervals were examined. First, their performance under fixed and random sampling of conditions (e.g., items, raters) was compared. When conditions were fixed, the confidence intervals were accurate for the kinds of data studied. When conditions were random, as few as 77% of the .95 confidence intervals included the parameter. Second, some researchers have recently questioned what constitute the necessary assumptions of these confidence intervals. In response, it was shown in this thesis that, contrary to what has been claimed, sphericity is not sufficient, although compound symmetry is. Compound symmetry may, in fact, be necessary, but no definitive proof of this can be offered at this time. Third, the performance of the two confidence intervals under random sampling of conditions was explored in more detail. Neither the type of confidence interval used nor heterogeneity of conditions means had any effect on the performance of the confidence intervals. However, heterogeneity of variances reduced the proportion of confidence intervals including the population value. This effect was most pronounced when the population value was high (.90) and when the number of conditions was low (5).

The following conclusions were reached. Researchers conducting Monte Carlo studies of coefficient alpha should simulate the type of sampling that they are most interested in, as results depend on the type of sampling used. The Feldt (1965) and Hakstian and Whalen (1976) confidence intervals are precise when conditions are fixed; however, when conditions are random and variances heterogeneous, these intervals should be used with caution, especially if the data contain fewer than 20 conditions. The two types of confidence intervals perform very similarly, and hence the choice of method is left to the researcher. Lastly, a call is made for more robust procedures to be developed, and one possible approach to such a robust procedure has been identified.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

I would like to thank some of the many people who have supported and encouraged me during the two years I have been working towards this thesis. First, Dr. Ralph Hakstian, my supervisor, for his support and encouragement, and his very helpful suggestions. Dr. Marsha Schroeder, for her insightful and friendly feedback. Dr. Nand Kishor, for taking time out of his sabbatical. Rob Davies, for his encouragement, tolerance, understanding, and the generous use of his computer. Linda Scratchley, for feedback on an earlier draft of this thesis. Rachel Fouladi, for her programming help, and Dr. James Steiger, for his generosity in lending me some of his subroutines. Page, for her invaluable help in unraveling Box (1954). Lastly, a special thank-you to Ross Wooley, who was always willing to listen.

GENERAL INTRODUCTION

Often in Psychology, Education, and Business, when a test is constructed, each item is intended to measure the same construct. A person taking such a test would be expected to obtain similar scores on different items. For example, if an examinee obtained a high score on the first item, this would suggest that the person lies near the upper end of the distribution of the construct being measured, and hence should obtain a high score on other items as well. If this were generally true, then most of the variance in total test scores would be due to differences between people, and very little would be due to random error. Such a test would be characterized as internally consistent.

Coefficient alpha (Cronbach, 1951), and its mathematically equivalent forms KR-20 (Kuder & Richardson, 1937) and Hoyt's estimate of reliability (Hoyt, 1941), are the most commonly used measures of internal consistency. Two methods of forming confidence intervals for coefficient alpha have been developed. The first is an exact procedure which was developed by Feldt (1965). The second is an approximate procedure based on a normalizing transformation, and was developed by Hakstian and Whalen (1976). The purpose of this thesis is twofold. First, two theoretical issues of interest to those researching the use of these confidence intervals are examined. These issues are (1) whether these confidence intervals perform similarly regardless of whether items are fixed or random, and (2) the assumptions underlying these confidence intervals. Second, this thesis explores the performance of these confidence intervals with various kinds of data which might be encountered in real-world research situations.

TECHNICAL BACKGROUND

Introduction to Reliability

Whenever a person is measured twice on some psychological or physical trait, the measurements are likely to be slightly different. A myriad of factors may influence the measurement: slight variations in the physical or psychological functioning of the individual, changes in the measurement situation, distractions, guessing, administration errors, scoring errors, and so forth. The consistency of repeated measurements is called reliability and is one of the main criteria used in evaluating the quality of a measuring device. Reliability is important because if measurements of the same person on the same construct produced widely different results, conclusions based on a single measurement would not be dependable.

Classical Test Score Theory provides a model for discussing reliability. According to this theory, any observed score can be viewed as the sum of the examinee's *true score* in the domain being measured and an *error score*, as

$$X = T + E,$$

where X is the observed score, T is the true score, and E is the random error score.

Classical test score theory makes several assumptions about these true and error scores (Lord & Novick, 1968, p. 68). First, true scores are unchanging over time and across situations. Second, errors can be either positive or negative, and their expected value is zero. Third, true scores and error scores are uncorrelated with each other.

Many results can be derived from these assumptions. Two of these will be presented here to clarify further the concepts of true score and error score. First,

$$E(X) = E(T + E) = E(T) + E(E) = E(T) = T$$

Thus, the true score is the expected value of the observed scores. Second,

$$\sigma_X^2 = \sigma_{T+E}^2 = \sigma_T^2 + \sigma_E^2 + 2\sigma_{TE} = \sigma_T^2 + \sigma_E^2.$$

Thus, because true scores and error scores are assumed to be uncorrelated, observed score variance is the sum of true score variance and error variance.

Using this model, reliability, ρ_x, is defined as the proportion of observed-score variance which is due to true differences between people (Cronbach, Rajaratnam, & Gleser, 1963):

(1) $$\rho_x = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}.$$

Because true scores are not observable, reliability is commonly estimated by correlating parallel measures. Measures are defined as parallel if they are indistinguishable statistically: they have the same mean and variance, the same intercorrelations with external criteria, and errors of measurement are uncorrelated with each other (Gulliksen, 1950). The following is a proof that the correlation between parallel measures, $\rho_{X_1X_2}$, is equal to the reliability of either measure:

$$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}$$

where $\sigma_{X_1X_2}$ is the covariance between the two measures, and σ_{X_i} is the standard deviation of the i^{th} measure. Using the classical test score model, the observed score is the sum of the true score and the error score. Substituting in, we have

$$\rho_{X_1X_2} = \frac{\sigma_{(T_1+E_1)(T_2+E_2)}}{\sigma_{X_1}\sigma_{X_2}},$$

where T_i is the true score on the i^{th} measure, and E_i is the error score on the i^{th} measure. Expanding, this gives

$$\text{(2)} \qquad \rho_{X_1X_2} = \frac{\sigma_{T_1T_2} + \sigma_{E_1T_2} + \sigma_{T_1E_2} + \sigma_{E_1E_2}}{\sigma_{X_1}\sigma_{X_2}}.$$

The last three terms in the numerator of the right side of Equation (2) are 0, by assumption, and hence Equation (2) simplifies as

$$\rho_{X_1X_2} = \frac{\sigma_{T_1T_2}}{\sigma_{X_1}\sigma_{X_2}}.$$

Because X_1 and X_2 are parallel measures, $T_1 = T_2$ and $\sigma_{X_1} = \sigma_{X_2}$. Therefore,

$$\rho_{X_1X_2} = \frac{\sigma_T^2}{\sigma_X^2}.$$

Thus, the correlation between two parallel measures equals the reliability of either measure.

Several different kinds of reliability (test-retest, alternate forms, inter-rater, and internal consistency reliability) can be estimated this way, by forming parallel measures in different ways. Each of these kinds of reliability takes a different approach to the quantification of the basic reliability definition in Equation (1). Most of these forms of reliability require that measurements be taken at more than one point in time. Internal consistency measures, however, allow estimation of the reliability of a measure based on a single measurement occasion.

Estimating Internal Consistency Reliability

Internal consistency reliability focuses on the similarity of scores on different parts of the same test. It is usually assumed that the different parts of a test are a more or less random sample from the population of test parts that span the construct of interest, although it is possible to view the test parts as fixed. These parts may consist of individual items, questions or tasks, or sets of these. Each part is intended to measure the same construct, and hence differences between scores on different parts represent error.

The simplest indexes of internal consistency are split-half indexes. To calculate split-half reliability, the test is first divided into two halves. This can be done in several different ways, such as separating odd and even numbered items, randomly assigning items to the two halves, or matching items for content or for difficulty. Each examinee is given a score on each half-test.

If the two halves are parallel, then the correlation between these halves is an estimate of the reliability of either half, as shown above. This will be less than the internal consistency of the full-length test, because longer tests are more reliable, in general, and more internally consistent, in particular. To estimate the internal consistency of the test as a whole, the Spearman-Brown Prophecy formula is used. The "stepped-up" reliability, $\rho_{xx'}$, is given by

$$\rho_{xx'} = \frac{2\rho_{AB}}{1+\rho_{AB}},$$

where ρ_{AB} is the correlation between the two halves of the test, and $\rho_{xx'}$ is the reliability of the full length test. If the two halves of the test are not parallel, the stepped-up reliability coefficient will underestimate the reliability of the test.

An alternative to correlating the two halves of the test is to use Rulon's (1939) method. In this method, the difference between the scores on the two halves of the test is calculated for each examinee. Then, the variance of these differences is calculated, and is used as an estimate of the variance of the error scores in Equation (1). This variance is compared to the variance of the total test scores, as

$$\rho_{xx'} = 1 - \frac{\sigma_D^2}{\sigma_X^2},$$

where $\rho_{xx'}$ is the reliability of the full length test, σ_D^2 is the variance of difference scores, and σ_X^2 is variance of total scores. This approach assumes "that the difference between the two true scores for the two half-tests is constant for all individuals studied, and that the errors of measurement in the two half-scores are chance errors, and hence uncorrelated" (Rulon, 1939, p. 101).

The problem with either of these split-half reliability indexes is that they depend on the particular split that is chosen. Test items can be divided into two halves in a large number of ways, and the different divisions will usually result in somewhat different reliability estimates.

The ANOVA approach presented by Hoyt (1941) avoids the problem of arbitrary test division. In the application of repeated-measures ANOVA to the *k*-item test context, *n* people are measured on each of k items. The observed score of Person i on Item j, X_{ij}, can be viewed as the sum of four independent components:

$$X_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij},$$

where μ is the grand mean, π_i is the person effect for Person i, τ_j is the item effect for Item j, and ε_{ij} is the residual or error component.

Then, using standard ANOVA terminology, $MS_{persons}$ estimates the variance of observed scores, and MS_{error} estimates the error variance (Hoyt, 1941). An estimate of the internal consistency reliability can then be calculated from these mean squares as

$$r_{xx'} = \frac{MS_{persons} - MS_{error}}{MS_{persons}}.$$

This estimate is derived by substituting observed mean squares for expected mean squares, as follows:

$$E[MS_{persons}] = k\sigma_p^2 + \sigma_e^2;$$

$$E[MS_{error}] = \sigma_e^2.$$

Thus,

$$\frac{E[MS_{persons}] - E[MS_{error}]}{E[MS_{persons}]} = \frac{k\sigma_p^2}{k\sigma_p^2 + \sigma_e^2}. \tag{3}$$

Equation (3) represents the population value of Hoyt's estimate of reliability, and will be denoted ρ_α. This is the reliability of the mean (or sum) of *k* measurements (Winer, 1971). This formula corresponds to the general equation for reliability, given by Equation (1), because the numerator is the variance due to people, and the denominator is the total score variance for a composite of *k* conditions.

When items are scored dichotomously, Hoyt's estimate of reliability is algebraically equivalent to a previously derived estimate of internal consistency, known as Kuder-Richardson (1937) formula 20 (KR-20). The formula for KR-20 is as follows:

$$\text{KR-20} = \left(\frac{k}{k-1}\right)\frac{SD_{tot}^2 - \sum_i^k p_i q_i}{SD_{tot}^2},$$

where k is the number of items, SD_{tot}^2 is the variance of the total scores, and $\sum_i^k p_i q_i$ is the sum of the item variances for dichotomously scored items (p_i is the proportion of people who got Item *i* right, and $q_i = 1 - p_i$ is the proportion of people who got Item *i* wrong).

Coefficient alpha (Cronbach, 1951) is algebraically equivalent to both Hoyt's estimate of reliability and KR-20. Cronbach simply replaced $\sum_i^k p_i q_i$ from KR-20 with the sum of the item variances, $\sum_i^k (SD_i^2)$. Coefficient alpha is estimated as

$$r_\alpha = \left(\frac{k}{k-1}\right)\frac{SD^2_{tot} - \sum_i^k (SD^2_i)}{SD^2_{tot}},$$

where SD^2_{tot} is the variance of the scores on the total test, SD^2_i is the variance on a particular item, Item *i*, and *k* is the number of items. To prevent confusion, the symbol α will be reserved for its usual designation as the probability of committing a Type I error. The symbol r_α will denote the sample value of coefficient alpha, and ρ_α the parameter value, as given by Equation (3).

Novick and Lewis (1967) showed that coefficient alpha is equal to the average Rulon (1939) split-half reliability over all possible divisions of a test.

The Relationship Between Coefficient Alpha and the Reliability of a Test

The above formulas (KR-20, Hoyt's formula, and coefficient alpha) were all derived to estimate the reliability of a test, where reliability is defined as in Equation (1). However, the parameter values of these statistics do not necessarily equal the reliability of a test. Some assumptions need to be made. In their derivation, Kuder and Richardson (1937) assumed that the inter-item correlation matrix was of unit rank and that the intercorrelations were constant. Hoyt (1941) assumed that errors were independent and equally variable. Cronbach (1951) did not state the assumptions he was making, but rather referred the reader to previous derivations. However, the assumptions made during these original derivations are too restrictive.

Novick and Lewis (1967) showed that, if items are fixed, then items need only be essentially tau-equivalent for coefficient alpha to equal the reliability of the composite, as given by Equation (1). Two items would be called essentially tau-equivalent if their errors were independent, and $T_{ij} = T_{ij'} + c$, for all i, where T_{ij} is the true score of Person i on Item j, $T_{ij'}$ is the true score of Person i on Item j', and c is a constant. If items are not essentially tau-equivalent, but errors are independent, then the value of ρ_{α} is a lower bound to the reliability of a test, as given by Equation (1). However, if the average covariance among errors is not zero, then estimates of true score variance will be biased. If the average covariance among errors is positive, then true score variance (and hence reliability) will be over-estimated. If the average covariance is negative, then true score variance and reliability will be under-estimated.

If items are randomly selected from the population, then the average error covariance among the items in the population is zero (Cronbach, 1995), and thus ρ_{α} equals the reliability of a test. Cronbach et al.'s (1963) derivation of coefficient alpha used Generalizability Theory and an ANOVA model developed by Cornfield and Tukey (1956). Cronbach et al. (1963) assumed that conditions were a random sample from some universe of conditions over which the researcher wishes to generalize. They also assumed that conditions were *experimentally independent* (the score of Person p on Condition i does not depend on whether he or she has or has not previously been observed in other conditions), and that observed scores are at the interval level of measurement. Readers should note that experimental independence does not imply *statistical independence* among conditions. By using the ANOVA model of Cornfield and Tukey (1956), Cronbach et al. (1963) were able to derive coefficient alpha with only the above assumptions; the usual assumptions of independent and equally variable errors were not needed. Thus, Cronbach et al (1963) were able to derive coefficient alpha as equaling the reliability of a composite, as given by Equation (1), without the usual assumptions of independent errors or homogeneous variances.

In summary, ρ_{α} equals the reliability of a test, as given by Equation (1), when items are fixed and essentially tau-equivalent, or when items are random (whether or not items are equivalent to each other in any way).

Generalizability Theory

Coefficient alpha is commonly interpreted in terms of Generalizability Theory (Cronbach et al., 1963). Thus, before discussing how coefficient alpha can be interpreted, let us begin with a brief introduction to the spirit of this theory. According to Cronbach et al. (1963), "[a]n investigator asks about the precision or reliability of a measure because he wishes to generalize from the observation in hand to some class of observations to which it belongs" (p. 144). Cronbach et al. (1963) referred to the class of observations to which one wishes to generalize as the universe of generalization. "Since a given measure may reasonably be generalized to many different universes, the investigator must specify the universe which is of interest to him before he can begin to study generalizability" (Cronbach et al., 1963, p. 144). Secondly, any measurement is taken under a specific combination of conditions, where "*Conditions* is a general term referring to particular test forms or stimuli, observers, occasions or situations of observation, etc." (Cronbach et al., 1963, p. 145). For example, when an intelligence test is administered, a particular person administers the test, at a particular time of day and time of year, in a particular location, using a particular set of instructions, and a particular form of the test. Any of these sources of error could be varied so that its influence on test scores could be assessed. For example, different test administrators could be used, to assess how much influence they have on test scores.

The sources of error which are studied depend on the universe of generalizability that interests the researcher. Generalizability Theory encourages researchers to think about the purposes of the measurement, and sources of error which will not be held constant during typical use of the instrument. This could lead the researcher to study sources of error that are not traditional reliability variables. For example, 10 randomly-selected supervisors for each of five different supervisory styles could be asked to give instructions and advice to a subordinate completing a particular task, and rate their effectiveness. Generalizability Theory could be used to assess the influence of different supervisory styles on effectiveness ratings. Such a study is as much an assessment of the reliability of effectiveness ratings, as it is a test of the influence of supervisory styles on job effectiveness.

Lastly, Generalizability Theory could be used to assess the effects of any number of sources of error (and their interactions) in a single study, unlike the traditional approach to reliability which considers each source of reliability separately. The present thesis has focused on studies involving only a single source of error.

Cronbach et al. (1963) dealt with two cases: (1) *matched* data, and (2) *unmatched* data. When data are matched, each person in the sample is observed on each condition studied. When data are unmatched, conditions are selected separately for each individual studied. In the unmatched case, with the exception of chance duplications, each person is observed under different conditions. Coefficient alpha corresponds to the matched case, and hence only the matched case will be considered here.

In classical derivations of coefficient alpha, only the reliability of the sum of k measurements is of interest. However, in Cronbach et al.'s derivation, several different quantities might be of interest. The researcher might be interested in the reliability of the mean or sum of k measurements, and for that purpose would use one of the usual formulas for coefficient alpha. Alternatively, the researcher might be interested in the reliability of a composite of k' measurements, when a set of k measurements was used in the original experiment. In this case, formula (29) from Cronbach et al. (1963) would be used. As well, the researcher could estimate the reliability of a single measure, as

$$r_{\alpha(1)} = \frac{MS_{persons} - MS_{error}}{MS_{persons} + (k-1)\, MS_{error}}. \tag{4}$$

$r_{\alpha(1)}$ is an estimate of the ratio of the expected true score variance to the expected observed score variance on a single condition, as

$$\rho_{\alpha(1)} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}.$$

Equation (4) can be used to estimate the reliability of a single condition in the universe, but can be used to estimate coefficients of generalizability as well. The coefficient of generalizability for some particular condition, Condition i, written $\rho^2_{M_i}$, gives the proportion of variance in scores on Condition i that is linearly predictable from the mean scores of people over all conditions in the universe. This can be estimated directly, using formula (19) from Cronbach et al. (1963); however, unless the number of conditions studied is large, or the variance-covariance matrix is of unit-rank, estimates obtained using formula (19) are subject to extreme sampling fluctuations. Because of this, Cronbach et al. (1963) recommend that $r_{\alpha(1)}$, an estimate of the reliability of a single condition in the universe, be used to estimate the coefficient of generalizability for any particular condition, Condition i. Equation (4) also provides a lower bound (equal when conditions are essentially parallel) to $E\rho^2_{M_i}$, the expected value of the coefficient of generalizability for each condition in the universe. According to Cronbach et al. (1963), $E\rho^2_{M_i}$ represents the coherence of a domain.

Generalizability Theory provides a very flexible and hence very powerful approach to reliability. For example, raters are rarely parallel. Their scores usually have different means and variances, and different intercorrelations with external variables. Different raters respond differently to the same aspects of the situation, and often pay attention to different environmental cues. However, if raters are randomly sampled from some specified group to which one would like to generalize, then no equivalence assumptions are required. One need not assume conditions are essentially parallel or essentially tau-equivalent. If three raters are studied, then the following reliabilities can be calculated: (a) the reliability of the ratings of each of these particular raters, (b) the reliability of the ratings of *any* single rater--not yet designated, (c) the reliability of the mean or sum of ratings by these or any other three raters, and (d) the reliability of the mean or sum of any specified number of ratings. The only requirement is that new raters are sampled from the same universe of generalization as the original raters.

The reader is reminded, however, that the power and flexibility of Generalizability Theory rests on the assumption of random sampling of conditions. If a researcher insists on viewing a set of conditions as fixed, then Cronbach et al.'s derivation cannot be used, and many of the above quantities can not be calculated.

Interpretation of Coefficient Alpha

Because coefficient alpha is usually interpreted in the context of Generalizability Theory, researchers usually assume that conditions are randomly sampled from the universe of admissible conditions, and subjects are randomly sampled from the population to which generalization is desired. In this context, the parameter value of coefficient alpha, ρ_α, is defined as the reliability of a test, as given by Equation (1). It is the expected true score variance divided by the expected observed score variance, for the mean or sum of scores on k conditions. Cronbach et al. (1963) showed that ρ_α is an upper bound on the proportion of observed score variance which is due to the first principle component underlying the population of conditions.

There are cases, however, when a fixed interpretation of conditions is more appropriate. In that case, ρ_α has be defined as

$$\rho_\alpha = \left(\frac{k}{k-1}\right)\frac{SD_{tot}^2 - \sum_i^k (SD_i^2)}{SD_{tot}^2},$$

where SD_{tot}^2 is the variance of the scores on the total test, calculated for the entire population of people, SD_i^2 is the variance on a particular item, Item i, calculated for the entire population of people, and k is the number of items (Novick & Lewis, 1967). In this case, ρ_α is equal to the reliability of a composite, as given by Equation (1) only when conditions are essentially tau-equivalent. Otherwise, ρ_α is a lower bound to the reliability of a composite.

Sample values of coefficient alpha, r_α, are the values of coefficient alpha for *n* people and *k* conditions, and can be calculated using the formulas given by Hoyt (1941), Cronbach (1951), or Kuder and Richardson (1937). In the context of randomly selected conditions, r_α estimates the proportion of total score variance on *any randomly parallel* composite of *k* measures that is due to true differences between people, where two sets of measures are considered to be randomly parallel if they were randomly sampled from the same universe of conditions. In the context of fixed conditions, r_α estimates the proportion of total score variance on the mean or sum of these particular *k* measures that is due to true differences between people.

