Statistical Inference for Coefficient Alpha

Leonard S. Feldt
The University of Iowa

David J. Woodruff
The American College Testing Program

Fathi A. Salih
The University of Iowa

Rigorous comparison of the reliability coefficients of several tests or measurement procedures requires a sampling theory for the coefficients. This paper summarizes the important aspects of the sampling theory for Cronbach's (1951) coefficient alpha, a widely used internal consistency coefficient. This theory enables researchers to test a specific numerical hypothesis about the population alpha and to obtain confidence intervals for the population coefficient. It also permits researchers to test the hypothesis of equality among several coefficients, either under the condition of independent samples or when the same sample has been used for all measurements. The procedures are illustrated numerically, and the assumptions and derivations underlying the theory are discussed.

When an estimate of the reliability of an educational or psychological instrument is needed and the parallel forms and test-retest approaches are impractical, investigators typically rely on internal consistency coefficients. For cognitive tests and affective scales, one of the most commonly used indices is Cronbach's (1951) coefficient alpha. This coefficient is also frequently employed in settings which involve raters or observers (Ebel, 1951). The purpose of this paper is to summarize the sampling theory for coefficient alpha and to illustrate the uses of this theory in evaluating reliability data.

The experimental problems for which the sampling theory is needed include the following: (1) to test the hypothesis that coefficient alpha equals a specified value for a given population; (2) to establish a confidence interval for the alpha coefficient; (3) to test the hypothesis of equality for two or more coefficients when the estimates are based on independent samples; (4) to test the hypothesis of equality when the observed coefficients are based on the same sample and hence are dependent; and (5) to obtain an unbiased estimate of the population value of alpha.

A test of a specific hypothesis is called for when a revised measurement procedure is compared to an established, accepted procedure. In most instances this statistical test would involve a directional alternative, typically that the new procedure is more reliable than the traditional procedure. However, in some applications the test might be two-tailed. Such an alternative might arise when changes in a measurement procedure make administration more efficient, but might affect reliability either positively or negatively.

Studies concerning differences among coefficients are not uncommon. Research on alternative meth-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 1, March 1987, pp. 93–103
© Copyright 1987 Applied Psychological Measurement Inc.
0146-6216/87/010093-11$1.80

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

ods of measuring a specified trait may well call for a test of the equality of alpha coefficients for the several methods. Evaluation of a training program designed to enhance interrater reliability may also demand a test of this null hypothesis. Refinement of an instrument may be assessed, in part, by comparing the reliabilities of several alternative versions. Choice of item types may depend on reliability considerations.

These problems of inference require the development of a sampling error theory for coefficient alpha. The first steps in this development occurred in the early 1960s, when Kristof (1963) and Feldt (1965) independently derived a transformation of the sample alpha coefficient, which is proven to be distributed as F. They showed how this result can be used to test hypotheses and generate confidence intervals for a single alpha coefficient.

Techniques for testing the equality of alpha coefficients were developed over the following 20-year period. The first situation to be considered was that of independent coefficients, that is, coefficients obtained from separate examinee samples. Feldt (1969) derived an F test for the two-coefficient case, and seven years later Hakstian and Whalen (1976) extended the methodology to any number of coefficients. Dependent or related coefficients—reliabilities based on the same sample—posed more complex statistical problems. Feldt (1980) resolved these problems for two coefficients; Woodruff and Feldt (in press) completed the cycle with a test of equality of m dependent coefficients. In each instance, the control of type I error was verified through computer-based monte carlo studies.

The present paper synthesizes this statistical theory for Cronbach's alpha. The principal objective is to make the procedures accessible to researchers and to provide numerical illustrations. For each situation the general outlines of the proofs and derivations are presented.

Inference for a Single Alpha Coefficient

Let X_{jp} denote the score of person p on item j. The test consists of n items or parts, and is administered to N persons. Let Y_p denote the total test score for person p, such that $Y_p = \sum_{j=1}^{n} X_{jp}$. The usual formula for the sample alpha coefficient ($\hat{\zeta}$) is

$$\hat{\zeta} = \left(\frac{n}{n-1}\right)\left(\frac{\hat{\sigma}_Y^2 - \sum_{j=1}^{n} \hat{\sigma}_{X_j}^2}{\hat{\sigma}_Y^2}\right) \quad . \tag{1}$$

In this formula $\hat{\sigma}_{X_j}^2$ represents the unbiased estimate of the variance for item j, and $\hat{\sigma}_Y^2$ the unbiased variance estimate for score Y. (The sample alpha coefficient is denoted by $\hat{\zeta}$ and its parameter value by ζ to avoid confusion with the symbolic representation of statistical significance levels, almost universally denoted in the statistical literature by α.)