Regardless of whether fixed or random conditions are being used, sample values of coefficient alpha are biased estimates of the population parameter. Two sources of bias exist (Cronbach et al., 1963). The first source of bias is the random sampling of people, and will affect the estimation of ρ_α whether conditions are fixed or random. The second source of bias is the random sampling of conditions. Thus, sample estimates of ρ_α will on average be slightly more biased under random sampling of conditions than when conditions are fixed, because two sources of bias are acting.

Inferential Procedures for Coefficient Alpha

Several different inferential procedures for coefficient alpha are available. The procedures use sample values of coefficient alpha, r_α, based on *k* conditions and *n* subjects, to make conclusions about the population values of coefficient alpha, ρ_α, as defined above for fixed or random conditions. As explained above, point estimation of ρ_α does not require equivalence assumptions about the conditions, if those conditions are randomly sampled from the universe of generalization. However, this non-reliance on equivalence assumptions does not apply to the various inferential procedures developed for coefficient alpha. These procedures do require some assumptions about the equivalence of conditions, regardless of whether conditions are viewed as random or fixed.

Kristof (1963) derived the sampling distribution of coefficient alpha, under the assumption of essentially parallel conditions. First, he showed that coefficient alpha is the maximum likelihood estimator of the reliability of a test when variances are assumed to be equal, means are possibly unequal, and errors are independent (i.e., conditions are essentially parallel). These assumptions imply that the variance-covariance matrix among conditions is compound symmetric (has equal variances and equal covariances). Then, assuming multivariate normality, he showed that

$$\frac{1-\rho_\alpha}{1-r_\alpha} \sim F_{(n-1),(n-1)(k-1)},$$

where ρ_α is the population value of coefficient alpha, r_α is the corresponding sample statistic. Using this, he developed a test of the hypothesis that $\rho_\alpha = a$.

Other inferential procedures for coefficient alpha followed. Feldt (1965) and Hakstian and Whalen (1976) developed confidence interval methods for coefficient alpha. There is a test of the equality of two alpha coefficients from independent samples (Feldt, 1969), and another one for matched samples (Feldt, 1980; Kristof, 1964). Related tests are also available for *k* independent alpha coefficients (Hakstian & Whalen, 1976), or *k* dependent alpha coefficients (Werts, Grandy, & Schabacher, 1980; Woodruff & Feldt, 1986). These procedures are based on the sampling distribution of coefficient alpha derived by Kristof (1963), and hence use the same assumptions as were used during the derivation of the sampling distribution. However, the necessity of using the precise assumptions given by

Kristof (1963) has been questioned (e.g., Eom, 1993). These inferential procedures clearly require some assumptions about the data; however, the precise nature of these assumptions is a point of debate in the literature at this time, and will be discussed further in Study 2.

The present thesis is concerned primarily with confidence interval procedures for coefficient alpha, and these will, therefore, be discussed in greater detail. Feldt (1965) developed the first confidence interval method to be examined in this study. He showed that, for essentially parallel conditions, an exact 100 (1 - α) % confidence interval for the parameter value of coefficient alpha, ρ_{α}, is given by:

$$\left[1-(1-r_{\alpha})F_{b}, 1-(1-r_{\alpha})F_{a}\right],$$

where F_a and F_b are respectively the $100(1\text{-}\alpha/2)$ and $100(\alpha/2)$ percentile points of the central F-distribution with (n - 1) and (n - 1)(k - 1) degrees of freedom. Feldt's proof is given in Appendix A. The difficulty with using these intervals is that they require two-tailed critical F values, which often are not available in standard statistical textbooks.

The second confidence interval method to be examined is based on asymptotic theory and hence is not exact, but it has the advantage of using critical values from the normal distribution, instead of the F-distribution, making it easier to apply. Hakstian and Whalen (1976) showed that:

$$P\{1-c^{*3}[(1-r_{\alpha})^{1/3}+Z_{1-\alpha/2}\sigma]^{3}<\rho_{\alpha}<1-c^{*3}[(1-r_{\alpha})^{1/3}-Z_{1-\alpha/2}\sigma]^{3}\}\cong 1-\alpha$$

where

$$\sigma=\frac{18k(n-1)(1-r_{\alpha})^{2/3}}{(k-1)(9n-11)},$$

$c^* = \frac{(9n-11)(k-1)}{9(n-1)(k-1)-2}$, and

$z_{1-\alpha/2}$ is the 100(1 - $\alpha/2$) percentile point of the standard normal distribution.

The derivation of this confidence interval method is given in Appendix B. Hakstian and Whalen (1976) imply that they assumed that conditions were essentially parallel.

Although Feldt (1965) and Hakstian and Whalen (1976) both assumed that conditions were being randomly sampled, Bay (1973) re-derived the Feldt confidence interval method without the assumption of random conditions. The assumption of random conditions does not appear to be necessary in the derivation of the Hakstian and Whalen confidence intervals, either, as long as conditions are essentially parallel. This is because conditions differ only in mean values, which are not utilized in the calculation of coefficient alpha, and hence fixed and random sampling of conditions are equivalent. Hence, the derivations of these confidence intervals are valid for either random or fixed conditions.

Purpose of this Thesis

As with any other statistical procedure, it is important to determine how accurate the confidence interval methods developed by Feldt (1965) and Hakstian and Whalen (1976) are with real-world data, where the assumptions made during the derivations may not be met. Thus, the purpose of this thesis is to examine the assumptions underlying these two confidence interval methods, and to investigate the robustness of the two methods to violations of their assumptions.

This thesis will focus on the random condition case, since traditional sources of error in reliability (e.g., items, forms, raters, judges, and times) are easily viewed as random effects, and generalization to underlying constructs or populations is usually desired. Furthermore, fixed conditions over which composites would be derived seem, to this author, to be rare in studies of single sources of error (where coefficient alpha is applied). Many fixed effects (sex, marital status, age) consist of mutually exclusive categories, and would not lead to the calculation of intraclass correlations. Other fixed facets (drug type, school in a district) are more likely to lead to analysis of condition effects than intraclass correlations. This author did find one example of a fixed effect which might be analyzed with an intraclass correlation--content areas of subtests (Brennan, 1983)--but fixed effects which are formed into composites seem unlikely to occur in studies of single sources of error. Thus, fixed conditions were not examined in this thesis. The assumption of multivariate normality was made throughout.

In Study 1, a preliminary study dedicated to sampling procedures is presented. This thesis, and much other research in this area, is based on a random-sampling interpretation of coefficient alpha. However, fixed conditions have been commonly used by researchers studying coefficient alpha and the performance of confidence intervals for coefficient alpha. Therefore, the question arises as to whether these two methods of sampling would lead to the same conclusions. If these two types of sampling do produce very similar results, then the type of sampling used in simulation studies is not very important. However, if they do not produce very similar results, then a researcher interested in the performance of the confidence intervals (or some other inferential procedure for coefficient alpha) under a particular type of sampling, should be sure to use that type of sampling in his or her simulations. Therefore, the first study presented in this thesis was a comparison of the performance of the confidence interval methods when conditions were fixed with their performance when conditions were random. When conditions are fixed, their statistical qualities need to be generated only once, and hence fixed conditions are easier to simulate than random conditions. The results of this first study were used to determine whether fixed or random conditions were simulated in later studies. This study will be of interest to other researchers working with coefficient alpha and its sampling properties, as they may wish to use fixed conditions in their simulations even though the random-sampling interpretation of coefficient alpha is the primary interest.

Study 2 was designed to eliminate some confusion about the assumptions underlying the confidence interval methods. Because of the relationship between coefficient alpha and repeated measures ANOVA, there is some confusion as to whether *sphericity*, a well-known ANOVA assumption, is sufficient, or whether the stronger assumption of *compound symmetry* (implied by essential parallelism of conditions) is required. A theoretical examination of the claim that sphericity is sufficient was performed first, followed by an empirical examination of the performance of the confidence intervals under sphericity and compound symmetry. As with the first study, this study will be of interest to other researchers working with coefficient alpha and its sampling properties.

In the third study, the performance of the confidence intervals under four different well-known measurement models was examined. As well, the effects of five variables on these confidence intervals was explored. This study will interest practitioners who wish to use the confidence interval procedures to inform them about the internal consistency of their measurements.

STUDY 1:
AN ISSUE CONCERNING MONTE CARLO STUDIES OF COEFFICIENT ALPHA: SAMPLING

Introduction

This first study is devoted to an issue of interest to psychometricians doing Monte Carlo studies in the area of reliability: does it matter what type of sampling is used when simulating data? Lord (1955) distinguished between three types of sampling, depending on whether subjects, conditions, or both were random. When subjects are random, a sample of subjects is randomly drawn from some larger population, and results are expected to generalize to that population. When they are fixed, these are the only subjects of interest, and generalization to a larger population is not desired. When conditions are random, the conditions used in the study are presumed to be a (more or less) random sample from some larger domain, and results are expected to generalize to that larger domain. Thus, conclusions about the construct *underlying* the conditions used are possible. When conditions are fixed, the conditions used in the study are the only ones of interest. Results are not intended to be related to any construct that might be presumed to underlie the conditions used.

Lord (1955) defined Type 1 sampling as the case when subjects are randomly sampled from their population, but conditions are fixed. Thus, if there existed a test which consisted of fixed items (and generalization to some larger construct was not intended), and this test was administered to several different randomly selected samples of subjects, this would be an example of Type 1 sampling. With Type 2 sampling, conditions are randomly sampled, but subjects are fixed. Thus, if different self-esteem tests were considered to be different independent random samples of conditions from the domain (or universe) of self-esteem, and these tests were administered to one particular set of subjects (and generalization to other people was not desired), then this would be an example of Type 2 sampling. Finally, the third type of sampling, Type 12 sampling, occurs when both subjects and conditions are randomly sampled. Thus, if different self-esteem tests were given to different randomly selected groups of subjects, then this would be an example of Type 12 sampling. Lord (1955) has shown that the Type 12 sampling variance for any statistic is approximately equal to the sum of the Type 1 and Type 2 sampling variances.

When Kristof (1963) developed the sampling distribution of coefficient alpha and when Feldt (1965) and Hakstian and Whalen (1976) developed their confidence interval procedures, they assumed Type 12 sampling, with both conditions and subjects random. This implies that they were interested in the random-sampling interpretations of coefficient alpha. However, modelling Type 12 sampling is more complicated than modelling Type 1 sampling. With Type 1 sampling only a single set of *k* conditions with the desired characteristics needs to be generated. With Type 12 sampling, a program must generate multiple samples of *k* conditions that represent independent random samples from a population with the desired characteristics.

Because coefficient alpha has a different sampling variance under Type 12 sampling than under Type 1 sampling, the performance of the confidence intervals will likely depend on the type of sampling used, unless conditions are parallel. When conditions are parallel, they have equal means, equal variances, and uncorrelated errors. In this situation, it does not matter whether conditions are held constant or are randomly sampled. In fact, because coefficient alpha is unaffected by differences in condition means, if conditions are essentially parallel (equal variances and uncorrelated errors), then Type 1 and Type 12 sampling variance of coefficient alpha will be identical.

Even if conditions are not essentially parallel, for a moderate number of conditions, the Type 2 sampling variance of coefficient alpha is quite small: Lord showed that the Type 2 sampling variance of coefficient alpha is on the order of $\frac{1}{k^3}$, where k is the number of conditions sampled. Because of this, Feldt (1965) and Hakstian and Whalen (1976) suggested that with moderate to large numbers of conditions, the differences between Type 1 sampling and Type 12 sampling of coefficient alpha will be small. It is thus common to see Type 1 sampling used in studies attempting to examine the inferential properties of coefficient alpha under Type 12 sampling (e.g., Feldt, 1965, 1969; Hakstian & Whalen, 1976; Sedere & Feldt, 1976; Woodruff & Feldt, 1986). None of these researchers has demonstrated that the confidence intervals perform similarly under the two types of sampling, when conditions are not essentially parallel. The purpose of this study was to determine whether the proportion of the confidence intervals that include the population value of coefficient alpha is very similar under Type 1 and Type 12 sampling, as often claimed.

Method

A Monte Carlo simulation experiment was conducted to compare the actual proportion of the .95 confidence intervals that include the population value of coefficient alpha when two sampling methods are used: Type 1 (conditions fixed, subjects random) and Type 12 (conditions and subjects both random). If the confidence intervals are providing precise interval estimation, 95% of the .95 confidence intervals will include the population value.

A double-precision FORTRAN program was written for this and subsequent experiments. The Choleski decomposition, random number generation, and sample selection subroutines were written by Dr. James Steiger, University of British Columbia, who generously lent them to the author for use on this thesis. All other subroutines were written by the author.

Kinds of Data Studied

This was a preliminary study. Its purpose was to determine whether Type 1 sampling produces results that are indistinguishable from those obtained with Type 12 sampling, for the kinds of data to be studied in the rest of this thesis. In Study 2, two kinds of data were examined: compound symmetric data with unequal condition means; and spherical data that were not compound symmetric, with unequal condition means. In Study 3, four kinds of data were studied, corresponding to four measurement models: parallel, essentially parallel, tau-equivalent, and essentially tau-equivalent.

For this first study, two kinds of data were selected: spherical (but not compound symmetric) data with unequal variances and means, like that used in Study 2, and data based on the tau-equivalent measurement model, which is like that used in Study 3. Data are spherical if in the population, the covariance between any two conditions is precisely the average of the variances of those two conditions minus some constant, where this constant is the same for all pairs of conditions. Sphericity appears to correspond to no known measurement model. However, the assumption of sphericity is important in repeated measures ANOVA, and because of the relationship between coefficient alpha and the F-test on people in repeated-measures ANOVA, some researchers (e.g., Eom, 1993) have claimed that sphericity is the assumption underlying the sampling distribution of coefficient alpha. Tau-equivalence, on the other hand, corresponds to a well-known measurement model. The assumption of tau-equivalence states that true scores are constant across all conditions. Thus, data which are tau-equivalent will have equal condition means, but because no assumptions are made about error variances, condition variances may be unequal. For this study, data with unequal variances were used.

Data were generated to maximize the observed differences between Type 1 and Type 12 sampling. As stated above, the Type 12 sampling variance of any statistic is approximately the sum of Type 1 sampling variance and Type 2 sampling variance (Lord, 1955). Thus, Type 1 and Type 12 sampling variance are most disparate when Type 1 sampling variance is low, and when Type 2 sampling variance is high. Type 1 sampling variance of coefficient alpha should be low when the sample size is large, and when the population value of coefficient alpha is high (Hakstian & Whalen, 1976). Type 2 sampling variance of coefficient alpha should be high when there are few conditions, and when these conditions are far from being essentially parallel (i.e., have heterogeneous variances, and perhaps correlated errors).

To determine that the computer program was working properly, a third kind of data which should produce accurate confidence intervals--parallel data--was simulated. Conditions had the same characteristics--means, variances, distributions, inter-correlations--and hence it did not matter if new conditions were selected for each sample, or if the same conditions were used throughout. Thus, only one type of sampling (Type 12 sampling) was used for the cells used to test the computer program. These data provided a good test of the computer program because they met all of the assumptions made by Feldt (1965) and Kristof (1963) during the development of the sampling distribution and confidence interval methods for coefficient alpha. These data also met the additional assumption of equal condition means.

Experimental Design

Combinations of levels of variables will be referred to as *cells*. The word *condition* will be reserved for reference to “particular test forms or stimuli, observers, occasions or situations of observation, etc" (Cronbach et al., 1963, p. 145).

Spherical (but not Compound Symmetric) and Tau-Equivalent Data

Eight cells each were simulated for the spherical and tau-equivalent data. These consisted of each combination of number of conditions (five or 20), population value of coefficient alpha (.60 or .90) and type of sampling (Type 1 or Type 12). The numbers of conditions were selected in large part because of specific data generation challenges to be discussed under *Data Generation.* The 5-condition cells were expected to evidence greater differences between Type 1 and Type 12 sampling, than were the 20-condition cells. Both high and low values of coefficient alpha were used because Type 1 sampling variance of coefficient alpha becomes smaller for more extreme values of coefficient alpha, but so too does Type 2 sampling variance.

Sample size (*n*) was set to 100, which represents a moderate sample size for a reliability experiment. Condition means were equal for the tau-equivalent data, and set to 0. For the spherical data, condition means were unequal, with a mean of 0 and a standard deviation of 20. In both cases, the means were normally distributed. Subject means were normally distributed with a mean of 0. For the spherical data, the standard deviation of subject means was not specified, as subject effects and errors were generated at the same time (see *Data Generation* below). For the tau-equivalent data, the variance due to subjects is regarded as "true score variance" and therefore was determined by the population value of coefficient alpha.

Condition variances had a mean of 100 and standard deviations (*SDVar*) of 15 for the tau-equivalent data and 10 for the spherical data. The data generation problems that lead to the selection of these values for *SDVar* will be discussed further in Data Generation. When *SDVar* was 15, most (about two-thirds of) condition variances were between 85 and 115 and almost all were between 70 and 130.

When variances were unequal, they were normally distributed in the population. The selection of the normal distribution requires further comment. If sample variances were being generated from one particular population variance, then these sample values would have a chi-square distribution. However, it was not sample variances that were being generated here, but rather population values. These population values could have had any distribution desired, and a normal distribution was selected. A chi-square distribution was rejected because of the higher proportion of high values: it appears to be the extreme values that cause problems with data generation.

Parallel Data

Four cells of parallel data were used as a check that the computer program was working properly. The levels of variables used for the parallel data are presented in Table 1 beside the levels of variables used for the spherical and tau-equivalent data.

Feldt (1965) and Hakstian and Whalen (1976) argued that the differences between Type 1 and Type 12 sampling are small; therefore, 20,000 replications were used for each of the 16 cells in the main design. For the parallel data used to test the computer program, 5,000 replications per cell were used. Both types of confidence intervals were calculated for each sample.

Data Generation

Data generation was carried out differently for the spherical data, as opposed to the tau-equivalent and parallel data. The generation of spherical data is considered first, followed by consideration of the generation of tau-equivalent data (and the parallel data used to check the program).

Table 1

Combinations of Variables and Levels of Variables Used in Study 1

Independent Variable	Spherical	Tau-Equivalent	Parallel
Average condition variance (*AveVar*)	100	100	100
SD of condition variances (*SDVar*)	10	15	0
SD of condition means (*SDMean*)	20	0	0
Number of conditions (*k*)	5, 20	5, 20	5, 20
Sample size (*n*)	100	100	100
Coefficient alpha (ρ_α)	.6, .9	.6, .9	.6, .9
Sampling	Type 1, Type 12	Type 1, Type 12	Type 12

Spherical Data

Generation of spherical data with unequal means and variances was broken into three steps. First, a spherical population variance-covariance matrix was randomly generated. This was then used to produce observations. Finally, differences in condition means were taken into account by adding randomly-generated condition effects to these observations. For cells using Type 12 sampling, all steps were repeated for each sample. For cells using Type 1 sampling, the population variance-covariance matrix and condition means were generated only once. Samples were generated by using that population variance-covariance matrix to generate observations, and then adding in the condition effects.

Generation of the population variance-covariance matrix.

The first step was to generate a spherical variance-covariance matrix, consisting of the population values of variances and covariances of *k* randomly-selected conditions. This was done by randomly generating variances and then calculating the necessary covariances so that the overall variance-covariance matrix was perfectly spherical and had the desired value of coefficient alpha.

To generate the variances, independent random normal variates with mean 0 and standard deviation 1 were generated. These were then multiplied by the desired standard deviation (*SDVar*) and added to the desired mean (*AveVar*). For example, for a 5-condition cell, with *AveVar* = 100, *SDVar* = 10, the randomly-generated variances might be 108.93, 92.98, 91.06, 113.22, and 96.05.

For the cells of the design using Type 1 sampling, the mean and standard deviation of these variances were then calculated to ensure that they were within .5% of the desired values. The randomly selected variances in the example above have mean 100.49 (the sample value of *AveVar*) and standard deviation 9.979 (the sample value of *SDVar*), and would be considered acceptable. If the randomly selected variances did not have close to the desired mean and standard deviation when Type 1 sampling was being used, then data generation was started again from the beginning and new variances were generated. This process was repeated until variances with the desired characteristics were generated.

Next, covariances were calculated which would produce a perfectly spherical matrix with the desired value of coefficient alpha. The covariances were given by

$$Cov(i,j) = \frac{Var(i)+Var(j)}{2} - c, \tag{5}$$

where c is $AveVar - \frac{\rho_\alpha AveVar}{k + \rho_\alpha - \rho_\alpha k}$, $AveVar$ is the average variance in the population, k is the number of conditions, and ρ_α is the population value of coefficient alpha. In our example, if the desired value of coefficient alpha is .60, then the spherical matrix which results is

$$\mathbf{V} = \begin{bmatrix} 108.93 & 24.03 & 23.07 & 34.15 & 25.57 \\ 24.03 & 92.98 & 15.10 & 26.18 & 17.59 \\ 23.07 & 15.10 & 91.06 & 25.22 & 16.63 \\ 34.15 & 26.18 & 25.22 & 113.22 & 27.71 \\ 25.57 & 17.59 & 16.63 & 27.71 & 96.05 \end{bmatrix}$$

Sometimes, the matrix which resulted was not positive-definite, and hence could not be used as population variance-covariance matrix. When this happened, the data generation process returned to the beginning.

If Type 1 sampling was being used, this matrix was examined to determine that the associated value of coefficient alpha was close to the desired value. It could differ slightly from the desired value because *AveVar* was used in the formulas leading to the calculation of the covariances, rather than the actual average of the randomly selected variances. If coefficient alpha was not within 1% of the desired value, then new variances were generated. Thus, if the population value of coefficient alpha was .60, acceptable values were those between [.594, .606]. For a population value of .90, acceptable values were those between [.891, .909]. The actual value of coefficient alpha for the example is .6046, and would have been considered acceptable.

Generation of observations.

This variance-covariance matrix, $\mathbf{V}$, was used to generate observations. First, the matrix $\mathbf{V}$ was decomposed using a Choleski decomposition, such that $\mathbf{V} = \mathbf{FF}'$. Then, nxk multivariate normal observations (where n is the number of subjects and k the number of conditions) were randomly generated with mean 0 and standard deviation 1, and post-multiplied by $\mathbf{F}'$. The new observations had the desired population variance-covariance matrix.

Proof:

Let $\mathbf{X}$ be a population of independent multivariate normal variables. Then, $E[\mathbf{X}'\mathbf{X}] = \mathbf{I}$. Let $\mathbf{X}_k$ be a randomly selected sample of k of these variables. Let $\mathbf{V}$ be the desired spherical variance-covariance matrix, and $\mathbf{V} = \mathbf{FF}'$. Take $\mathbf{Y} = \mathbf{X}_k\mathbf{F}'$. Then, $E[\mathbf{Y}'\mathbf{Y}]$ $= E[[\mathbf{X}_k\mathbf{F}']'[\mathbf{X}_k\mathbf{F}']]$ $= E[\mathbf{FX}_k'\mathbf{X}_k\mathbf{F}']$ $= \mathbf{F}E[\mathbf{X}_k'\mathbf{X}_k]\mathbf{F}'$ $= \mathbf{FIF}' = \mathbf{FF}' = \mathbf{V}$.

These observations represent the sum of the effects due to true differences between people, and the effects of random errors. The variance-covariance matrix used, $\mathbf{V}$, can be viewed as the sum of the variance-covariance matrix due to people and the matrix due to errors. The variances were the sums of the variance due to people and the variance due to errors. Because true scores are constant across conditions, the covariance between true scores on different conditions equals the variance due to people, and hence the covariances between observations were the sums of the variance due to people and the covariance between errors. Generating people effects and errors separately would have been possible. However, because people effects and errors are independent and additive, nothing is lost by generating them at the same time, based on their combined variances and covariances. Therefore, this approach, which is simpler and faster, was used.

Generation of condition effects.

Finally, differences in condition means were taken into account by adding randomly-generated condition effects to the observations just generated. First, independent random normal variates with mean 0 and standard deviation 1 were generated. These were multiplied by the desired standard deviation (*SDMean*). These condition effects were then added to the observations. Thus, if the randomly-generated condition effects were 13.32, 25.19, -19.49, -20.51, and .04, then 13.32 would have been added to the observations in Condition 1 for all subjects, 25.19 would have been added to the observations in Condition 2, -19.49 for Condition 3, and so forth.

For cells using Type 1 sampling, these condition effects were examined to ensure that their mean and standard deviation were within .5% of the desired values. The condition effects given above had mean -.29, and standard deviation 20.07, which are within .5% of the desired values of 0 and 20. These condition effects would have been judged acceptable. If they had not been, new condition effects would have been generated.

Problems encountered in data generation.

When Type 12 sampling was used with data with unequal variances, some of the matrices were unacceptable and were discarded, as discussed above. If this had happened frequently, the samples might have been considerably biased, making it impossible to determine the cause of any apparent differences between cells. To make interpretation clear, 100% (or nearly 100%) of the randomly-generated variance-covariance matrices should be acceptable. This was therefore controlled by careful selection of the number of conditions and the degree of heterogeneity of the condition variances. Preliminary studies suggested that with the standard deviation of the variances (*SDVar*) equal to 10, very few, if any, of these matrices would be unacceptable for 5- and 20-condition cells, and hence these values were used. With higher values of *SDVar*, this was unlikely. See the results of the preliminary studies in Appendix C.

Tau-Equivalent Data (and Parallel Data)

With tau-equivalent data with unequal condition variances, the data generation process was broken into four steps: generation of conditions, generation of subjects, generation of errors, and addition of these components. For cells using Type 12 sampling, all four steps were repeated for each sample value of coefficient alpha. For cells using Type 1 sampling, condition means and variances were generated only once. Only the last three steps were repeated: samples were obtained by randomly generating subjects and errors (whose variances were fixed) and adding these to the fixed condition means.

The parallel data used to test the computer program was generated using the same process as the tau-equivalent data, with one slight variation, discussed below.

Generation of conditions.

With tau-equivalent data, true scores are assumed to be constant across conditions. However, error variances may be unequal, and in the data studied here, they were unequal. To generate error variances, the desired mean and standard deviation of these variances are needed. The desired mean (*AveErr)* is the remainder after the variance due to true differences between people (*TVar*) has been subtracted from the average total variance: $AveErr = AveVar - TVar$, where

$$TVar = \frac{\rho_\alpha AveVar}{k + \rho_\alpha - \rho_\alpha k}, \tag{6}$$

AveVar is the average condition variance, k is the number of conditions, and ρ_α is the population value of alpha. The desired standard deviation of the error variances is simply the desired standard deviation of the total variances (*SDVar*). To generate error variances, first, independent random normal variates with mean 0 and standard deviation 1 were generated. These were multiplied by the desired standard deviation (*SDVar*), and added to the desired mean (*AveErr*). Because variances are non-negative, if any negative values were randomly generated at this stage, then a new set of variances was generated.

For the cells using Type 1 sampling, the error variances were examined to ensure that their mean and standard deviation were within .5% of the desired values. If they were not, then new error variances were generated. The population value of coefficient alpha was also examined. Unless it was within 1% of the desired value, new error variances were generated.

For the parallel data used to test the computer program, error variances were equal, and were set to *AveErr*.

Generation of subjects.

To randomly generate the true scores of subjects, independent random normal variates were generated with a mean of 0 and a standard deviation of 1. These were multiplied by the desired standard deviation (the square root of *TVar*, as given in Equation (6)).

Generation of errors.

Independent random normal variates (*nk* of them) were generated with a mean of 0 and a standard deviation of 1. These were multiplied by the desired standard deviation: the square root of the error variance for that condition, which was randomly generated in the first step.

Addition of components.

The three components were then added together: $X_{ij} = \rho_i + \tau_j + \varepsilon_{ij}$, where X_{ij} is the observed score of Person *i* on Condition *j*, ρ_i is the person effect for Person *i*, τ_j is the condition effect for Condition *j*, and ε_{ij} is the residual or error component.

Problems encountered in data generation.

For cells using Type 12 sampling, there were some data generation problems. When the number of conditions was small, and the population value of coefficient alpha was large, the average error variance was small. In such a case, if the error variances were made to be quite heterogeneous, some of the randomly-generated numbers were negative. When this occurred, new variances were generated. If this had occurred frequently, samples might have been considerably biased. Preliminary studies were used to select numbers of conditions and values of *SDVar* for which this would occur infrequently (less than 5% of the time). See Appendix D for results from the preliminary studies.

Data Analysis

For each sample, the Hakstian and Whalen (1976) and Feldt (1965) confidence intervals were calculated. These confidence intervals were compared to the population value of coefficient alpha, to determine whether they included it. In the case of Type 12 sampling, the population value of coefficient alpha was defined as its desired value (.60 or .90). In the case of Type 1 sampling, the actual value of coefficient alpha in the randomly selected fixed population of conditions was used. Summaries of the numbers and proportions of confidence intervals that included the population value were made for each cell in the design. These were compared across cells using statistical tests for equality of proportions.

Results

In this section, the results of the Monte Carlo study are presented. The results with the parallel data are presented first, to demonstrate the accuracy of the computer program. These are followed by the results for the spherical data, and finally the results for the tau-equivalent data.

Parallel Data

Parallel data were simulated to test the accuracy of the computer program. These data should result in precise confidence intervals. As can be seen from Table 2, the four cells run did produce accurate confidence intervals. In none of these cells did the proportion of the confidence intervals that included the population value differ significantly from nominal .95.

Table 2

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Parallel Data

Conditions	Alpha	H & W CI	Feldt CI
5	.60	.9504	.9506
5	.90	.9466	.9446
20	.60	.9462	.9456
20	.90	.9480	.9480

Note. Conditions = the number of conditions; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 5,000 replications per cell.

Spherical Data

Before testing for differences between Type 1 and Type 12 sampling cells, several data checks were done to ensure that the data were generated as intended and that no confounds were introduced.

Data Checks

Matrices that were not positive definite.

As discussed under *Problems encountered in data generation*, it is possible to randomly generate spherical matrices that are not positive definite and hence must be discarded. If this had happened frequently, then Type 12 sampling results might have been be considerably biased. In three out of the four cells, this problem *never* occurred. In the one cell in which matrices occasionally were not positive definite ($k = 20$, $\rho_\alpha = .60$), this problem occurred only 1.52% of the time, and thus it is unlikely that this phenomenon caused any differences in the results of the Type 1 and Type 12 sampling cells.

Levels for Type 1 sampling.

For cells using Type 1 sampling, it is important that the fixed population conditions had close to the desired characteristics. Variances should have had close to their desired mean and variance (*AveVar* and *VarVar,* which is the square of *SDVar*). Condition means should have had close to their desired mean and variance (0 and *VarMean*, which is the square of *SDMean*). Finally, the population value of coefficient alpha should have been close to its desired value (ρ_α). If these values were not close, then comparisons between Type 1 and Type 12 sampling would be confounded by those differences. The actual values used were in close agreement with the desired values. See Appendix E.

Levels for Type 12 sampling.

For cells using Type 12 sampling, population characteristics of conditions were randomly selected for each sample. These population characteristics, when averaged across replications, should have been close to the desired values. Thus, the average and variance of the condition means should have been close to their desired values (0 and *VarMean*, respectively). Similarly, the average and variance of condition variances should have been close to their desired values (*AveVar* and *VarVar*, respectively). These values were close to the desired values, and hence discrepancies from desired values could not account for any differences in the results of Type 1 and Type 12 sampling cells. See Appendix F.

Sample values of coefficient alpha.

Across replications, the average of the sample values of coefficient alpha should have been close to the desired population value. If the average of the sample values of coefficient alpha were very different from the desired value of coefficient alpha, this could have affected the performance of the confidence intervals and created differences between Type 1 and Type 12 sampling cells.

From Table 3, sample values of coefficient alpha appear to be slightly negatively biased. This bias has been noted elsewhere (e.g., Kristof, 1963; Cronbach et al., 1963). This bias appears to be justly slightly larger for cells with low values of population coefficient alpha. As well, it appears to be slightly larger for Type 12 sampling than Type 1 sampling, which could create some small differences between Type 1 and Type 12 sampling results.

Table 3

Average Across Replications of the Sample Values of Coefficient Alpha, for Spherical Data

Sampling	Conditions	Population Alpha	Sample Alpha	Bias
Type 1	5	.6046	.5968	-.0078
Type 1	5	.9005	.8985	-.0020
Type 12	5	.6000	.5852	-.0148
Type 12	5	.9000	.8976	-.0024
Type 1	20	.5999	.5913	-.0086
Type 1	20	.8996	.8976	-.0020
Type 12	20	.6000	.5773	-.0227
Type 12	20	.9000	.8975	-.0025

Note. Sampling = the type of sampling; Conditions = the number of conditions; Bias = the difference between the average sample value and the desired population value.

Comparison of Type 1 and Type 12 Sampling

Table 4 presents the results of the Monte Carlo study for spherical data. The proportions of the confidence intervals which included the population value were compared across Type 1 and Type 12 sampling. As can be seen from the table, Type 1 and Type 12 sampling did produce significantly different results for spherical data, with the differences being most pronounced when the population value of coefficient alpha was low (.60), and when the number of conditions was high (20). As well, differences from nominal level were much more pronounced with Type 12 sampling than with Type 1 sampling.

These observed differences cannot be accounted for by inadequate data generation or by confounding variables. Although some differential bias of sample estimates for the two types of sampling was observed, these differences are unlikely to have created the large observed differences between cells using Type 1 and Type 12 sampling.

Tau-Equivalent Data

Before testing for differences between Type 1 and Type 12 sampling cells, several data checks were done to ensure that the data were generated as intended and that no confounds were introduced.

Data Checks

Negative variances with Type 12 sampling.

As discussed under *Problems encountered in data generation*, it is possible to randomly generate negative numbers for the error variances. Because variances are non-negative, when a negative value was generated for a variance, that set of error variances was discarded, and a new set generated. The proportion of sets of variances that included one

Table 4

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Spherical Data

Conditions	Alpha	H&W CI		Feldt CI	
		Type 1	Type 12	Type 1	Type 12
5	.60	.95165*	.88715	.95185*	.88670
5	.90	.95180*	.93520	.95265*	.93605
20	.60	.95415*	.77055	.95485*	.77050
20	.90	.95295*	.93005	.95265*	.93005

* indicates that Type 1 and Type 12 sampling produced significantly ($p < .05$) different results.

Note. Conditions = the number of conditions; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 20,000 replications per cell.

or more negative value was 0 for the cells with ρ_α = .60, .041 for the cell with ρ_α = .90 and $k = 5$, and .00005 for the cell with ρ_α = .90 and $k = 20$. Because variances were discarded so rarely, this could not account for any differences between the results of Type 1 and Type 12 sampling cells.

Levels for Type 1 sampling.

For cells using Type 1 sampling, it is important that the fixed population conditions had close to the desired characteristics. Variances should have had close to their desired mean and variance (*AveVar* and *VarVar,* which is the square of *SDVar*). Means should have had close to their desired mean and variance (0 and *VarMean*, which is the square of *SDMean*). Finally, the population value of coefficient alpha should have been close to its desired value (ρ_α). If these values were not close, then comparisons between Type 1 and Type 12 sampling would be confounded by those differences. The actual values used were in close agreement with the desired values. See Appendix G.

Levels for Type 12 sampling.

For cells using Type 12 sampling, population characteristics of the conditions were randomly selected for each sample. These population characteristics, when averaged across replications, should have been close to the desired values. Thus, the average and variance of the condition means should have been close to their desired values (0 and *VarMean*, respectively). Similarly, the average and variance of condition variances should have been close to their desired values (*AveVar* and *VarVar*, respectively). All but one of these values were close to the desired values. The average variance of the variances for the cells with $k = 5$ and ρ_α = .90 appears to have been slightly smaller than the desired value; the average value was 211.42 compared to the desired value of 225. See Appendix H.

Sample values of coefficient alpha.

Across replications, sample values of coefficient alpha should have had an average close to the desired population value. If the average of the sample values of coefficient alpha were very different from the desired value of coefficient alpha, this could have affected the performance of the confidence intervals and created differences between Type 1 and Type 12 sampling cells. From Table 5, sample values of coefficient alpha appear to be slightly negatively biased. The bias appears to have been slightly larger for the small value of coefficient alpha. However, no differences between the amount of bias for Type 1 and Type 12 sampling are observed. Hence, this bias cannot account for any differences between Type 1 and Type 12 sampling results.

Comparison of Type 1 and Type 12 Sampling

Table 6 presents the results of the Monte Carlo study for tau-equivalent data. The proportion of the confidence intervals that included the population value were compared across Type 1 and Type 12 sampling. As can be seen from the table, Type 1 and Type 12 sampling did produce significantly different results for tau-equivalent data, with the differences being most pronounced when the number of conditions was small, and when the population value of coefficient alpha was high. This is the opposite of the results obtained with spherical data. Differences from the nominal level were much more pronounced with Type 12 sampling than with Type 1 sampling, as was the case for spherical data.

Discussion and Conclusions

Type 1 sampling and Type 12 sampling often produced substantially different results, for both spherical and tau-equivalent data. Alternative explanations of the observed differences were sought, and in only two of the 14 significant comparisons was any

Table 5

Average Across Replications of the Sample Values of Coefficient Alpha, for Tau-Equivalent Data

Sampling	Conditions	Population Alpha	Sample Alpha	Bias
Type 1	5	.6008	.5933	-.0075
Type 1	5	.9000	.8978	-.0022
Type 12	5	.6000	.5926	-.0074
Type 12	5	.9000	.8972	-.0028
Type 1	20	.5997	.5918	-.0079
Type 1	20	.8996	.8976	-.0020
Type 12	20	.6000	.5918	-.0082
Type 12	20	.9000	.8982	-.0018

Note. Sampling = type of sampling; Conditions = number of conditions; Bias = the difference between the average of the sample values and the desired population value.

Table 6

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Tau-Equivalent Data

Conditions	Alpha	H&W CI		Feldt CI	
		Type 1	Type 12	Type 1	Type 12
5	.60	.94900*	.93690	.94805*	.93603
5	.90	.94865*	.82490	.94875*	.82345
20	.60	.94890	.95050	.94865	.94495
20	.90	.95185*	.93840	.95180*	.93840

* indicates that Type 1 and Type 12 sampling produced significantly ($p < .05$) different results.

Note. Conditions = the number of conditions; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 20,000 replications per cell.

alternative explanation plausible. These were the comparisons, for both types of confidence interval, between Type 1 and Type 12 sampling for tau-equivalent data with $k = 5$ and $\rho_{\alpha} = .90$. Under Type 12 sampling, the average variance of the variances was 211.42, which is noticeably smaller than the desired value of 225. In this cell 4.1% of the sets of error variances were discarded because they had at least one negative value. Assuming then that the proportions of confidence intervals that included the population value under Type 12 sampling may have been biased by as much as 4.1%, significant differences between the Type 1 and Type 12 sampling cells still exist, for both types of confidence interval.

Thus, in 14 of the 16 comparisons, significant differences, sometimes substantial ones, were found. For spherical data, these differences were greatest when the number of conditions was large, and when the population value of coefficient alpha was small. These differences may have been partly caused by the increased bias associated with Type 12 sampling. The opposite occurred when tau-equivalent data were used: differences were largest for the smaller number of conditions and for the larger value of coefficient alpha. In this case, no differences in bias were observed for the two types of sampling. Because of this interaction, no generalizations can be made about when Type 1 sampling may produce similar results to Type 12 sampling.

It is interesting to note that differences were larger for the 20-condition cells than the 5-condition cells for spherical data. Lord (1955) showed that Type 2 sampling variance of coefficient alpha is of the order of $\frac{1}{k^3}$, and hence should decrease as k increases. However, if Type 1 sampling variance also decreased, then the relative difference between Type 1 and Type 12 sampling variance may not decrease as k increases, as has been assumed.

These results lead to the conclusion that future studies of inferential procedures for coefficient alpha should use Type 1 sampling if they are interested in fixed conditions, and Type 12 sampling if they are interested in random conditions. In general, the cells using Type 1 sampling produced fairly accurate confidence intervals. In most cases, the proportion of the confidence intervals that included the population value was within sampling error of nominal levels. However, the cells using Type 12 sampling sometimes performed poorly, with as few as 77% of the .95 confidence intervals including the population value. It is unfortunate that the confidence intervals are more accurate with the less common type of sampling--Type 1 sampling. If the type of sampling is not taken into account when evaluating the performance of the confidence intervals, serious mistakes will be made in evaluating their strengths and weaknesses, and the results of different studies may appear to conflict greatly.

Because of this finding, careful attention was paid to the type of sampling used in the two remaining studies in this thesis. Since this thesis is concerned with random sampling of conditions, Type 12 sampling was used for Studies 2 and 3. The preliminary results in Appendices C and D were used to determine which combinations of numbers of conditions and *SDVar* to use to overcome the previously mentioned problems associated with data generation under Type 12 sampling.

This first study also found that, with both kinds of data, differences from nominal level were much more pronounced when Type 12 sampling was used. Thus, the past use of Type 1 sampling as a substitute for Type 12 sampling may have given researchers an erroneous impression that inferential procedures for coefficient alpha were robust to violation of some of their assumptions when Type 12 sampling is used. As discussed above, when conditions are essentially parallel, Type 1 and Type 12 sampling will produce identical results. Several studies have used conditions which are essentially parallel, and their results are applicable to either Type 1 or Type 12 sampling (e.g., Alsawalmeh & Feldt, 1992, Bay, 1973; parts of Eom, 1993; Feldt, 1980; Woodruff & Feldt, 1986). However, other studies have deliberately used conditions which were not essentially parallel. Hakstian and Whalen (1976) and Feldt (1965, 1969) used dichotomous data with unequal variances. Eom (1993) used unequal correlations among the errors for most of his cells. For these studies, the type of sampling used in the simulations becomes important, as results for the two types of sampling can be expected to differ.

STUDY 2:

AN EXAMINATION OF THEORETICAL ASSUMPTIONS

Introduction

Cronbach at al. (1963) showed that if random sampling of conditions is assumed, then coefficient alpha can be interpreted as an intraclass correlation coefficient with very few additional assumptions. However, distributional assumptions are needed to understand the sampling distribution of coefficient alpha and to develop confidence intervals for it. In the past, there has been some confusion regarding these assumptions. Feldt (1965) clearly laid out the assumptions he was using in the development of his confidence interval method, but later researchers have contradicted him. Feldt stated the following assumptions:

(i) The score of subject *i* (*i*=1, ... *N*) on item *j* (*j*=1, ... *k*) may be represented as $X_{ij} = \mu + a_j + t_i + e_{ij}$, where the notation is defined as follows.

μ = the mean condition score over the entire population of subjects and item.

a_j = the amount by which the mean examinee score on item *j* deviates from μ. The quantity a_j reflects, in deviation form, the relative difficulty of the item for the population of examinees.

Since a_j is a deviation score, $\sum_{j=1}^{\infty} a_j = 0$.

t_i = the amount by which the mean item score for examinee i deviates from μ. The quantity t_i reflects, in deviation form, the true ability of the examinee in the domain defined by the population of items. Since t_i is a deviation score, $\sum_{i=1}^{\infty} t_i = 0$.

e_{ij} = the interaction effect of item j with subject i, an effect presumed wholly the result of measurement error. For examinee i, $\sum_{i=1}^{\infty} e_{ij} = 0$; for item j, $\sum_{j=1}^{\infty} e_{ij} = 0$.

(ii) The N subjects are assumed to be a random sample from the examinee population.

(iii) The k items are assumed to be a random sample from the population of items.

(iv) Over the entire population of examinees, the quantity t_i is assumed to be normally distributed.

(v) Over the entire examinees-by-item matrix, the e_{ij} are assumed normally distributed, independently of each other and of t_i.

(vi) For any infinite subpopulation of examinees and items, σ_e^2 is assumed equal to σ_e^2 for any other subpopulation.

(Feldt, 1965, p. 359)

In the same paper, Feldt (1965) clarified the last assumption, noting that scores on each item (condition) have variance $\sigma_t^2 + \sigma_e^2$. He also identifies these as the assumptions of the two-way random effects ANOVA model. Hakstian and Whalen (1976) also stated that their confidence interval method has the same assumptions as the two-way random effects ANOVA model; hence, they were apparently in agreement with Feldt about the assumptions that should be made for the confidence intervals.

Let us examine these assumptions more closely. Assumption (i) is a statement of the basic two-way ANOVA model, in which the interaction term has been omitted. This is the model used by other researchers of coefficient alpha, such as Hoyt (1941). Assumptions (ii) and (iii) specify that both subjects and conditions are random, and hence that Type 12 sampling is being used. Assumption (iv) specifies that subject effects--and hence subjects' true scores--are normally distributed. Assumption (v) includes three sub-assumptions: (a) that errors are normally distributed, (b) that error scores are independent of subject effects and hence of true scores, and (c) that errors are independent of each other. Finally, assumption (vi) states that σ_e^2 is constant.

The assumptions (v) that errors are independent of each other and (vi) that errors are equally variable, together imply that the variance-covariance matrix among the observations has a specific form, known as *compound symmetry.* In general, the variance-covariance matrix among observations has the following form:

$$\mathbf{A} = \begin{bmatrix} \sigma_p^2+\sigma_{e1}^2 & \sigma_p^2+\sigma_{e1e2} & \cdots & \sigma_p^2+\sigma_{e1ek} \\ \sigma_p^2+\sigma_{e1e2} & \sigma_p^2+\sigma_{e2}^2 & \cdots & \sigma_p^2+\sigma_{e2ek} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \sigma_p^2+\sigma_{e1ek} & \sigma_p^2+\sigma_{e2ek} & \cdots & \sigma_p^2+\sigma_{ek}^2 \end{bmatrix}$$

where σ_p^2 is the variance due to people, σ_{ei}^2 is the variance of the errors for Condition *i*, σ_{eiej} is the covariance of the errors for Conditions *i* and *j*. However, when errors are independent,$\sigma_{eiej} = 0$ for all $i \neq j$. When errors are equally variable, $\sigma_{ei}^2 = \sigma_{ej}^2$ for all *i* and *j*. Hence, the variance-covariance matrix of observations has equal variances and equal covariances, and is called compound symmetric. In the present case, this matrix would has the following form:

$$\begin{bmatrix} \sigma_p^2+\sigma_e^2 & \sigma_p^2 & \cdots & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2+\sigma_e^2 & \cdots & \sigma_p^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ \sigma_p^2 & \sigma_p^2 & \cdots & \sigma_p^2+\sigma_e^2 \end{bmatrix}$$

It should be noted that the assumptions listed above and identified by Feldt (1965) as the assumptions of the two-way random effects ANOVA model are now known to be sufficient but not necessary to conduct a test for equality of condition means under that model. These assumptions will therefore be referred to as the combined assumptions of compound symmetry, normality, and random sampling.

Feldt (1965) implied that compound symmetry was necessary for the confidence intervals for coefficient alpha. Hakstian and Whalen (1976) appear to have agreed with him. However, later work in the related area of repeated-measures ANOVA (to be discussed later) showed that compound symmetry was not necessary in that context. Because of this finding, subsequent researchers of coefficient alpha (e.g., Eom, 1993) have indicated that a weaker assumption than compound symmetry--that of *sphericity*--is sufficient for the sampling distribution of coefficient alpha. Sphericity is the assumption that the population covariance between two conditions is equal to the average of the variances of those two conditions minus a constant, as in Equation (5). Sphericity can also be defined in matrix notation as follows:

$$C'\Sigma C = \lambda \mathrm{I},$$

where C is a k x (k-1) orthonormal contrast matrix, Σ is the k x k population variance-covariance matrix, λ is a non-zero constant, and I is the (k-1) x (k-1) identity matrix.

Using the first definition of sphericity, given by Equation (5), compound symmetry can be viewed as the combination of sphericity and equality of variance. Thus, a spherical matrix that is not compound symmetric will violate the assumption of equality of variance. It will also violate the assumption of independent errors, because true differences between people can only account for constant non-negative covariances among conditions. Inequality of covariances are caused by unequal covariances among errors. Examples of spherical and compound symmetric variance-covariance matrices are given below:

Spherical

$$\begin{bmatrix} 100 & 75 & 75 & 125 \\ 75 & 150 & 100 & 150 \\ 75 & 100 & 150 & 150 \\ 125 & 150 & 150 & 250 \end{bmatrix}$$

Compound Symmetric

$$\begin{bmatrix} 100 & 50 & 50 & 50 \\ 50 & 100 & 50 & 50 \\ 50 & 50 & 100 & 50 \\ 50 & 50 & 50 & 100 \end{bmatrix}$$

The first hint that sphericity might be sufficient for the confidence intervals for coefficient alpha was found in Schroeder and Hakstian (1990). They were working with two facet generalizability coefficients, which are extensions of coefficient alpha that incorporate two different sources of error at the same time. They made the assumption of local circularity (which is a specific form of sphericity). No proof of the sufficiency of this assumption was given: the reader was referred to the repeated-measures ANOVA literature (i.e., Rouanet & Lepine, 1970). More directly, when Eom (1993) was working with the sampling distribution of coefficient alpha, he stated that sphericity, and not the more restrictive assumption of compound symmetry, is the assumption. He, too, offered no proof, but again referred the reader to the repeated-measures ANOVA literature, in this case Box (1954).

Because it was later work on coefficient alpha that contradicted Feldt's claims as to the assumptions of the confidence intervals, later articles by Feldt on other inferential procedures for coefficient alpha were examined to determine whether he had changed his mind about the assumptions of the sampling distribution of coefficient alpha. In his later articles, he restated the assumptions used in his earlier article (Feldt, 1969; 1980; 1992), and was firm in the claim that errors were independent and equally variable. Thus, Feldt implied that compound symmetry is necessary, but other researchers have disagreed, basing their disagreement on the repeated-measures ANOVA literature. Because of this, the repeated-measures ANOVA literature was examined for support for the claims that sphericity is sufficient for the confidence intervals for coefficient alpha.

The necessity of the assumptions of random sampling of people, and of normally distributed errors have not been questioned, and hence will be made throughout this thesis.

Repeated-Measures ANOVA

In repeated-measures ANOVA, each of *n* objects is measured on each of *k* variables. In the context of reliability, the objects would be people and the variables would be conditions. The effects of people and conditions can be tested by forming F-ratios for each. These F-ratios are calculated as follows:

$$F_{persons} = \frac{MS_{persons}}{MS_{error}}, \text{ and}$$

$$F_{conditions} = \frac{MS_{conditions}}{MS_{error}},$$

where

$$MS_{persons} = \frac{\sum_{i}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2}{n-1},$$

$$MS_{conditions} = \frac{\sum_{j}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2}{k-1}, \text{ and}$$

$$MS_{error} = \frac{\sum_{j}(\bar{x}_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2}{(n-1)(k-1)}.$$

In addition, X_{ij} =the score of the i^{th} person on the j^{th} condition, n is the number of subjects, and k is the number of conditions.

These observed F-ratios are then referenced to the central F-distribution with the appropriate degrees of freedom--(n-1), (n-1)(k-1) for the test on people, and (k-1), (n-1)(k-1) for the test on conditions--to determine if the effects of people and conditions are significantly different from zero. Thus, if the observed F-ratio exceeds the critical F-value, the null hypothesis of no effects is rejected.

The observed F-ratio for people is closely related to coefficient alpha:

$$r_{\alpha} = 1 - \frac{1}{F_{persons}}.$$

Hence, the distribution of coefficient alpha will be closely related to the distribution of this F-ratio. In fact, it was this relationship which allowed Feldt (1965) to derive the sampling distribution of coefficient alpha. Because of the close relationship between coefficient alpha and the F-test on people in repeated-measures ANOVA, this F-test was examined more closely.

Early texts stated that compound symmetry was necessary for repeated-measures ANOVA F-tests to be valid (e.g., Gaito, 1961; Winer, 1962). However, in 1970, Rouanet and Lepine showed that compound symmetry was not necessary, for $F_{conditions}$ to be F-distributed with (k-1) and (k-1)(n-1) degrees of freedom, although it was sufficient. They showed that the less restrictive assumption of sphericity is both necessary and sufficient. However, their work did not address the F-test for people. Huynh and Feldt (1970) also showed that sphericity was necessary and sufficient for the F-test for conditions, in repeated-measures ANOVA. They, too, omitted any discussion of the F-test on people. This omission is not surprising, because researchers are rarely interested in the F-test on people.

A third classic paper on the assumptions of repeated-measures ANOVA is that by Box (1954). In this paper, Box analyzed the separate effects of unequal variances and correlated errors on the F-tests for columns (conditions) and rows (people), using sophisticated matrix algebra. Several of Box's findings are relevant to the confidence intervals for coefficient alpha. First, he showed that correlated errors affect the expected values of the sums of squares of both people and errors. If the average correlation among the errors is positive, the F-test for people will be positively biased, and if the average correlation is negative, it will be negatively biased. Second, Box showed that the distributions of both sums of squares would be affected by inequality of variances and correlated errors. Nowhere in his paper did Box suggest that sphericity will result in the sums of squares being distributed in the usually-assumed manner.

Third, Box (1954) proved that the sums of squares for people and the sums of squares for errors will rarely be independent when there are unequal variances and correlated errors. A necessary condition for their independence is that $v_{s.}$, the average of the entries in the s^{th} row or column of the variance-covariance matrix among the errors, is constant. Box noted two cases in which $v_{s.}$ would be constant and, hence, the two sums of squares would be independent (although other cases of course might exist). The first occurs when the errors are uncorrelated and have equal variances; in such a case, the variance-covariance matrix among observations would be compound symmetric. The second case occurs "when the observations are circularly correlated" (Box, 1954, p. 489). Many current writers use the terms "circularity" and "sphericity" interchangeably; some others adopt the distinction that circularity is the property of the original variance-covariance matrix, when the transformed variance-covariance matrix, $C'\Sigma C$, is equal to λI, and hence is spherical. Apparently, because of these current uses of the term "circularity", Eom (1993) interpreted Box's claim to mean that sphericity of the error variance-covariance matrix (and hence of the observation variance-covariance matrix) is a necessary condition for the independence of the sums of squares for people and error. This clearly is not what Box intended. A quick look at an arbitrarily chosen spherical matrix, such as the one below, will reveal that the average of the elements is not necessarily constant across rows.

$$\begin{bmatrix} 200 & 125 & 100 \\ 125 & 150 & 75 \\ 100 & 75 & 100 \end{bmatrix}$$

Some other interpretation of "circularly correlated" is probably implied. Other uses of "circular" are common. One recent reference, Gifi (1990), referred to variance-covariance matrices which exhibited "circular contiguity". In this case, the elements of each row are the same, but the elements shift over one column in each successive row. Thus, for example, the following matrix exhibits circular contiguity:

$$\mathbf{B} = \begin{bmatrix} a & a & b & b & b & a \\ a & a & a & b & b & b \\ b & a & a & a & b & b \\ b & b & a & a & a & b \\ b & b & b & a & a & a \\ a & b & b & b & a & a \end{bmatrix} .$$

Because the elements of each row are the same, it is obvious that the average of the elements of each row is a constant. Thus, it might be that Box (1954) was referring to a pattern of correlations such as that displayed in the matrix **B** above. The plausibility of this interpretation is enhanced when one realizes that very few forms of variance-covariance matrixes are likely to satisfy the requirement that $v_{s.}$ be constant.

Thus, Box (1954) has shown that $v_{s.}$ being constant is a necessary condition for the independence of the sums of squares for people and errors. This condition is necessary, but not sufficient, for the F-test on people to be valid. Sphericity is not sufficient for $v_{s.}$ to be constant. Box also showed that the distributions of the sums of squares for people and errors will be influenced by correlated errors and unequal variances, and nowhere in his paper does Box suggest that sphericity would make the F-test valid. Thus, the Box (1954) paper provides no support for the claim that sphericity is sufficient for the F-test on people.

On the other hand, three necessary assumptions of the F-test on people are known. First, Box (1954) showed that unless the average correlation among the errors is zero, the F-test will be biased. Second, Box showed that unless $v_{s.}$ is constant, the sums of squares for people and the sums of squares for the errors will be correlated. These conditions are unlikely to both be met unless the variance-covariance matrix of observations is compound symmetric. Furthermore, Courrege and Rouanet (cited in Rouanet & Lepine, 1970) showed that the sums of squares for error will not be distributed as a central χ^2 unless the variance-covariance matrix is spherical (recall that compound symmetry is a special case of sphericity). Thus, it seems likely that the variance-covariance matrix must be compound symmetric for the F-ratio for people to be distributed as a central F with (*n*-

1) and (n-1)(k-1) degrees of freedom, although no definitive proof of this can be offered at this time.

In summary, no proof has been found for the claim that sphericity is sufficient for the F-ratio for people to be distributed as a central F with (n-1) and (n-1)(k-1) degrees of freedom under the null hypothesis. The more restrictive assumption of compound symmetry, however, is sufficient. It appears to be one of only a very limited number of assumptions which would meet all of the necessary conditions: (a) the average of the entries in the rows of the variance-covariance matrix among the errors is constant; (b) the average correlation among the errors is zero; and (c) and population variance-covariance matrix is spherical. This work can be directly applied to Feldt's sampling theory for coefficient alpha (and hence the confidence intervals for it) because the sampling theory was based directly on the distribution of the repeated-measures ANOVA F-test for people.

Thus, compound symmetry is sufficient for the confidence intervals for coefficient alpha, but no proof could be found that it is necessary. Sphericity, on the other hand, is necessary, but no proof could be found that it is sufficient. Therefore, an empirical study was conducted to determine whether sphericity is sufficient, or if data must be compound symmetric.

Method

A Monte Carlo simulation experiment was conducted to compare the actual proportions of the .95 confidence intervals which included the population values of coefficient alpha, using the two confidence interval methods with data that conform to the two sets of assumptions. A double precision FORTRAN program, described in Study 1, was used. The two kinds of data--(a) spherical, but not compound symmetric, and (b) compound symmetric--were compared to determine which, if either, set of assumptions was sufficient for the two confidence interval methods.

Kinds of Data Studied

Two kinds of data were studied: (a) data which met the assumption of compound symmetry, and (b) data in which the variance-covariance matrix was spherical but had unequal variances, and hence was not compound symmetric. First, compound symmetric data were constructed. Because no assumption of equal condition means was made by Feldt (1965), unequal condition means were used. Second, data was constructed that were equivalent to the first kind of data, except that the inter-condition variance-covariance matrix was spherical but not compound symmetric. These latter data sets had both unequal means and unequal variances.

Experimental Design

Study 1 indicated that Type 12 sampling needed to be used for this study. Because of this, the preliminary studies in Appendix C were used to determine the best number of conditions and *SDVar* to use for the spherical (but not compound symmetric) data. For the spherical data that were not compound symmetric, variances were unequal with a mean of 100 and a standard deviation of 10. Under sphericity, covariances were determined by the population value of coefficient alpha, the number of conditions, and the variances, as in Equation (5). The variance due to people was not specified.

For the compound symmetric data, variances were equal and set to 100. Numbers of conditions were matched with those for the spherical data that were not compound symmetric. The variance due to people is considered true score variance and was determined by the population value of coefficient alpha, the number of conditions, and the average variance, as in Equation (6).

For both kinds of data, two numbers of conditions were used: 5 and 20. Condition means were unequal, with a mean of 0 and standard deviation of 20. The sample size was 100. Population values of alpha were set to .60, .75, and .90.

Table 7 provides a summary of the variables and levels used.

To ensure that the slightest departure from nominal levels would be detected, 20,000 replications of each cell were run.

Data Generation

Data generation was carried out differently for the two kinds of data. Generation of the spherical data that were not compound symmetric will be described first, followed by a description of the generation of compound symmetric data.

Spherical Data that were not Compound Symmetric

Generation of spherical data with unequal means and variances was broken into three steps. First, a spherical population variance-covariance matrix was randomly generated. This was then used to produce observations. Finally, differences in condition means were taken into account by adding randomly generated condition effects to these observations. Because Type 12 sampling was used, all steps were repeated for each sample. If the reader is familiar with this data generation process from Study 1, this section can be ignored.

Generation of the population variance-covariance matrix.

The first step was to generate a spherical variance-covariance matrix, consisting of the population values of variances and covariances of k randomly selected conditions. This

Table 7

Combinations of Variables and Levels of Variables Used in Study 2

Independent Variable	C.S.	Spherical
Average Condition Variance (*AveVar*)	100	100
SD of Condition Variances (*SDVar*)	0	10
SD of Condition Means (*SDMean*)	20	20
Number of Conditions (*k*)	5,20	5, 20
Sample Size (*n*)	100	100
Coefficient Alpha (ρ_α)	.6, .75, .90	.6, .75, .90
Type of Sampling	Type 12	Type 12

Note. C.S. = Cells where the variance-covariance matrix among conditions was compound symmetric; Spherical = Cells where the variance-covariance matrix among conditions was spherical but not compound symmetric.

was done by randomly generating variances and then calculating the necessary covariances so that the overall variance-covariance matrix was perfectly spherical and had the desired value of coefficient alpha.

To generate the variances, independent random normal variates with mean 0 and standard deviation 1 were randomly generated. These were then multiplied by the desired standard deviation (*SDVar*) and added to the desired mean (*AveVar*). For example, for a 5-condition cell, with *AveVar* = 100, *SDVar* = 10, the randomly generated variances might be 108.93, 92.98, 91.06, 113.22, and 96.05.

Next, covariances were calculated that would produce a perfectly spherical matrix with the desired value of coefficient alpha. The covariances were given by Equation (5). In our example, if the desired value of coefficient alpha is .60, then the spherical matrix which results is

$$\mathbf{V} = \begin{bmatrix} 108.93 & 24.03 & 23.07 & 34.15 & 25.57 \\ 24.03 & 92.98 & 15.10 & 26.18 & 17.59 \\ 23.07 & 15.10 & 91.06 & 25.22 & 16.63 \\ 34.15 & 26.18 & 25.22 & 113.22 & 27.71 \\ 25.57 & 17.59 & 16.63 & 27.71 & 96.05 \end{bmatrix}$$

When a matrix that was not positive-definite resulted, the data generation process returned to the beginning.

Generation of observations.

This variance-covariance matrix, $\mathbf{V}$, was used to generate observations. First, the matrix $\mathbf{V}$ was decomposed using a Choleski decomposition, such that $\mathbf{V} = \mathbf{F}\mathbf{F}'$. Then, *n*x*k* multivariate normal observations (where n is the number of subjects and k the number of conditions) were randomly generated with mean 0 and standard deviation 1, and post-multiplied by $\mathbf{F}'$.

These observations represent the sum of the effects due to true differences between people, and the effects of random errors. The variance-covariance matrix, $\mathbf{V}$, can be viewed as the sum of the variance-covariance matrix due to true differences between people and the matrix due to errors. The variances were the sums of the variance due to true differences between people and the variance due to errors. The covariances were the sums of the variance due to true differences between people and the covariance between errors. See matrix **A** on page 53.

Generation of condition effects.

Finally, differences in condition means were taken into account by adding randomly generated condition effects to the observations just generated. First, independent random normal variates with mean 0 and standard deviation 1 were generated. These were multiplied by the desired standard deviation (*SDMean*). These condition effects were then added to the observations. Thus, if the randomly generated condition effects were 13.32, 25.19, -19.49, -20.51, and .04, then 13.22 would have been added to the observations in condition 1 for all subjects, 25.19 would have been added to the observations in condition 2, -19.49 for condition 3, and so forth.

Problems encountered in data generation.

When Type 12 sampling was used for data with unequal variances, some of the matrices were unacceptable and were discarded, as discussed above. If this had happened frequently, the samples might have been considerably biased, making it impossible to determine the cause of any apparent differences between cells. To make interpretation clear, 100% (or nearly 100%) of the randomly generated variance-covariance matrices should be acceptable. This was therefore controlled by careful selection of the number of conditions and the degree of heterogeneity of the condition variances. Preliminary studies suggested that with the standard deviation of the variances (*SDVar*) equal to 10 very few, if any, of these matrices would be unacceptable for 5- and 20-condition cells, and hence these values were used. With higher values of *SDVar*, more than 5% of the matrices were likely to be unacceptable, for some cells. See the results of the preliminary studies in Appendix C.

Compound Symmetric Data

Compound symmetric data would result from parallel or essentially parallel conditions: true scores for different conditions might differ by a constant, but error variances are assumed to be equal and errors independent. Thus, conditions may have unequal means (and in this case they did), but they must have equal variances. The data generation process was broken into four steps: generation of conditions, generation of subjects, generation of errors, and addition of these components. The first two steps can be thought of as the random selection of conditions and subjects from their respective populations. The third step can be thought of as the collection of sample data for these conditions and subjects. Because Type 12 sampling was used, all four steps were repeated for each sample value of coefficient alpha calculated.

Generation of conditions.

To generate conditions, unequal means were randomly generated. Independent random normal variates with mean 0 and standard deviation 1 were generated. These were multiplied by the desired standard deviation (*SDMean*).

Generation of subjects.

To randomly generate the true scores of subjects, independent random normal variates were generated with a mean of 0 and a standard deviation of 1. These were multiplied by the desired standard deviation (the square root of *TVar*, as given in Equation (6)).

Generation of errors.

Independent random normal variates (*nk* of them) were generated with a mean of 0 and a standard deviation of 1. These were multiplied by the desired standard deviation (the square root of the error variance for that condition), which was randomly generated in the first step.

Addition of components.

The three components were then added together: $X_{ij} = \pi_i + \tau_j + \varepsilon_{ij}$, where X_{ij} is the observed score of Person *i* on Condition *j*, π_i is the person effect for Person *i*, τ_j is the condition effect for Condition *j*, and ε_{ij} is the residual or error component.

Problems encountered in data generation.

No data generation problems were encountered with this kind of data.

Data Analysis

For each sample, the Hakstian and Whalen (1976) and Feldt (1965) confidence intervals were calculated. These confidence intervals were compared to the desired population value of coefficient alpha to determine if they included that value. Summaries of the numbers and proportions of confidence intervals that included the population value were made for each type of confidence interval, for each cell in the design. These were compared to nominal levels using statistical tests on proportions.

Results

In this section, the results of the Monte Carlo studies are presented. The results for the data meeting the assumption of compound symmetry are presented first. Compound symmetry is implied by the assumptions made by Kristof (1963) and Feldt (1965) during the development of the sampling theory of coefficient alpha and its confidence intervals. Data which met this assumption were therefore expected to produce precise confidence intervals. Next, the results for the spherical non-compound symmetric data are presented.

Compound Symmetric Data

Before testing whether these data produced confidence intervals that captured the population value of coefficient alpha at the nominal level, several data checks were done to ensure that the data were generated as intended and that no confounds were introduced.

Data Checks

Levels of independent variables.

Because Type 12 sampling was used, population characteristics of conditions were randomly selected for each sample. These population characteristics, when averaged across replications, should have been close to the desired values. Thus, the average and variance of the condition means should have been close to their desired values (0 and *VarMean*, respectively). Similarly, the average and variance of condition variances should have been close to their desired values (*AveVar* and *VarVar*, respectively). This was the case; these values were close to the desired values in all cells used. See Appendix I.

Sample values of coefficient alpha.

Across replications, sample values of coefficient alpha should have had an average close to the desired population value. If the average of the sample values of coefficient alpha was very different from the desired value of coefficient alpha, this could have affected the performance of the confidence intervals. From Table 8, sample values of coefficient alpha appear to have been slightly negatively biased. The bias does not appear to be large enough to greatly affect the accuracy of the confidence intervals. However, it is interesting to note that this bias appears to be largest for small values of population coefficient alpha.

Accuracy of the Confidence Intervals

Table 9 presents the results of the Monte Carlo study for compound symmetric data. The proportions of the confidence intervals that included the population value were compared to the nominal level (.95), using standard tests on proportions. None of the cells was significantly different from nominal. These significance tests possessed high levels of power, based as they were on 20,000 replications. In fact, the result which departed most from nominal levels was .94795, a mere .00205 different from nominal .95.

Spherical Data that were not Compound Symmetric

Before testing whether the spherical non-compound symmetric data produced confidence intervals that captured the population value of coefficient alpha at the nominal level, several data checks were done to ensure that the data were generated as intended and that no confounds were introduced.

Table 8

Average Across Replications of the Sample Values of Coefficient Alpha, for Compound Symmetric Data

Conditions	Population Alpha	Sample Alpha	Bias
5	.60	.5911	-.0089
5	.75	.7451	-.0049
5	.90	.8980	-.0020
20	.60	.5922	-.0078
20	.75	.7447	-.0053
20	.90	.8980	-.0020

Note. Conditions = number of conditions; Bias = the difference between average sample value and desired population value.

Table 9

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Compound Symmetric Data

Conditions	Alpha	H&W CI	Feldt CI
5	.60	.94835	.94870
5	.75	.95165	.95145
5	.90	.95050	.95070
20	.60	.95050	.95070
20	.75	.94830	.94795
20	.90	.94815	.94820

Note. Conditions = the number of conditions; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 20,000 replications per cell.

Data Checks

Matrices that were not positive definite.

As discussed under data generation problems, it is possible to randomly generate spherical matrices that are not positive definite and hence must be discarded. If this had happened frequently, then the results might have been be considerably biased. The proportions of randomly generated matrices that were not positive definite were usually zero and always small, and could not account for any differences between observed and nominal levels. See Appendix J.

Levels of independent variables.

The average and variance of the condition means were close to their desired values (0 and *VarMean*, respectively). Similarly, the average and variance of condition variances were close to their desired values (*AveVar* and *VarVar*, respectively). See Appendix K.

Sample values of coefficient alpha.

The averages across replications of sample values of coefficient alpha should have been close to the desired population values. From Table 10, sample values of coefficient alpha appear to have been slightly negatively biased. In some cases this bias appears to have been quite large. The bias was greatest for low values of coefficient alpha, and for larger numbers of conditions.

Accuracy of the Confidence Intervals

Table 11 presents the results of the Monte Carlo study for spherical data that were not compound symmetric. The proportions of the confidence intervals that included the population value were compared to the nominal level (.95). As can be seen from the table, spherical data produced inaccurate confidence intervals in every cell in the design.

Table 10

Average Across Replications of the Sample Values of Coefficient Alpha, for Spherical (but not Compound Symmetric) Data

Conditions	Population Alpha	Sample Alpha	Bias
5	.60	.5870	-.0130
5	.75	.7426	-.0074
5	.90	.8977	-.0023
20	.60	.5776	-.0224
20	.75	.7407	-.0093
20	.90	.8973	-.0027

Note. Conditions = the number of conditions; Bias = the difference between average sample value and desired population value.

Table 11

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Spherical (but not Compound Symmetric) Data

Conditions	Alpha	H&W CI	Feldt CI
5	.60	.88775*	.88815*
5	.75	.91460*	.91495*
5	.90	.93235*	.93270*
20	.60	.77075*	.77075*
20	.75	.85935*	.85935*
20	.90	.92900*	.92895*

* indicates that the proportion is significantly ($p < .05$) different from .95.

Note. Conditions = the number of conditions; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 20,000 replications per cell.

The confidence intervals were least accurate with small population values of coefficient alpha (.60) and with the larger number of conditions (20). This result may have been aided, to a slight extent, by the greater bias of the sample values of coefficient alpha for the smaller values of coefficient alpha and larger numbers of conditions.

Discussion and Conclusions

All of the cells using compound symmetric data produced accurate confidence intervals. This assumption appears to be sufficient for the confidence intervals to be precise, as expected. None of the cells based on the spherical data that were not compound symmetric produced accurate confidence intervals. In some cases, these confidence intervals were quite inaccurate; in one case, as few as 77% of the .95 confidence intervals included the population value of coefficient alpha. There were no indications that data generation problems had caused these inaccuracies. Thus, it appears that sphericity alone is not sufficient for precise confidence intervals for coefficient alpha.

In the case of the spherical data that were not compound symmetric, it was found that the sample estimates of the population value of coefficient alpha were in some cases quite biased. Furthermore, the degree of inaccuracy of the confidence intervals appears to correspond to the size of the bias. Thus, it is possible that if this bias could be corrected, these confidence intervals might be *somewhat* more accurate. The degree of improvement which might be realized from this kind of adjustment cannot be estimated, because the sampling error of coefficient alpha was not calculated. It is quite possible that even with this adjustment, the confidence intervals would capture the population value of coefficient alpha much less often than desired.

These findings show that sphericity is not sufficient for the confidence intervals for coefficient alpha, although the assumption of compound symmetry is. Given the combined evidence from the literature review and the present empirical study, it seems likely that the confidence intervals require the assumption of compound symmetry, but no definitive proof of this hypothesis can be offered at this time.

STUDY 3:
AN EXAMINATION OF THE PERFORMANCE OF CONFIDENCE INTERVALS FOR COEFFICIENT ALPHA UNDER WELL-KNOWN MEASUREMENT MODELS

Introduction

This study focuses on issues of interest to the practitioner who wishes to use the confidence intervals for coefficient alpha to tell him or her about the consistency of measures. These issues are primarily the following: when can these confidence intervals be expected to work, and are there any differences in the performance of the two types of confidence intervals being studied?

Many distinctive kinds of data exist, and it is impossible to address them all in a single thesis. This study focused on four well-known measurement models: parallelism, essential parallelism, tau-equivalence, and essential tau-equivalence. Each of these is described briefly.

If two conditions are parallel, then every person's true score on the second condition is the same as his or her true score on the first condition. As well, the errors on the two conditions are independent, and have the same variance. The two conditions are indistinguishable statistically, and the variance-covariance matrix of observations among several parallel conditions will be compound symmetric. Parallelism is a theoretical assumption that is rarely met in the real world. However, the assumption of parallelism is often made. For example, the usual methods of calculating test-retest reliability and alternate forms reliability assume that the two sets of scores are parallel.

If two conditions are essentially parallel, every person's true score on the second condition is equal to his or her true score on the first condition, plus some constant. The constant is the same for all people. As with parallel conditions, the errors are independent and equally variable, and the variance-covariance matrix of several essentially parallel conditions will be compound symmetric. The only difference between parallel and essentially parallel conditions is that with essentially parallel conditions, means may differ.

The assumption of essential parallelism, like that of parallelism, is largely a theoretical assumption. However, it is slightly more realistic than that of strict parallelism. For example, if the conditions were short essay questions, condition means might differ because some questions were easier than others; however, unless some conditions had means which were near zero or near perfect, one might suppose that condition variances could be equal. Essential parallelism is not an assumption which is often made, despite being more easily satisfied than the assumption of parallelism. It is included here for completeness, and because it leads to compound symmetry of the variance-covariance matrix without making any additional assumptions.

With tau-equivalent conditions, while true scores are equal for all conditions and errors are independent, the errors may not be equally variable. Thus, conditions are of equal difficulty, but are not interchangeable. Some conditions measure more exactly what one is trying to measure; while others incorporate more sources of error. Hence, the variance-covariance matrix has equal covariances but possibly unequal variances. The difference between parallelism and tau-equivalence is that with tau-equivalent conditions, condition variances may differ.

Finally, if conditions are essentially tau-equivalent, then a subject's true scores on different conditions may differ by some constant, and errors, though independent, may not be equally variable. Thus, both means and variances can be unequal. As with the tau-equivalent conditions, the variance-covariance matrix has equal covariances but possibly unequal variances. The difference between essential tau-equivalence and tau-equivalence is that condition means may differ.

Essential tau-equivalence is seen as a reasonable measurement model (de Gruijter & van der Kamp, 1984). With a carefully selected set of conditions administered in random order, the assumption of uncorrelated errors is quite reasonable. Furthermore, in many cases, the other three measurement models will be obviously inappropriate because condition means and variances are heterogeneous.

It should be noted that all parallel conditions are also essentially parallel, tau-equivalent, and essentially tau-equivalent, and that all essentially parallel and all tau-equivalent conditions are also essentially tau-equivalent. However, for ease of reference throughout this study, "essentially parallel" conditions means conditions which are essentially parallel but not parallel. Similarly, "essentially tau-equivalent" conditions means essentially tau-equivalent conditions which are not tau-equivalent or essentially parallel. In this way, the four categories considered were non-overlapping. See Table 12.

Method

A Monte Carlo study was conducted to compare the actual proportions of the .95 confidence intervals that included the population value of coefficient alpha, using the two confidence interval methods, for data conforming to the four measurement models. As

Table 12

Relationship of the Four Measurement Models to Equality of Condition Means and Variances

		Condition Means	
		Equal	Unequal
Condition Variances	Equal	Parallel	Essentially Parallel
	Unequal	Tau-Equivalent	Essentially Tau-Equivalent

well, the effects of type of confidence interval, number of conditions, population value of coefficient alpha, and heterogeneity of condition means and variances were examined. A double precision FORTRAN program, described in Study 1, was used.

Kinds of Data Studied

Four kinds of data were studied: parallel, essentially parallel, tau-equivalent, and essentially tau-equivalent. These are described above. All data considered were multivariate normal and continuous.

Experimental Design

Some characteristics of the data were the same for all four measurement models. Population values of coefficient alpha were .60, .75, and .90. The proportion of variance due to people was determined by the population value of coefficient alpha. The sample size was 100. The number of conditions were five and 20.

The standard deviations of the condition means and variances defined the four kinds of data studied. For the parallel data, both condition means and condition variances were held constant. Thus, condition means had a mean of 0, and a standard deviation of 0, while condition variances had a mean of 100, and a standard deviation of 0. For essentially parallel data, condition means varied, but condition variances were held constant. Thus, condition means had a mean of 0, and standard deviations of 10 or 20, while condition variances were set to 100. For tau-equivalent data, condition means were held constant (mean 0, standard deviation 0), while condition variances were unequal: for the 5-condition cells, they had a mean of 100 and standard deviations of 10 and 15; for the 20-condition cells, they had a mean of 100 and standard deviations of 10, 15 and 25. Finally, for the essentially tau-equivalent data, condition means and variances were both unequal. Condition means had a mean of 0, and standard deviations of 10 and 20, while condition variances had a mean of 100, and standard deviations of 10 and 15 for the 5-condition cells, and 10, 15, and 25 for the 20-condition cells. Table 13 provides a summary of the variables and levels of variables used.

Study 1 indicated that Type 12 sampling needed to be used for this study. Because of this, the levels of numbers of conditions and the standard deviations of variances were selected based on preliminary studies (see Appendix D) to minimize data generation problems.

Data Generation

The four different kinds of data were generated using the same data generation process. This process was broken into four steps: generation of conditions, generation of subjects, generation of errors, and addition of these components. The first two steps can be thought of as the random selection of conditions and subjects from their respective populations. The third step can be thought of as the collection of sample data for these conditions and subjects. Because Type 12 sampling was used, all four steps were repeated for each sample value of coefficient alpha calculated.

Generation of conditions.

Generation of conditions involved randomly generating means and variances for each condition. Depending on the kind of data being simulated, either means or variances or both may have been constant. If so, these were set to their average values.

If conditions had different variances (as they did with tau-equivalent and essentially tau-equivalent data), then the reason for this was that their error variances were unequal. To randomly generate error variances, independent random normal variates with mean 0 and

Table 13

Combinations of Variables and Levels of Variables Used in Study 3

	Parallel	Essentially Parallel	Tau-Equivalent	Essentially Tau-Equivalent
Average of Means (*AveMean*)	0	0	0	0
SD of Means *(SDMean)*	0	10, 20	0	10, 20
Average of Variances *(AveVar)*	100	100	100	100
Number of Conditions (*k*)	5, 20	5, 20	5, 20	5, 20
Sample Size (*n*)	100	100	100	100
Coefficient Alpha (ρ_α)	.60, .75, .90	.60, .75, .90	.60, .75, .90	.60, .75, .90
SD of Variances *(SDVar)*				
For five Condition cells:	0	0	10, 15	10, 15
For 20 Condition cells:	0	0	10, 15, 25	10, 15, 25
Type of Sampling	Type 12	Type 12	Type 12	Type 12

standard deviation 1 were generated. These were multiplied by the desired standard deviation (*SDVar*), and were added to the desired mean (*AveErr*). *AveErr* was calculated based on the average total variance (*AveVar*), the number of conditions and the population value of alpha (ρ_α): $AveErr = AveVar - TVar$, where *TVar* was calculated according to Equation (6).

If condition means were unequal (as they were with essentially parallel and essentially tau-equivalent data), then condition effects and hence condition means were randomly generated. First, *k* independent random normal variates with mean 0 and standard deviation 1 were generated, and then these were multiplied by the desired standard deviation (*SDMean*).

Generation of subjects.

To randomly generate the true scores of subjects, *n* independent random normal variates were generated with a mean of 0 and a standard deviation of 1. These were multiplied by the desired standard deviation (the square root of *TVar,* which is given by Equation (6)).

Generation of errors.

To generate errors, *nk* independent random normal variates were generated with a mean of 0 and a standard deviation of 1. These were multiplied by the desired standard deviation: the square root of the error variance for that condition, which was randomly generated (or set to 100) in the first step.

Addition of components.

The three components were then added together: $X_{ij} = \pi_i + \tau_j + \varepsilon_{ij}$, where X_{ij} is the observed score of Person *i* on Condition *j*, π_i is the person effect for Person *i*, τ_j is the condition effect for Condition *j*, and ε_{ij} is the residual or error component.

Problems encountered in data generation.

When unequal variances were called for, negative numbers were sometimes generated for the error variances. When this occurred, new variances were generated. If this had occurred frequently, samples might have been considerably biased. Therefore, preliminary studies were run, and were used to select numbers of conditions and values of *SDVar* for which this would not occur very often (less than 5% of the time).

For each cell, 5,000 replications were run. The number of replications was reduced (from that of Studies 1 and 2) so that a larger number of cells could be run, and a greater range of effects assessed.

Data Analysis

For each sample, the Hakstian and Whalen (1976) and Feldt (1965) confidence intervals were calculated. These confidence intervals were compared to the desired population value of coefficient alpha, to determine whether they included the population value. Summaries of the numbers and proportions of confidence intervals that included the population value were made for each cell in the design.

Results

In this section, the results of the Monte Carlo study are presented. Before assessing the effects of the different kinds of data and independent variables on the performance of the confidence intervals, several data checks were done.

Data Checks

Negative variances.

As discussed under data generation problems, negative numbers were sometimes generated for the variances. The proportion of sets of variances that were discarded because one or more of them was negative was generally very small, and usually less than 1%. For three of the 20-condition cells, these proportions were slightly over the .05 level, which was the criterion for choosing combinations of *SDVar* and the number of conditions. However, the highest proportion was only .0663, which is still small. Thus, there is no reason to believe that this has significantly biased the samples. See Appendix L.

Levels of independent variables.

To ensure that independent variables were manipulated as desired, the average and variance of condition means and variances were stored for each sample, and then averaged over the 5,000 samples in each cell. These values were usually very close to the desired values. However, there were two exceptions. For ρ_α = .90, *SDVar* = 15, and k = 5, the variances of the variances were slightly smaller on average than desired. The same occurred for ρ_α = .90, *SDVar* = 25, and k = 20. The fact that variances were slightly less heterogeneous than desired can be attributed to the discarding of sets of variances with negative values, which on average would have larger variances than sets of variances with no negative values. This was done about 5% of the time for each of the two cells in question. However, in no case is the bias large enough to confuse interpretation: samples still had close to the desired characteristics. Between cell differences in *SDVar* were much larger than the differences between desired and average observed values of *SDVar*. See Appendix M.

Sample values of coefficient alpha.

Previous researchers (Cronbach et al., 1963; Kristof, 1963) have noted that sample values of coefficient alpha are biased. This bias might have affected the performance of the confidence intervals. Tables 14 and 15 present the average of the sample values of coefficient alpha for each cell in the design. For both the 5- and 20-condition cells, this bias appears to be greatest for low population values of coefficient alpha.

Accuracy of the Confidence Intervals for the Four Measurement Models

The results of the Monte Carlo study were examined to determine the accuracy of the confidence intervals for the four measurement models. Each of these statistical tests on the proportions was run at $p = .01$, to control the family-wise error rate. In none of the six cells with parallel data and none of the 12 cells with essentially parallel data did the actual confidence coefficient differ significantly from nominal (see Tables 16 and 17). This was as expected, because parallelism and essential parallelism both meet the assumption of compound symmetry. Averaging over cells and type of confidence intervals, .9484 of the samples under parallelism, and .9475 of the samples under essential parallelism captured the population value. These values were not significantly different ($z = .5598$, $p > .05$). On the other hand, tau-equivalent and essentially tau-equivalent data did not produce accurate confidence intervals (see Tables 18 and 19). Often, the proportion of the confidence intervals that captured the population

Table 14

Average Across Replications of the Sample Values of Coefficient Alpha, for Five-Condition Cells

SDMean	*SDVar*	Population Alpha	Sample Alpha	Bias
0	0	.60	.5922	-.0078
0	0	.75	.7448	-.0052
0	0	.90	.8978	-.0022
0	10	.60	.5921	-.0079
0	10	.75	.7453	-.0047
0	10	.90	.8979	-.0021
0	15	.60	.5924	-.0076
0	15	.75	.7453	-.0047
0	15	.90	.8979	-.0021
10	0	.60	.5916	-.0084
10	0	.75	.7454	-.0046
10	0	.90	.8981	-.0019
10	10	.60	.6003	.0003
10	10	.75	.7457	-.0043
10	10	.90	.8980	-.0020
10	15	.60	.5929	-.0071
10	15	.75	.7467	-.0033
10	15	.90	.8976	-.0024

con't

Table 14 *con't*

SDMean	*SDVar*	Population Alpha	Sample Alpha	Bias
20	0	.60	.5935	-.0065
20	0	.75	.7446	-.0054
20	0	.90	.8980	-.0020
20	10	.60	.6005	.0005
20	10	.75	.7445	-.0055
20	10	.90	.8978	-.0022
20	15	.60	.5919	-.0081
20	15	.75	.7465	-.0035
20	15	.90	.8973	-.0027

Note. SDMean = the standard deviation of condition means; *SDVar* = the standard deviation of condition variances; Bias = the difference between the average of the sample values and the desired population value.

Table 15

Average Across Replications of the Sample Values of Coefficient Alpha, for 20-Condition Cells

SDMean	*SDVar*	Population Alpha	Sample Alpha	Bias
0	0	.60	.5899	-.0101
0	0	.75	.7456	-.0044
0	0	.90	.8978	-.0022
0	10	.60	.5902	-.0098
0	10	.75	.7447	-.0053
0	10	.90	.8979	-.0021
0	15	.60	.5916	-.0084
0	15	.75	.7449	-.0051
0	15	.90	.8978	-.0022
0	25	.60	.5933	-.0067
0	25	.75	.7455	-.0045
0	25	.90	.9874	-.0026
10	0	.60	.5910	-.0090
10	0	.75	.7440	-.0060
10	0	.90	.8980	-.0020
10	10	.60	.5926	-.0084
10	10	.75	.7459	-.0041
10	10	.90	.8981	-.0019

con't

Table 15 *con't*

SDMean	*SDVar*	Population Alpha	Sample Alpha	Bias
10	15	.60	.5927	-.0083
10	15	.75	.7450	-.0050
10	15	.90	.8980	-.0020
10	25	.60	.5936	-.0064
10	25	.75	.7444	-.0056
10	25	.90	.8978	-.0022
20	0	.60	.5923	-.0087
20	0	.75	.7450	-.0050
20	0	.90	.8979	-.0021
20	10	.60	.5912	-.0088
20	10	.75	.7444	-.0056
20	10	.90	.8979	-.0021
20	15	.60	.5912	-.0088
20	15	.75	.7456	-.0044
20	15	.90	.8982	-.0018
20	25	.60	.5923	-.0077
20	25	.75	.7453	-.0047
20	25	.90	.8978	-.0022

Note. SDMean = the standard deviation of condition means; *SDVar* = the standard deviation of condition variances; Bias = the difference between the average of the sample values and the desired population value.

Table 16

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Parallel Data

Conditions	Alpha	H & W CI	Feldt CI
5	.60	.9504	.9506
5	.75	.9514	.9506
5	.90	.9466	.9446
20	.60	.9462	.9456
20	.75	.9488	.9500
20	.90	.9480	.9480

Note. Conditions = the number of conditions; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 5,000 replications per cell.

Table 17

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Essentially Parallel Data

Conditions	*SDMean*	Alpha	H & W CI	Feldt CI
5	10	.60	.9502	.9510
5	10	.75	.9488	.9504
5	10	.90	.9496	.9498
5	20	.60	.9428	.9432
5	20	.75	.9442	.9423
5	20	.90	.9458	.9468
20	10	.60	.9512	.9514
20	10	.75	.9438	.9444
20	10	.90	.9428	.9436
20	20	.60	.9502	.9498
20	20	.75	.9496	.9494
20	20	.90	.9498	.9496

Note. Conditions = the number of conditions; *SDMean* = the standard deviation of the condition means; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 5,000 replications per cell.

Table 18

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Tau-Equivalent Data

Conditions	*SDVar*	Alpha	H & W CI	Feldt CI
5	10	.60	.9458	.9464
5	10	.75	.9308*	.9302*
5	10	.90	.8840*	.8848*
5	15	.60	.9372*	.9372*
5	15	.75	.9126*	.9148*
5	15	.90	.8270*	.8264*
20	10	.60	.9520	.9522
20	10	.75	.9494	.9492
20	10	.90	.9488	.9488
20	15	.60	.9378*	.9380*
20	15	.75	.9414*	.9422
20	15	.90	.9414*	.9414*
20	25	.60	.9432	.9432
20	25	.75	.9430	.9436
20	25	.90	.9248*	.9526

* indicates that the observed proportion was significantly ($p < .01$) different from the nominal proportion of .95.

Note. Conditions = the number of conditions; *SDVar* = the standard deviation of the condition variances; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 5,000 replications per cell.

Table 19

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, for Essentially Tau-Equivalent Data

Conditions	*SDMean*	*SDVar*	Alpha	H & W CI	Feldt CI
5	10	10	.60	.9460	.9456
5	10	10	.75	.9344*	.9340*
5	10	10	.90	.8866*	.8892*
5	10	15	.60	.9358*	.9376*
5	10	15	.75	.9110*	.9108*
5	10	15	.90	.8278*	.8288*
5	20	10	.60	.9472	.9456
5	20	10	.75	.9326*	.9328*
5	20	10	.90	.8830*	.8828*
5	20	15	.60	.9392*	.9392*
5	20	15	.75	.9126*	.9148*
5	20	15	.90	.8302*	.8294*
20	10	10	.60	.9444	.9444
20	10	10	.75	.9462	.9460
20	10	10	.90	.9526	.9518
20	10	15	.60	.9484	.9488
20	10	15	.75	.9414*	.9416*
20	10	15	.90	.9428	.9432
20	10	25	.60	.9430	.9434

con't

Table 19 *con't*

Conditions	*SDMean*	*SDVar*	Alpha	H & W CI	Feldt CI
20	10	25	.75	.9354*	.9344*
20	10	25	.90	.9174*	.9178*
20	20	10	.60	.9496	.9488
20	20	10	.75	.9496	.9496
20	20	10	.90	.9432	.9430
20	20	15	.60	.9470	.9470
20	20	15	.75	.9422	.9412*
20	20	15	.90	.9412*	.9410*
20	20	25	.60	.9436	.9450
20	20	25	.75	.9386*	.9382*
20	20	25	.90	.9218*	.9222*

* indicates that the observed proportion was significantly ($p < .01$) different from the nominal proportion of .95.

Note. Conditions = the number of conditions; *SDMean* = the standard deviation of condition means; *SDVar* = the standard deviation of condition variances; Alpha = the desired population value of coefficient alpha; H&W CI = the proportion of the Hakstian and Whalen (1976) confidence intervals that included the population value of coefficient alpha; Feldt CI = the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha. Proportions were based on 5,000 replications per cell.

value of coefficient alpha was significantly ($p < .01$) different from the nominal level of .95. These differences were in some cases substantial. For the tau-equivalent data, in one case as few as 82.64% of the confidence intervals included the population value. For the essentially tau-equivalent data, this figure went as low as 82.78%. For the tau-equivalent data, the average over cells and type of confidence interval was .9290, and for essentially tau-equivalent it was .9279. These values were not significantly different (z = .6399, $p > .05$).

Effect of Type of Confidence Interval

The Hakstian and Whalen (1976) and Feldt (1965) confidence intervals were compared to determine if either was more accurate. Perusal of Tables 16, 17, 18 and 19 indicates that the two types of confidence interval performed similarly under each of the kinds of data studied. Averaged over the 63 cells, the proportions of the confidence intervals that included the population value were .9336 for the Hakstian and Whalen confidence intervals, and .9337 for the Feldt confidence intervals. Thus, the differences between the two types of confidence intervals are very small, and are unlikely to make any practical difference in a research situation.

Effect of Inequality of Condition Means

Perusal of Tables 16, 17, 18 and 19 suggests that inequality of condition means had no effect on the accuracy of the confidence intervals. This impression was confirmed by use of a statistical test. First, the mean of the proportions of the confidence intervals that included the population value was calculated for each level of *SDMean*. Averaging across the type of confidence interval used, these averages were .9206, .9215, and .9197 for *SDMean* = 0, 10, and 20, respectively, for the 5-condition cells. For 20-condition cells, these averages were .9439, .9429, and .9434 for *SDMean* = 0, 10, and 20, respectively. These averages are based on 45,000 samples for each of the 5-condition cells, and 60,000 samples for the 20-condition cells. The three proportions for $k = 5$ were compared, using standard tests on proportions and a family wise error rate of .05; differences between the proportions were non-significant. This process was repeated for $k = 20$, and, again, the differences were non-significant.

Effects of Inequality of Condition Variances, Number of Conditions, and Population Value of Coefficient Alpha

Results were then collapsed across levels of *SDMean* and type of confidence interval (see Table 20). Each cell in this table is based on 15,000 samples. A graph of the relationship between population coefficient alpha, *SDVar*, and the proportion of the confidence intervals that included the population value was then constructed for each level of number of conditions (see Figures 1 and 2).

From Table 20 and Figures 1 and 2, it appears that the proportion of the confidence intervals that included the population value of coefficient alpha generally decreased as variances became more heterogeneous. As well, when unequal variances were present, this proportion decreased as the population value of coefficient alpha increased. Both of these effects were much greater for the 5-condition cells than for the 20-condition cells.

These results cannot be accounted for by the bias in the sample values of coefficient alpha, which were largest for small population values of coefficient alpha. Nor can they be accounted for by the slightly higher proportion of randomly generated sets of variances that contained a negative value, in the cells with high *SDVar* and ρ_{α}. Those sets of variances that were rejected appear, on average, to have had more heterogeneous variances than those that were retained. Heterogeneous variances appear to result in

Table 20

Proportion of the .95 Confidence Intervals that Included the Population Value of Coefficient Alpha, Collapsed Across Levels of SDMean and Type of Confidence Interval

Conditions	*SDVar*	Alpha	Proportion
5	0	.60	.948033
5	0	.75	.947950
5	0	.90	.947200
5	10	.60	.946100
5	10	.75	.932467*
5	10	.90	.885067*
5	15	.60	.937700*
5	15	.75	.912767*
5	15	.90	.828267*
20	0	.60	.949067
20	0	.75	.947667
20	0	.90	.946967
20	10	.60	.948567
20	10	.75	.948333
20	10	.90	.948033
20	15	.60	.944500*
20	15	.75	.941667*
20	15	.90	.941833*

con't

Table 20 *con't*

Conditions	*SDVar*	Alpha	Proportion
20	25	.60	.943567*
20	25	.75	.938867*
20	25	.90	.921600*

* indicates that the observed proportion was significantly ($p < .01$) different from the nominal proportion of .95.

Note. *SDMean* = the standard deviation of the means; Conditions = the number of conditions; *SDVar* = the standard deviation of the variances; Alpha = the desired population value of coefficient alpha; Proportion = the average proportion of the confidence intervals that included the population value of coefficient alpha. Proportions were based on 15,000 replications per cell.

Figure 1

Confidence Intervals Capturing Population Value

5-Condition Cells

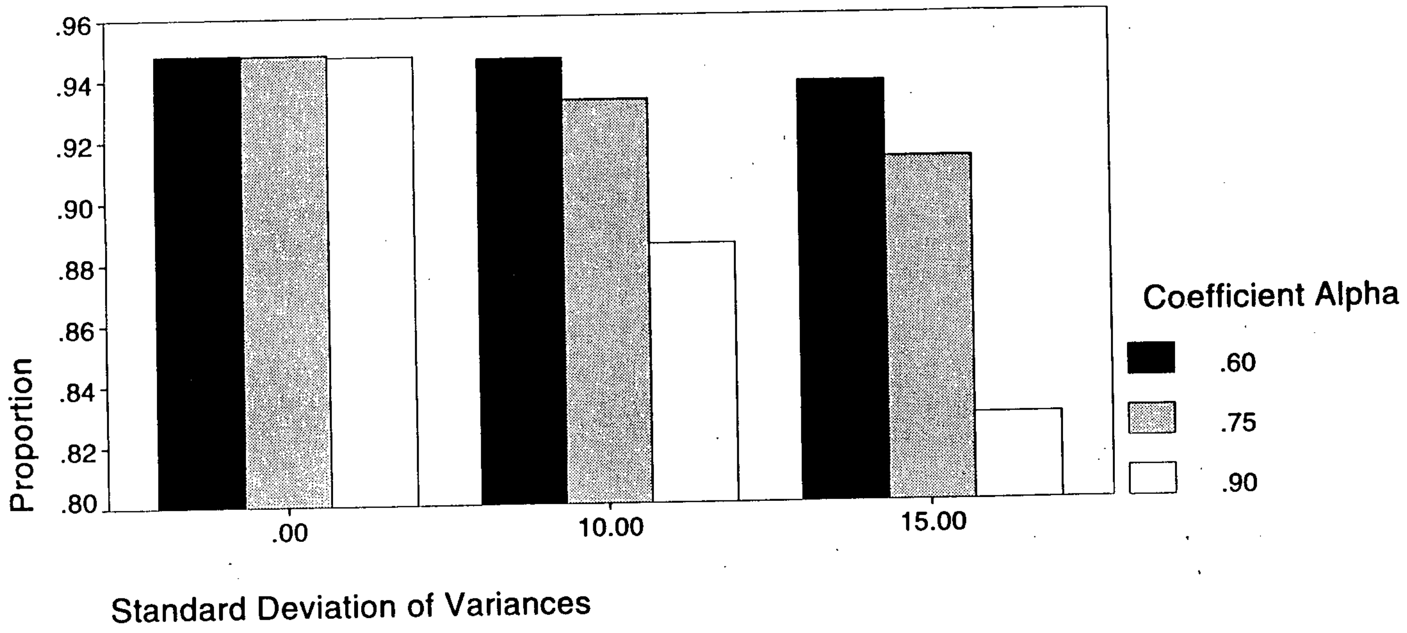
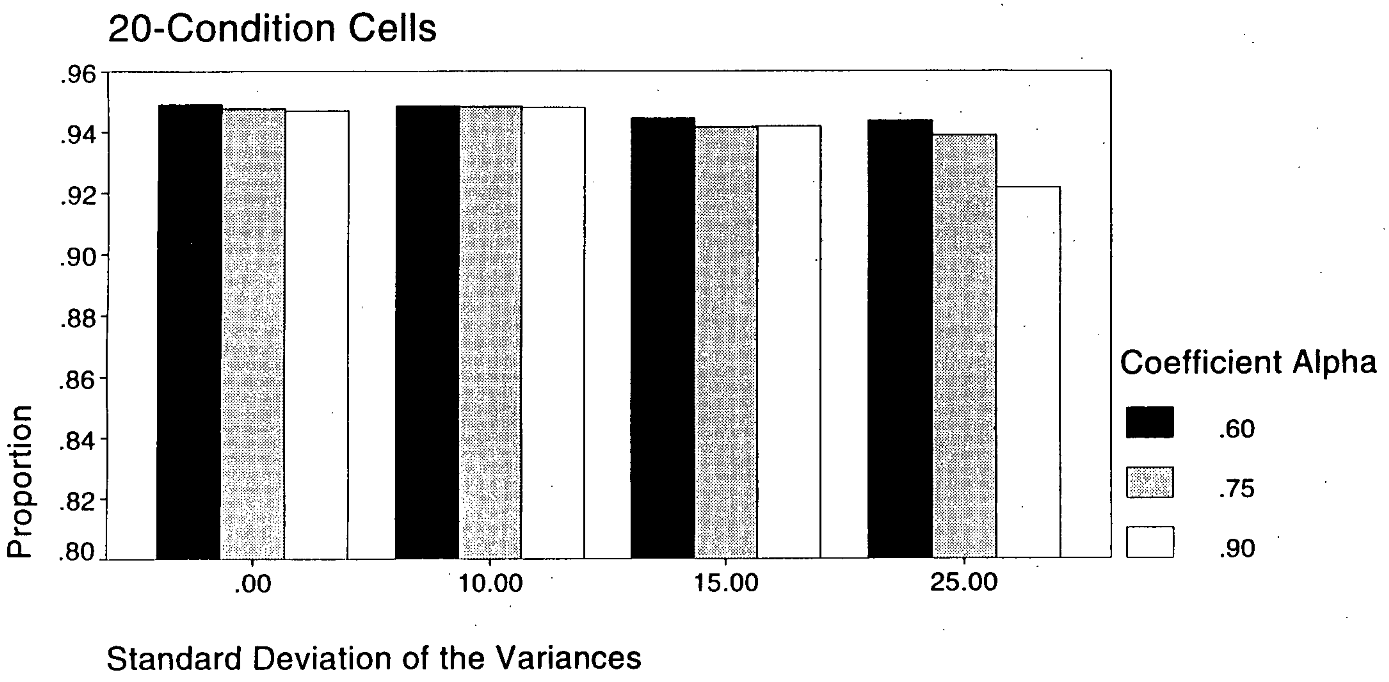


Figure 2

Confidence Intervals Capturing Population Value



fewer of the confidence intervals including ρ_α. Thus, cells where a larger proportion of the variance sets were rejected might be expected to contain ρ_α *more* frequently, not less frequently as was the case. In summary, neither data generation problems nor the sample bias of coefficient alpha can account for the findings of this study.

Discussion and Conclusions

In this study, parallel and essentially parallel data--data which met the assumption of compound symmetry--produced accurate confidence intervals. Tau-equivalent and essentially tau-equivalent data, on the other hand, did not produce accurate confidence intervals. Heterogeneous variances resulted in fewer of the confidence intervals capturing the population value. This problem was worst when measures were highly reliable (ρ_α = .90), and when the number of conditions was low (k = 5). When the number of conditions was high (k = 20), the confidence intervals performed reasonably well at all times. The least robust cell studied with k = 20 found .9174 of the confidence intervals included the population value of coefficient alpha. When the population value of coefficient alpha was .60, the confidence intervals also performed reasonably well, regardless of the degree of variance heterogeneity: in the least robust of the ρ_α = .60 cells studied, .9358 of the confidence intervals included the population value. Inequality of condition means had no effect on the performance of the confidence intervals.

The type of confidence interval used--the Hakstian and Whalen (1976) confidence interval or the Feldt (1965) confidence interval--had no effect on the performance of the confidence intervals. This is an important finding because the approximate confidence interval method developed by Hakstian and Whalen (1976) is easier to use than the Feldt confidence interval method. This is because the latter requires two-tailed critical F-values (the required degrees of freedom are not usually found in conventional tables), whereas the former requires only standard normal critical values.

Several data checks were done to ensure that the data were being generated as desired and that no confounds were being introduced. It was concluded that no data generation problem could account for the findings.

This study leads to the following conclusions. First, if condition variances are equal and the variance-covariance matrix among conditions is compound symmetric, then the confidence intervals will perform as intended. If condition variances are unequal, then the confidence intervals will perform fairly well if the number of conditions is large (20 or more), or the reliability is low (around .60). Otherwise, these confidence intervals must be considered highly suspect. Significantly fewer than 95% of the .95 confidence intervals will include the population value of coefficient alpha. This figure could run as low as 80%, or even lower for a highly reliable composite measure with few conditions.

GENERAL DISCUSSION AND CONCLUSIONS

Summary of Findings

This thesis began with three basic questions regarding the procedures proposed by Feldt (1965) and Hakstian and Whalen (1976) for setting confidence intervals for coefficient alpha. First, is the proportion of confidence intervals that include the population value of coefficient alpha the same for Type 1 and Type 12 sampling? Second, what are the assumptions underlying these confidence intervals and other inferential techniques for coefficient alpha? Third, how well do these confidence intervals perform with various kinds of multivariate normal data? Three Monte Carlo studies were conducted to explore these questions.

Study 1

In the first study, two kinds of data were simulated: spherical (but not compound symmetric) data and tau-equivalent data with unequal variances. The performance of the confidence intervals with each kind of data was compared under Type 1 and Type 12 sampling. The two types of sampling produced significantly different results in 14 of the 16 cells studied. Very close to 95% of the .95 confidence intervals included the population value of coefficient alpha under Type 1 sampling, with both kinds of data. For Type 12 sampling, however, the proportion was usually noticeably smaller than .95. This finding cannot be attributed to inadequate programming, as Type 12 sampling produced nominal level results with data meeting the requirements of parallelism.

Lord's (1955) work can be used to explain the finding that confidence intervals capture the population value less often under Type 12 sampling than under Type 1 sampling. As the reader will recall, Lord showed that the sampling variance of any statistic under Type 12 sampling is approximately equal to the sum of its sampling variance under Type 1 sampling and its sampling variance under Type 2 sampling. If a statistic's Type 2 sampling variance is greater than zero, then its Type 12 sampling variance will be larger than its Type 1 sampling variance. Because of this, sample values of coefficient alpha will be more heterogeneous under Type 12 sampling than under Type 1 sampling, and hence confidence intervals will capture the population value of coefficient alpha less often.

When will the Type 2 sampling variance of coefficient alpha be greater than zero? A statistic's Type 2 sampling variance *might* be greater than zero any time conditions are not statistically identical. However, one factor, differences in condition means, is known not to affect sample values of coefficient alpha, and hence will not affect its sampling variance. Thus, with the exception of allowing differences in condition means, any time conditions are not statistically identical, Type 12 sampling variance may be greater than Type 1 sampling variance, and the confidence intervals for coefficient alpha may capture the population value less often under Type 12 sampling than under Type 1 sampling. Future Monte Carlo research on inferential procedures for coefficient alpha should accommodate this finding: if conditions differ in any way except in mean values, then Type 1 sampling should be simulated if one is interested in fixed conditions, and Type 12 sampling should be simulated if one is interested in random conditions.

Study 2

In the second study, the assumptions underlying the confidence intervals were examined. Previously, some researchers have suggested that sphericity might be sufficient for these confidence intervals. However, since no theoretical support for this claim was found in the literature, a simulation experiment was conducted, using Type 12 sampling.

Two kinds of data were used. The first kind of data met the assumptions of Kristof (1963) and Feldt (1965), which imply that the variance-covariance matrix among conditions is compound symmetric. The second kind of data had a spherical (but not compound symmetric) variance-covariance matrix, but was otherwise matched with the first kind of data. The accuracy of the confidence intervals was assessed. Spherical data which were not compound symmetric did not produce accurate confidence intervals; in one cell, only 77% of the .95 confidence intervals included the population value. This contrasts markedly with the performance of the compound symmetric data: in all cells that used compound symmetric data, the proportion of confidence intervals that included the population value was within sampling error of nominal .95.

This study showed that sphericity is not sufficient to produce precise confidence intervals for coefficient alpha. Future research on the performance of inferential procedures for coefficient alpha should not focus on the effects of sphericity and non-sphericity. This study further suggests that compound symmetry may be necessary for these inferential procedures. The literature review found that, in order for coefficient alpha to have the sampling distribution that Kristof (1963) derived, the variance-covariance matrix among conditions must be spherical, the average correlation among the errors must be zero, and $v_{s.}$, the average of the entries in the s^{th} row or column of the variance-covariance matrix among the errors, must be constant. The last two of these requirements are violated by spherical data which are not also compound symmetric. In order that these three requirements all be met, it seems likely that the variance-covariance matrix must be compound symmetric, but no proof of this can be given at the present time.

Study 3

In the third study, the performance of the confidence intervals under four well-known measurement models was studied. The measurement models were parallelism, essential parallelism, tau-equivalence, and essential tau-equivalence. The accuracy of the confidence intervals under Type 12 sampling of each kind of data was assessed. Parallelism and essential parallelism resulted in accurate confidence intervals. This finding was expected, as these kinds of data meet the assumptions made by Feldt (1965) and Kristof (1963). Tau-equivalence and essential tau-equivalence, however, did not result in accurate confidence intervals: the confidence intervals captured the population value less often than desired.

The poor performance of the confidence intervals with tau-equivalent and essentially tau-equivalent data is not surprising. These kinds of data have unequal variances but independent errors. Hence, $v_{s.}$ is not constant, and the mean squares for people and errors are not independent. Furthermore, the variance-covariance matrices are non-spherical, and hence the mean square for errors is not distributed as a chi-square variate with the usual degrees of freedom. The parallel and essentially parallel data, on the other hand, have constant $v_{s.}$ and spherical variance-covariance matrices.

The effects of unequal condition means, unequal condition variances, type of confidence interval, number of conditions, and population value of coefficient alpha were also assessed. No differences were found in how well the Feldt (1965) and Hakstian and Whalen (1976) confidence intervals performed. Differences in condition means also had no affect on the confidence intervals. Heterogeneous variances did affect the performance of the confidence intervals; as variances became more heterogeneous, fewer of the confidence intervals captured the population value. This effect was most pronounced when the population value of coefficient alpha was high, and when the number of conditions was low. However, as the number of conditions increased, the confidence intervals generally became more accurate. In the worst cell with 20 conditions, 91.74% of the .95 confidence intervals captured the population value, compared to 82.64% in the worst cell with five conditions.

Implications for the Use of Confidence Intervals for Coefficient Alpha

Researchers interested in using confidence intervals for coefficient alpha might have two questions. First, is one method better than the other? Second, when can the confidence intervals be used?

The answer to the first question is No. The Feldt (1965) and Hakstian and Whalen (1976) confidence interval methods gave nearly identical results in every cell studied. They can be used interchangeably. The Hakstian and Whalen (1976) method may be preferred because it is somewhat easier to use: it requires only standard normal critical values, not critical values from a central F-distribution, as does the Feldt (1965) method.

The answer to the second question is more complicated. Study 1 suggests that if a researcher is interested in coefficient alpha as an index of the internal consistency of a fixed set of conditions, then the two confidence intervals may work quite well, even when assumptions are violated. This finding cannot be explained at this time. It is true, of course, that with Type 1 sampling, conditions *are* statistically identical *across* samples, although not necessarily *within* samples: the same, possibly heterogeneous, set of conditions are given to each sample of people. Heterogeneity of conditions across samples may lead to greater sampling variance of coefficient alpha, and hence to fewer of the confidence intervals including the population value. It may be this heterogeneity, and not heterogeneity within samples, that is most important. However, there is nothing in the derivation of the confidence intervals which suggests why this should be the case. More detailed studies of the performance of these confidence intervals under Type 1 sampling need to be conducted with more varied kinds of data to ensure their robustness under a wide variety of situations. As well, theoretical work is needed to understand the robustness of these confidence intervals under Type 1 sampling.

In contrast, when conditions are random, Studies 2 and 3 have shown that researchers must be careful when using the two confidence interval methods. The only kinds of data which always resulted in precise confidence intervals were parallel and essentially parallel data. It will be recalled that all data which are parallel are also essentially parallel. Thus, if one is dealing with coefficient alpha in the context of random conditions, one should test to see whether one's data meet the assumptions of essential parallelism. When data are essentially parallel, errors are independent and equally variable, and true scores differ only by a constant (if at all). Because of this, the variance-covariance matrix must be compound symmetric. Wilk (1946) developed a test of compound symmetry. Let

$$L = \frac{|S|}{(s^2)^k (1-r)^{k-1} [1+(k-1)]r}, \tag{7}$$

where $|S|$ is the determinant of the sample variance-covariance matrix, s^2 is the average sample variance, r is ratio of the average sample covariance to the average sample variance, and k is the number of conditions. Wilk showed that $-n\ln(L)$ is distributed as a χ^2 with f_1 degrees of freedom, for large samples, where f_1 = $(k^2 + k - 4)/2$ and n is the number of subjects. A modified version of this test was given by Box (1949, 1950), who showed that $-C\ln(L)$ is distributed as a χ^2 with f_1 degrees of freedom, where C =

$(n-1)-\frac{k(k+1)^2(2k-3)}{6(k-1)(k^2+k-4)}$. If this test is significant, then the researcher should conclude that the conditions are not essentially parallel. However, if conditions do satisfy the assumption of essential parallelism (i.e., the test is non-significant), then the confidence intervals can be expected to perform as intended.

This thesis also shed light on the performance of these confidence intervals when conditions are not essentially parallel, if errors are independent. Therefore, if data are not essentially parallel, one should test whether the assumption of independent errors is reasonable. Because errors are assumed to be multivariate normal, a test of their independence is equivalent to a test of their lack of intercorrelation. However, neither true scores nor errors are directly observable, and the hence testing the hypothesis that all correlations among errors are zero is complicated. If conditions are random, then, in the population, the average covariance among the errors is zero (Cronbach, 1995), and hence a test of the equality of the covariances is equivalent to a test of the independence of the errors. This is a pattern hypothesis that can be tested using the SePath (Structural Equation / Path Analysis) module of STATISTICA (StatSoft, 1995). However, if conditions are fixed the hypothesis of uncorrelated errors can not be fully tested. This hypothesis can be broken down into two subcomponents: the hypothesis that the average covariance among the errors is zero, and the hypothesis that the covariances among the errors are equal. It is the first of these two subcomponents that can not be tested. Novick and Lewis (1967) showed that the average correlation among errors for fixed conditions will be *estimated* to be -1/(k-1), regardless of the parameter value of the average correlation. Therefore, a test of the hypothesis that the average covariance among the errors is zero is therefore not possible. However, the second subcomponent can be tested using SePath, as above. If this hypothesis is rejected, then the assumption of independent errors is untenable. However, if this latter hypothesis is not rejected, we still cannot be sure we have independent errors, as the assumption that the average covariance among the errors is zero cannot be tested.

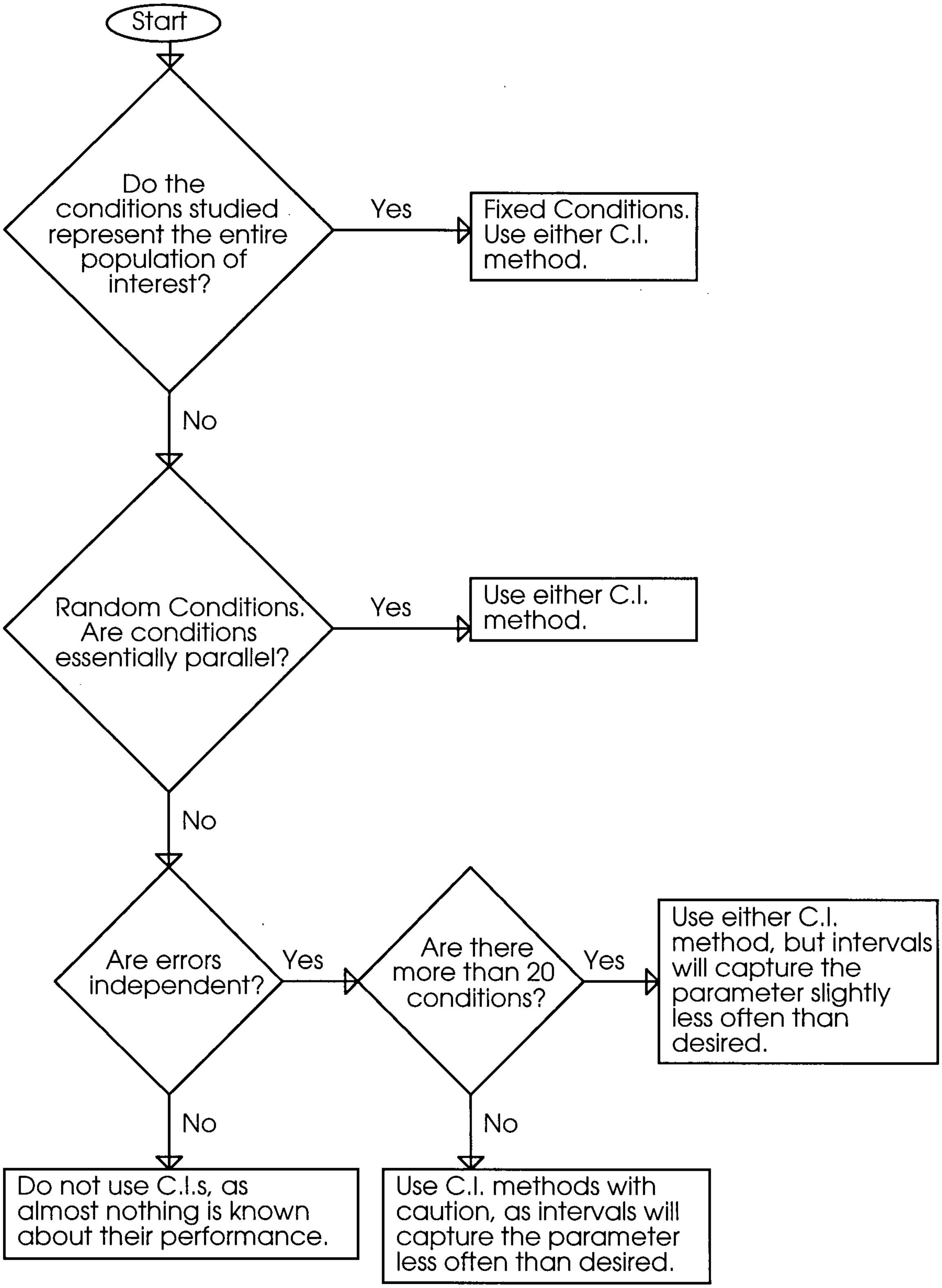
If conditions are not essentially parallel, but the test for equal covariance among the errors cannot be rejected, then the performance of these confidence intervals can be predicted. If there are 20 or more conditions, the confidence intervals will likely perform reasonably well. Probably slightly fewer than 95% of the .95 confidence intervals will include the population value (unless the population value of coefficient alpha is above .90, this percentage will not likely fall below 92%). If a researcher has fewer than 20 measures, the confidence intervals can be expected to perform adequately only when the population value of coefficient alpha is quite low.

On the other hand, if independent errors cannot be assumed (i.e., the test for equal covariances among is rejected), then these confidence interval methods should not be used at all. Correlated errors may occur for a variety of reasons. For example, serial correlations--where subjects receive more similar scores on adjacent conditions--may be observed any time conditions are administered in the same order to every subject. Correlated errors may also occur when more than one factor underlies performance on different conditions. More research is needed to determine what effect correlated errors have on the performance of these confidence intervals.

The process of deciding when to use confidence intervals for coefficient alpha is summarized in flowchart form in Figure 3.

Figure 3

Deciding on Appropriate Use of Confidence Intervals for Coefficient Alpha


Start
Do the conditions studied represent the entire population of interest?
Yes
Fixed Conditions. Use either C.I. method.
No
Random Conditions. Are conditions essentially parallel?
Yes
Use either C.I. method.
No
Are errors independent?
Yes
Are there more than 20 conditions?
Yes
Use either C.I. method, but intervals will capture the parameter slightly less often than desired.
No
Do not use C.I.s, as almost nothing is known about their performance.
No
Use C.I. methods with caution, as intervals will capture the parameter less often than desired.

Future Research

Future research could profitably explore three areas. First, some research on the robustness of these confidence interval methods to violation of other assumptions has been done, but this work should be expanded. Research with non-normal data was done by Bay (1973), Hakstian and Whalen (1976), and Feldt (1965). Bay (1973) examined the effects of non-normality on the sampling distribution of coefficient alpha, under Type 1 sampling. He found, for example, that "[n]on-zero kurtosis of the true score distribution substantially affects the sampling distribution and standard error of reliability estimates" (p. 57). Hakstian and Whalen (1976) and Feldt (1965) showed that dichotomous data with unequal difficulty levels and variances result in quite accurate confidence intervals under Type 1 sampling, when 20 or more conditions are used. These studies should be repeated under Type 12 sampling.

No published research has explored the effects of violation of the assumption of independent errors. Correlated errors will likely preclude the condition that the average covariance among the errors be constant for any row or column in the error variance-covariance matrix. Because of this, $MS_{persons}$ and MS_{error} will be dependent (Box, 1954), and thus the performance of the confidence intervals may be affected.

Some evidence of the effect of correlated errors on the performance of these confidence intervals can be found from this thesis. Recall the results in Study 2 for spherical data with *SDMean* 20 and *SDVar* 10. The proportion of the confidence intervals that included the population value *decreased* as coefficient alpha decreased and as the number of conditions increased. These are the exact opposite of the results for the essentially tau-equivalent data with *SDMean* 20 and *SDVar* 10 from Study 3. For these latter data, the proportion of the confidence intervals that included that population value *increased* as coefficient alpha decreased and as the number of conditions increased. The only difference between these two data sets was that the spherical data had unequal covariances. Recall that true differences between people can account for constant positive covariances among conditions. Unequal population covariances are the result of correlated errors. This example shows how important the assumption of independent errors is. Monte Carlo studies and theoretical work could be done to determine how correlated errors affect the performance of the confidence intervals, both in isolation and in combination with other assumption violations.

The second area future research could explore is the effect of assumption violation on other inferential procedures for coefficient alpha. Some limited research in this area has been done (e.g., Feldt, 1969; 1980; Woodruff & Feldt, 1986). However, these studies have used Type 1 sampling, and research involving Type 12 sampling is still needed. The expected performance of these procedures under Type 12 sampling can to some extent be inferred from the findings of this and other studies of confidence intervals for coefficient alpha under Type 12 sampling. Thus, for example, other inferential procedures for coefficient alpha will likely not be robust to heterogeneity of variance and correlated errors, under Type 12 sampling, but the extent of the problem will not be known until each procedure is tested in turn.

The third area future research could explore is the development of more robust inferential techniques for coefficient alpha. Three approaches to doing this are briefly explored below. Recall that, in Study 2, the size of the bias in the sample estimates of coefficient alpha appeared to correspond to the degree of inaccuracy in the confidence intervals for coefficient alpha. This suggested that quantifying the degree of bias involved in sampling of conditions, and adjusting the estimate of coefficient alpha before applying the confidence intervals might make the confidence intervals somewhat more accurate.

However, in Study 3, the size of bias did not correspond to the degree of inaccuracy of the confidence intervals, and hence this technique would not be uniformly helpful.

Alternative approaches to making the confidence intervals more robust are suggested by the fact that, when the confidence intervals were not precise under Type 12 sampling, they captured the population value *less often* than desired. It is possible that reducing the degrees of freedom associated with the confidence intervals might widen the confidence intervals appropriately, and result in greater robustness. A similar problem is sometimes encountered in the *k*-group ANOVA, where unequal variances and unequal group sizes result in the Behren's-Fisher problem. In this situation, the degrees of freedom of MS_{within} is reduced using Satterthwaite's (1941) degrees of freedom. Thus, the second approach to making the confidence intervals more robust is to use a chi-square distribution with modified degrees of freedom to approximate the distribution of MS_{error} when error variances are heterogeneous. However, an examination of this method shows that the Satterthwaite degrees of freedom are not directly applicable in the repeated-measures case for two reasons. First, the maximum Satterthwaite degrees of freedom are *k*(*n*-1)--corresponding to the unadjusted degrees of freedom of MS_{within} --not (*k*-1)(*n*-1)--corresponding to the unadjusted degrees of freedom of MS_{error}. Second, the Satterthwaite degrees of freedom are based on the assumption that the variance of the composite (in this case MS_{error}) is the sum of the variance of its components (the error variance for the conditions). This is not an assumption which is easily justified when components are measured on the same people and are likely to covary. Furthermore, if some modified degrees of freedom could be developed for MS_{error}, MS_{error} and MS_{people} would still likely be dependent when variances are unequal or errors correlated (Box, 1954). Substantial effort would therefore be needed to apply this approach to the confidence intervals for coefficient alpha.

Making an inferential procedure more robust by reducing the degrees of freedom is done in a second context as well. In repeated-measures ANOVA, the F-test for conditions requires sphericity. When sphericity is not present, both degrees of freedom are reduced by a correction factor, epsilon, which is a measure of the degree of non-sphericity (Box, 1954). Thus, a third approach to developing more robust confidence intervals for coefficient alpha would be to quantify the degree of assumption violation and use that as a correction for both degrees of freedom. When variances are equal and errors independent, as assumed by Kristof (1963) and Feldt (1965), the variance-covariance matrix among conditions is compound symmetric. Wilk's (1946) generalized likelihood ratio statistic for testing compound symmetry, *L*, is given by Equation (7). This statistic equals 1 when the sample variance-covariance matrix is precisely compound symmetric, and has a minimum value of 0. *L* could be used directly to adjust the numerator and denominator degrees of freedom, but likely would provide a poor fit. Monte Carlo studies could be used to examine various transformations of *L*, to find one which provides sufficiently precise confidence intervals under a wide variety of situations.

Final Remarks

Inferential procedures for coefficient alpha are, at present, an under-utilized research tool. Further research with them is needed to determine when their use should be encouraged, and what steps can be taken to make them more robust. This thesis has added to efforts to do this in four ways. First, this thesis has shown that Type 1 and Type 12 sampling produce different results and should be treated separately, and provides a model of how Type 12 sampling can be simulated. Second, this thesis has shown that sphericity is not a sufficient assumption for these inferential techniques, and should not be the basis of robustness studies. Third, this thesis has shown that the two confidence interval methods studied produce very similar results, and can be used interchangeably. Finally, this thesis has shown that current confidence interval methods for coefficient alpha are not, in general, robust to violation of the assumptions of homogeneous variances and independent errors. Other inferential techniques for coefficient alpha are based on the same sampling distribution, and hence likely also lack robustness. Thus, this thesis has indicated that more robust procedures need development, and one possible approach to doing this has been identified.

REFERENCES

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

Box, G.E.P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika, 35*, 317-346.

Box, G.E.P. (1950). Problems in the analysis of growth and wear curves. *Biometrika, 6*, 362-389.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484-498.

Brennan, R.L. (1983). *Elements of Generalizability Theory*. Iowa: ACT.

Cornfield, J., & Tukey, J.W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 17*, 907-949.

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

Cronbach, L.J. (1995). *Personal Communication.* October 4

Cronbach, L.J., Rajaratnam, N., & Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*, 137-163.

Eom, H.J. (1993). *The interactive effects of data categorization and noncircularity on the sampling distribution of generalizability coefficients in analysis of variance models: An empirical investigation.* Unpublished doctoral dissertation, University of British Columbia, Vancouver.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 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*, 99-105.

Feldt, L.S. (1992). Test of the hypothesis that the interclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement, 16*, 195-205.

Gaito, J. (1961). Repeated measurements designs and counterbalancing. *Psychological Bulletin, 58,* 46-54.

Gifi, A. (1990). *Non-linear multivariate analysis.* New York: Wiley.

de Gruijter, D.N.M., & van der Kamp, L.J.Th. (1984). *Statistical models in Psychological and Educational Testing.* Lisse: Swets & Zeitlinger.

Gulliksen, H. (1950). *Theory of Mental Tests.* New York: Wiley & Sons.

Hakstian, A.R., & Whalen, T.E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160.

Huynh, H., & Feldt, L.S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association, 65,* 1582-1589.

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.

Kristof, W. (1964). Testing differences between reliability coefficients. *The British Journal of Mathematical and Statistical Psychology, 17,* 105-111.

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

Lord, F.M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika, 20*, 1-22.

Lord, F.M., & Novick, M.R. (1968*). Statistical Theories of Mental Test Scores.* Reading, Massachusetts: Addison-Wesley.

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

Paulson, E. (1942). An approximate normalization of the analysis of variance distribution. *Annals of Mathematical Statistics, 13*, 233-235.

Rouanet, H., & Lepine, D. (1970). Comparison between treatments in a repeated-measures design: ANOVA and multivariate methods. *The British Journal of Mathematical and Statistical Psychology, 23*, 147-163.

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

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika, 6*, 309-316.

Schroeder, M.L., & Hakstian, A.R. (1990). Inferential procedures for multifaceted coefficients of generalizability. *Psychometrika, 55*, 429-447.

Sedere, M.U., & Feldt, L.S. (1976). The sampling distributions of the Kristof reliability coefficient, the Feldt coefficient, and Guttman's lambda-2. *Journal of Educational Measurement, 14*, 53-62.

StatSoft, Inc. (1995). *STATISTICA for Windows* [Computer program manual]. Tulsa, OK: StatSoft, Inc.

Werts, C.E., Grandy, J., & Schabacker, W.H. (1980). A confirmatory approach to calibrating congeneric measures. *Multivariate Behavioral Research, 15*, 109-122.

Wilks, S.S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution, *Annals of Mathematical Statistics, 17*, 257-281.

Winer, B.J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Winer, B.J. (1971). *Statistical principles in experimental design (2nd ed.).* New York: McGraw-Hill.

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

Notation

r_α	sample value of coefficient alpha
ρ_α	population value of coefficient alpha
n	number of subjects
k	number of conditions
AveVar	average of condition variances
SDVar	standard deviation of condition variances
VarVar	variance of condition variances
AveMean	average of condition means
SDMean	standard deviation of condition means
VarMean	variance of condition variances
TVar	true score variance or variance due to true differences between people
AveErr	average error variance
Σ	population variance-covariance matrix

APPENDICES

Appendix A:

Feldt's (1965) Confidence Interval Method

Feldt (1965) showed that a 100 (1-α) % confidence interval for coefficient alpha is given by $\left[1-(1-\alpha)F_b, 1-(1-\alpha)F_a\right]$ where F_a and F_b are respectively the 100(1-$\alpha/2$) and 100($\alpha/2$) percentile points of the central F-distribution with (n-1) and (n-1)(k-1) degrees of freedom. His proof is given below.

Feldt began his proof with a relationship known from repeated measures ANOVA:

$$\frac{MS_{persons} \Big/ (\sigma_e^2 + k\sigma_t^2)}{MS_{error} \Big/ \sigma_e^2} \approx F_{(n-1),(n-1)(k-1)},$$

where $MS_{persons}$ is the mean square due to people, and equals $\frac{\sum_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2}{n-1}$, and MS_{error} is the mean square to due error, and equals $\frac{\sum_j (\bar{x}_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2}{(n-1)(k-1)}$, n is the number of subjects, k is the number of conditions, σ_e^2 is the error variance and σ_t^2 is the true score variance or the variance due to people.

He referred to the ratio $\frac{MS_{persons}}{MS_{error}}$ as F_{obs}, and the ratio $\frac{\sigma_e^2 + k\sigma_t^2}{\sigma_e^2}$, which equals $\frac{E[MS_{persons}]}{E[MS_{error}]}$, as F_{pop}. Then $\frac{Fobs}{Fpop} \approx F_{(n-1),(n-1)(k-1)}$. Therefore,

(8) $$P[F_a < \frac{F_{obs}}{F_{pop}} < F_b]=1-\alpha$$

where F_a and F_b are respectively the $100(1\text{-}\alpha/2)$ and $100(\alpha/2)$ percentile points of the central F-distribution with (n-1) and (n-1)(k-1) degrees of freedom. Equation (8) can be rearranged as

$$P[\frac{1}{F_b} < \frac{F_{pop}}{F_{obs}} < \frac{1}{F_a}]=1-\alpha$$

$$P[\frac{F_{obs}}{F_b} < F_{pop} < \frac{F_{obs}}{F_a}]=1-\alpha .$$

Then, using the relationship $F_{obs} = \frac{MS_{persons}}{MS_{error}} = \frac{1}{1-r_\alpha}$, and defining $\rho_\alpha = 1 - \frac{1}{F_{pop}}$, we have

$$P[\frac{1}{(1-r_\alpha)F_b} < \frac{1}{(1-\rho_\alpha)} < \frac{1}{(1-r_\alpha)F_a}]=1-\alpha$$

$$P[1-(1-r_\alpha)F_b < \rho_\alpha <1-(1-r_\alpha)F_a]=1-\alpha$$

Feldt's proof is somewhat circuitous. The same confidence interval method can be derived more simply. Kristof (1963) showed $\frac{1-\rho_\alpha}{1-r_\alpha} \approx F_{(n-1),(n-1)(k-1)}$. Thus,

(9) $$P[F_a < \frac{1-\rho_\alpha}{1-r_\alpha} < F_b]=1-\alpha ,$$

where F_a and F_b are defined as above. Equation (9) can be rearranged as

$$P[(1-r_\alpha)F_a < (1-\rho_\alpha) < (1-r_\alpha)F_b] = 1-\alpha$$

$$P[1-(1-r_\alpha)F_b < \rho_\alpha < 1-(1-r_\alpha)F_a] = 1-\alpha$$

Feldt (1965) stated the following assumptions, during the derivation of these confidence intervals:

(i) The score of subject *i* (*i*=1, ... *N*) on item *j* (*j*=1, ... *k*) may be represented as $X_{ij} = \mu + a_j + t_i + e_{ij}$, where the notation is defined as follows.

μ = the mean item score over the entire population of subjects and items.

a_j = the amount by which the mean examinee score on item *j* deviates from μ. The quantity a_j reflects, in deviation form, the relative difficulty of the item for the population of examinees.

Since a_j is a deviation score, $\sum_{j=1}^{\infty} a_j = 0$

t_i = the amount by which the mean item score for examinee *i* deviates from μ. The quantity t_i reflects, in deviation form, the true ability of the examinee in the domain defined by the population of items. Since t_i is a deviation score, $\sum_{i=1}^{\infty} t_i = 0$

e_{ij} = the interaction effect of item *j* with subject *i*, an effect presumed wholly the result of measurement error. For examinee *i*, $\sum_{i=1}^{\infty} e_{ij} = 0$; for item *j*, $\sum_{j=1}^{\infty} e_{ij} = 0$.

(ii) The *N* subjects are assumed to be a random sample from the examinee population.

(iii) The *k* items are assumed to be a random sample from the population of items.

(iv) Over the entire population of examinees, the quantity t_i is assumed to be normally distributed.

(v) Over the entire examinees-by-item matrix, the e_{ij} are assumed normally distributed, independently of each other and of t_i.

(vi) For any infinite subpopulation of examinees and items, σ_e^2 is assumed equal to σ_e^2 for any other subpopulation.

(p. 359)

Appendix B:

Hakstian and Whalen's (1976) Confidence Interval Method

Hakstian and Whalen (1976) showed:

$$P\{1-c^{*3}[(1-r_{\alpha})^{1/3}+Z_{1-\alpha/2}\sigma]^3<\rho_{\alpha}<$$
$$1-c^{*3}[(1-r_{\alpha})^{1/3}-Z_{1-\alpha/2}\sigma]^3\}\cong 1-\alpha$$

where $\sigma=\dfrac{18k(n-1)(1-r_{\alpha})^{2/3}}{(k-1)(9n-11)}$, and $c^{*}=\dfrac{(9n-11)(k-1)}{9(n-1)(k-1)-2}$.

Proof:

They began with the relationship, known from repeated measures ANOVA, that

(10) $$\frac{1-\rho_{\alpha}}{1-r_{\alpha}}\approx F_{(n-1),(n-1)(k-1)}.$$

Using Paulson's (1942) normalizing transformation for F-variables, Equation (10) implies that

(11) $$(CF^{1/3}-\mu^{*})(\sigma^{*)})^{-1}\cong N(0,1),$$

where

$$C=\frac{9n-11}{9(n-1)},$$

$$\mu^{*}=1-\frac{2}{9(n-1)(k-1)},$$

and the symbol "$\cong$" is to be read "is approximately distributed as". From Equation (11) it follows that

$$(1-r_{\alpha})^{1/3} \cong N(\mu,\sigma^{2}), \tag{12}$$

where

$$\mu = C^{-1}(1-\rho_{\alpha})^{1/3}\mu^{*}, \tag{13}$$

and

$$\sigma^{2} = 2kC^{-1}(1-\rho_{\alpha})^{2/3}[9(n-1)(k-1)]^{-1}, \tag{14}$$

n = the number of subjects, and k = the number of conditions.

The expression for the variance (14) includes the unknown population parameter of coefficient alpha. The sample value, r_{α}, is substituted in, and once simplified, the variance estimate is given by

$$\sigma^{2} = \frac{18k(n-1)(1-r_{\alpha})^{2/3}}{(k-1)(9n-11)}. \tag{15}$$

From (12), (13) and (15), confidence intervals for coefficient alpha can be given:

$$\Pr(1-[C^{*^{3}}[1-r_{\alpha})^{1/3}+z_{1-\alpha/2}\sigma]^{3} < \rho_{\alpha} < 1-[C^{*^{3}}[1-r_{\alpha})^{1/3}-z_{1-\alpha/2}\sigma]^{3}\} \cong 1-\alpha,$$

where

$$C^{*} = \frac{C}{\mu^{*}} = \frac{(9n-11)(k-1)}{9(n-1)(k-1)-2}.$$

Hakstian and Whalen (1976) stated that they were making the assumptions of the two-way random effects ANOVA design, but did not clarify further.

Appendix C:

Preliminary Studies of Spherical Data

The percentage of the randomly generated matrices under Type 12 sampling that were not positive definite and the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha, for Spherical data.

Conditions	*SDVar*	Alpha	Not Positive-Definite	Include
10	10	.60	0%	.838
10	10	.75	0%	.906
10	10	.90	0%	.918
10	15	.60	.4%	.710
10	15	.75	0%	.836
10	15	.90	0%	.914
10	20	.60	7.5%	.662
10	20	.75	1.1%	.766
10	20	.90	0%	.888
15	10	.60	.2%	.746
15	10	.75	0%	.890
15	10	.90	0%	.920

con't

Conditions	*SDVar*	Alpha	Not Positive-Definite	Include
15	15	.60	9.1%	.698
15	15	.75	1.4%	.814
15	15	.90	0%	.910
15	20	.60	51%	.602
15	20	.75	28.3%	.766
15	20	.90	2.9%	.868
20	5	.60	0%	.894
20	5	.75	0%	.930
20	5	.90	0%	.952
20	10	.60	1.18%	.780
20	10	.75	0%	.862
20	10	.90	0%	.922

Note. The average variance was 100, the standard deviation of condition means was 20, and the sample size was 100. Conditions = the number of conditions; SDVar = the standard deviation of variances; Alpha = the population value of coefficient alpha; Not Positive-Definite = the percentage of randomly-generated matrices that were not positive-definite; Include = the proportion of Feldt (1965) confidence intervals that included the population value. Five hundred replications were used per cell.

Appendix D:

Preliminary Studies of Essentially Tau-Equivalent Data

The percentage of the randomly generated sets of unequal variances under Type 12 sampling that included at least one negative value and the proportion of the Feldt (1965) confidence intervals that included the population value of coefficient alpha, for Essentially Tau-Equivalent data.

Conditions	*SDVar*	Alpha	Negative	Include
5	10	.60	0%	.954
5	10	.75	0%	.942
5	10	.90	0%	.864
5	15	.60	0%	.938
5	15	.75	0%	.910
5	15	.90	3.8%	.812
5	20	.60	0%	.916
5	20	.75	.6%	.906
5	20	.90	20.2%	.812
7	15	.60	0%	.950
7	15	.75	0%	.918
7	15	.90	1%	.878

con't

Conditions	*SDVar*	Alpha	Negative	Include
7	20	.60	0%	.930
7	20	.75	.2%	.908
7	20	.90	11.5%	.802
10	10	.60	0%	.962
10	10	.75	0%	.954
10	10	.90	0%	.940
10	15	.60	0%	.938
10	15	.75	0%	.928
10	15	.90	.2%	.908
10	20	.60	0%	.938
10	20	.75	0%	.930
10	20	.90	3.8%	.900
15	10	.60	0%	.954
15	10	.75	0%	.956
15	10	.90	0%	.916
15	15	.60	0%	.952
15	15	.75	0%	.952
15	15	.90	0%	.942

con't

Conditions	*SDVar*	Alpha	Negative	Include
15	20	.60	0%	.934
15	20	.75	0%	.940
15	20	.90	1%	.910
20	10	.60	0%	.946
20	10	.75	0%	.940
20	10	.90	0%	.934
20	15	.60	0%	.950
20	15	.75	0%	.942
20	15	.90	0%	.950
20	20	.60	0%	.950
20	20	.75	0%	.934
20	20	.90	.4%	.934
20	25	.60	0%	.944
20	25	.75	1.99%	.950
20	25	.90	4.9%	.906

Note. The average variance was 100, the standard deviation of condition means was 20, and the sample size was 100.

Conditions = the number of conditions; SDVar = the standard deviation of variances; Alpha = the population value of coefficient alpha; Negative = the percentage of randomly-generated sets of variances that included at least one negative value; Include = the proportion of Feldt (1965) confidence intervals that included the population value. Five hundred replications were used for each cell.

Appendix E

Actual Population Values of Coefficient Alpha, the Average and Variance of the Condition Variances, and the Average and Variance of Condition Means for Type 1 Sampling, for Spherical Data - Study 1

		Actual Values				
Conditions	Desired Alpha	Alpha	AveVar	VarVar	AveMean	VarMean
5	.60	.6046	100.45	99.58	-.2912	402.91
5	.90	.9005	100.36	99.16	-.4601	397.52
20	.60	.5999	100.00	100.04	-.1451	397.81
20	.90	.8996	99.88	100.33	-.0014	402.63

Note. Desired values were *AveVar = 100, SDVar = 10 (VarVar = 100), AveMean = 0, SDMean = 20 (VarMean = 400).* Conditions = the number of conditions; Desired Alpha = the desired value of coefficient alpha in the population; Alpha = the sample value of coefficient alpha; *AveVar* = the average of condition variances; *VarVar* = the variance of the condition variances; *AveMean* = the average of condition means; *VarMean* = the variance of the condition means.

Appendix F

Average Across Replications of the Average and Variance of Condition Variances and the Average and Variance of the Condition Means, for Type 12 Sampling Conditions, for Spherical Data - Study 1

		Averages Across Replications			
Conditions	Alpha	*AveVar*	*VarVar*	*AveMean*	*VarMean*
5	.60	99.97	99.76	-.0098	400.41
5	.90	100.02	99.81	-.0588	396.32
20	.60	100.05	99.24	-.0300	399.63
20	.90	100.00	100.14	-.0104	401.60

Note. Desired values were *AveVar = 100, SDVar = 10 (VarVar = 100), AveMean = 0, SDMean = 20 (VarMean = 400).* Conditions = the number of conditions; Alpha = the population value of coefficient alpha; *AveVar* = the average of condition variances; V*arVar* = the variance of the condition variances; *AveMean* = the average of condition means; *VarMean* = the variance of the condition means.

Appendix G

Actual Population Values of Coefficient Alpha, the Average and Variance of the Condition Variances, for Type 1 Sampling, for Tau-Equivalent Data - Study 1

		Actual Values		
Conditions	Desired Alpha	Alpha	*AveVar*	*VarVar*
5	.60	.6008	99.75	223.66
5	.90	.9000	99.98	225.54
20	.60	.5997	100.13	226.75
20	.90	.8996	100.32	223.35

Note. Desired values were *AveVar = 100, SDVar = 15 (VarVar = 225), AveMean = 0, SDMean = 0 (VarMean = 0).* Conditions = the number of conditions; Desired Alpha = the desired population value of coefficient alpha; Alpha = the sample value of coefficient alpha; *AveVa*r = the average of condition variances; *VarVar* = the variances of condition variances.

Appendix H

Average Across Replications of the Average and Variance of Condition Variances, for Type 12 Sampling Conditions, for Tau-Equivalent Data - Study 1

		Averages Across Replications	
Conditions	Alpha	AveVar	VarVar
5	.60	100.02	225.54
5	.90	100.38	211.42
20	.60	100.02	225.39
20	.90	100.05	225.37

Note. Desired values were *AveVar = 100, SDVar = 15 (VarVar = 225), AveMean = 0, SDMean = 0 (VarMean = 0).* Conditions = the number of conditions; Alpha = the population value of coefficient alpha; *AveVar* = the average of condition variances; *VarVar* = the variances of condition variances.

Appendix I

Average Across Replications of the Average and Variance of Condition Means, for Compound Symmetric Data - Study 2

		Averages Across Replications	
Conditions	Alpha	*AveMean*	*VarMean*
5	.60	-.0346	399.49
5	.75	-.0356	396.26
5	.90	-.0246	401.96
20	.60	-.0353	400.47
20	.75	-.0190	399.40
20	.90	.0116	399.59

Note. Desired values were *AveVar = 100, SDVar = 0, AveMean = 0, SDMean = 20 (VarMean =400).* Conditions = the number of conditions; Alpha = the population value of coefficient alpha; *AveVar* = the average of condition variances; *VarVar* = the variances of condition variances; *AveMean* = the average of condition means; *VarMean* = the variance of condition means.

Appendix J

Proportion of Randomly Generated Matrices That Were Not Positive Definite, for Spherical Data that were not Compound Symmetric - Study 2

Conditions	Alpha	Not Positive-Definite
5	.60	0
5	.75	0
5	.90	0
20	.60	.0137
20	.75	.0001
20	.90	0

Note. Conditions = the number of conditions; Alpha = the population value of coefficient alpha; Not Positive-Definite = the proportion of randomly-generated matrices that were not positive-definite and hence unusable.

Appendix K

Average Across Replications of the Average and Variance of Condition Variances and the Average and Variance of the Condition Means, for Spherical (but not Compound Symmetric) Data - Study 2

		Averages Across Replications			
Conditions	Alpha	*AveVar*	*VarVar*	*AveMean*	*VarMean*
5	.60	100.03	99.91	.0580	397.49
5	.75	99.97	99.96	-.0190	402.53
5	.90	100.04	99.40	-.0150	398.92
20	.60	100.05	99.03	-.0581	400.41
20	.75	100.000	100.16	-.0239	401.37
20	.90	99.98	100.16	.0029	400.33

Note. Desired values were *AveVar = 100, SDVar = 10 (VarVar = 100), AveMean = 0, SDMean = 20 (VarMean = 400).* Condition = the number of conditions; Alpha = the population value of coefficient alpha; *AveVar* = the average of condition variances; *VarVar* = the variance of the condition variances; *AveMean* = the average of condition means; *VarMean* = the variance of the condition means.

Appendix L

Proportion of Randomly Generated Sets of Unequal Variances that Included at Least One Negative Value - Study 3

Cells with Five Conditions

SDMean	*SDVar*	Alpha	Negative
0	10	.60	0
0	10	.75	0
0	10	.90	.0008
0	15	.60	0
0	15	.75	.0002
0	15	.90	.0381
10	10	.60	0
10	10	.75	0
10	10	.90	.0006
10	15	.60	0
10	15	.75	.0002
10	15	.90	.0434
20	10	.60	0
20	10	.75	0
20	10	.90	.0006

con't

SDMean	*SDVar*	Alpha	Negative
20	15	.60	0
20	15	.75	.0004
20	15	.90	.0440

Note. *SDMean* = the standard deviation of condition means; *SDVar* = the standard deviation of condition variances; Alpha = the population value of coefficient alpha; Negative = the proportion of sets of variances which included at least one negative value.

Cells with 20 Conditions

SDMean	*SDVar*	Alpha	Negative
0	10	.60	0
0	10	.75	0
0	10	.90	0
0	15	.60	0
0	15	.75	0
0	15	.90	.0002
0	25	.60	.0020
0	25	.75	.0050
0	25	.90	.0663
10	10	.60	0
10	10	.75	0
10	10	.90	0
10	15	.60	0
10	15	.75	0
10	15	.90	.0002
10	25	.60	.0028
10	25	.75	.0046
10	25	.90	.0521
20	10	.60	0
20	10	.75	0
20	10	.90	0

con't

SDMean	*SDVar*	Alpha	Negative
20	15	.60	0
20	15	.75	0
20	15	.90	0
20	25	.60	.0034
20	25	.75	.0038
20	25	.90	.0573

Note. SDMean = the standard deviation of condition means; *SDVar* = the standard deviation of condition variances; Alpha = the population value of coefficient alpha; Negative = the proportion of sets of variances which included at least one negative value.

Appendix M

Average Across Replications of the Average and Variance of Condition Variances and Means - Study 3

Cells with Five Conditions

	Desired Values				Averages Across Replications			
Alpha	AveM	VarM	AveV	VarV	AveM	VarM	AveV	VarV
.60	0	0	100	0	0	0	100	0
.75	0	0	100	0	0	0	100	0
.90	0	0	100	0	0	0	100	0
.60	0	0	100	100	0	0	99.95	100.27
.75	0	0	100	100	0	0	100.03	100.71
.90	0	0	100	100	0	0	100.06	98.65
.60	0	0	100	225	0	0	99.86	225.16
.75	0	0	100	225	0	0	100.03	224.77
.90	0	0	100	225	0	0	100.26	212.59
.60	0	100	100	0	-.1343	99.55	100	0
.75	0	100	100	0	-.1170	100.95	100	0
.90	0	100	100	0	-.0609	99.43	100	0
.60	0	100	100	100	-.0284	98.92	100.02	99.26
.75	0	100	100	100	.0371	100.98	99.93	99.45
.90	0	100	100	100	-.0775	100.15	99.99	99.82

con't

	Desired Values				Averages Across Replications			
Alpha	AveM	VarM	AveV	VarV	AveM	VarM	AveV	VarV
.60	0	100	100	225	.0844	102.67	99.99	221.97
.75	0	100	100	225	.0144	98.88	100.03	224.88
.90	0	100	100	225	-.0751	99.43	100.27	212.50
.60	0	400	100	0	-.1308	404.17	100	0
.75	0	400	100	0	-.0728	400.75	100	0
.90	0	400	100	0	-.1668	406.13	100	0
.60	0	400	100	100	-.1425	400.04	99.94	100.33
.75	0	400	100	100	-.0275	404.62	99.86	101.68
.90	0	400	100	100	-.1209	398.06	100.03	98.62
.60	0	400	100	225	-.0388	398.07	99.93	223.81
.75	0	400	100	225	.0564	397.99	99.84	222.82
.90	0	400	100	225	-.0270	405.39	100.39	211.54

Note. Alpha = the population value of coefficient alpha; *AveV* = the average of condition variances; *VarV* = the variances of condition variances; *AveM* = the average of condition means; *VarM* = the variance of condition means.

Cells with 20 Conditions

	Desired Values				Averages Across Replications			
Alpha	AveM	VarM	AveV	VarV	AveM	VarM	AveV	VarV
.60	0	0	100	0	0	0	100	0
.75	0	0	100	0	0	0	100	0
.90	0	0	100	0	0	0	100	0
.60	0	0	100	100	0	0	100.00	99.450
.75	0	0	100	100	0	0	100.02	99.99
.90	0	0	100	100	0	0	99.96	99.73
.60	0	0	100	225	0	0	99.99	224.95
.75	0	0	100	225	0	0	100.06	224.87
.90	0	0	100	225	0	0	99.95	223.88
.60	0	0	100	625	0	0	100.08	621.48
.75	0	0	100	625	0	0	99.99	626.71
.90	0	0	100	625	0	0	100.17	609.96
.60	0	100	100	0	.0162	99.22	100	0
.75	0	100	100	0	.0322	99.81	100	0
.90	0	100	100	0	-.0264	100.54	100	0
.60	0	100	100	100	-.0207	99.22	100.01	100.11
.75	0	100	100	100	-.0511	99.64	99.97	100.46
.90	0	100	100	100	.0159	100.97	99.96	99.81

con't

	Desired Values				Averages Across Replications			
Alpha	AveM	VarM	AveV	VarV	AveM	VarM	AveV	VarV
.60	0	100	100	225	.0205	100.26	99.97	225.65
.75	0	100	100	225	.0032	99.53	99.99	226.15
.90	0	100	100	225	.0258	99.23	99.92	224.15
.60	0	100	100	625	.0212	100.77	100.00	626.35
.75	0	100	100	625	.0135	99.10	99.73	625.92
.90	0	100	100	625	.0355	100.14	100.28	608.35
.60	0	400	100	0	.1241	398.92	100	0
.75	0	400	100	0	.0224	401.90	100	0
.90	0	400	100	0	-.0387	402.92	100	0
.60	0	400	100	100	.0219	397.17	99.96	100.31
.75	0	400	100	100	.0154	399.03	99.97	100.55
.90	0	400	100	100	.1225	401.75	100.04	99.80
.60	0	400	100	225	-.0572	397.93	100.02	226.22
.75	0	400	100	225	-.0268	396.12	99.99	225.08
.90	0	400	100	225	.0961	400.88	99.88	224.17
.60	0	400	100	625	.0089	402.46	100.00	626.68
.75	0	400	100	625	.0073	400.98	100.38	619.43
.90	0	400	100	625	-.0155	398.37	100.11	609.25

Note. Alpha = the population value of coefficient alpha; *AveV* = the average of condition variances; *VarV* = the variances of condition variances; *AveM* = the average of condition means; *VarM* = the variance of condition means.