Following Hoyt (1941), an alternate formula for $\hat{\zeta}$ may be derived by considering the responses of the N persons on the n items as observations in a two-way persons $\times$ items analysis of variance (ANOVA) with one observation per cell. Within this framework, a formula for $\hat{\zeta}$ is

$$\hat{\zeta} = \frac{\text{MS(P)} - \text{MS(PI)}}{\text{MS(P)}} = 1 - \frac{\text{MS(PI)}}{\text{MS(P)}} \quad , \tag{2}$$

where MS(P) denotes the mean square for persons and MS(PI) denotes the mean square for persons $\times$ items interaction. When applied to the setting in which n raters evaluate N persons, $\hat{\zeta}$ can be used as measure of interrater agreement, with differences among rater means *not* considered measurement error. In such a case, raters substitute for items in Equation 2.

Let E denote expected value, and in particular let the expected values for MS(P) and MS(PI) be

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

denoted as E[MS(P)] and E[MS(PI)], respectively. The population value of coefficient alpha is then defined as

$$\zeta = \frac{\text{E[MS(P)]} - \text{E[MS(PI)]}}{\text{E[MS(P)]}} = 1 - \frac{\text{E[MS(PI)]}}{\text{E[MS(P)]}} \quad . \tag{3}$$

Assume that items and persons are randomly sampled from their respective domains and that the persons × items interaction effects and the residual errors are independently and normally distributed with homogeneous variance. Under these conditions, often referred to as the Type II ANOVA model, Kristof (1963) and Feldt (1965) independently proved that the following statistic is distributed as F with degrees of freedom (DF) of $N - 1$ and $(n - 1)(N - 1)$:

$$\frac{1 - \zeta}{1 - \hat{\zeta}} = \frac{\text{MS(P)/E[MS(P)]}}{\text{MS(PI)/E[MS(PI)]}} \quad . \tag{4}$$

The proof that $(1 - \zeta)/(1 - \hat{\zeta})$ is an F variable follows from the fact that under the assumed ANOVA model, MS(P)/E[MS(P)] is distributed as a chi-square variable divided by its DF, $N - 1$. Likewise, MS(PI)/E[MS(PI)] is distributed as a chi-square variable divided by its DF, $(n - 1)(N - 1)$. Under the assumed model these chi-squares are independent. Therefore, their ratio (Equation 4) is distributed as a central F with DF of $N - 1$ and $(n - 1)(N - 1)$.

This distribution theory for $(1 - \zeta)/(1 - \hat{\zeta})$ may be used to formulate a test of a specific numerical hypothesis and to derive a confidence interval for a population alpha coefficient. To test the null hypothesis H_0: $\zeta = \zeta_0$ against a two-tailed alternative at the α level of significance, let $F(\alpha/2)$ denote the $100(\alpha/2)$ percentile and $F(1 - \alpha/2)$ the $100(1 - \alpha/2)$ percentile of the central F with $N - 1$ and $(n - 1)(N - 1)$ as its DF. The null hypothesis is rejected if

$$\hat{\zeta} < 1 - \frac{(1 - \zeta_0)}{F(\alpha/2)} \quad \text{or} \quad \hat{\zeta} > 1 - \frac{(1 - \zeta_0)}{F(1 - \alpha/2)} \quad . \tag{5}$$

If a one-tailed test at the α level of significance is desired, $\alpha/2$ is replaced by α in the appropriate critical value.

The upper and lower endpoints of a $100(1 - \alpha)$ percent interval for ζ are given respectively by

$$\zeta_U = 1 - [(1 - \hat{\zeta})F(\alpha/2)] \tag{6}$$

and

$$\zeta_L = 1 - [(1 - \hat{\zeta})F(1 - \alpha/2)] \quad . \tag{7}$$

If a one-sided $100(1 - \alpha)$ percent interval is desired, $\alpha/2$ is replaced by α in the appropriate endpoint.

The following example illustrates the procedures outlined above. Suppose a researcher used 41 examinees to obtain an estimate of .790 for the alpha coefficient of a 26-item test. The relevant F distribution has DF of 40 and 1,000, for which the 5th and 95th percentiles are .66 and 1.41. The 90% confidence interval (bounded below and above) has

$$\zeta_U = 1 - (1 - .79)(.66) = .861 \tag{8}$$

and

$$\zeta_L = 1 - (1 - .79)(1.41) = .704 \quad . \tag{9}$$

A one-tailed test of H_0: $\zeta_0 = .70$, with H_{alt}: $\zeta_0 > .70$ and $\alpha = .05$, would require only a lower bound for the critical region. By Equation 5, the critical region (CR) is

$$\text{CR} > 1 - \frac{(1 - .70)}{F(.95)} = 1 - \frac{(1 - .70)}{1.41} = .787 \quad . \tag{10}$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Since the observed coefficient alpha of .790 exceeds the lower bound of the critical region, $\zeta_0 = .70$ may be rejected.

The expected value of $\hat{\zeta}$, $E(\hat{\zeta})$, and the bias in $\hat{\zeta}$ can be deduced from the fact that $(1 - \hat{\zeta})/(1 - \zeta)$ is also distributed as F, but with DF of $(n - 1)(N - 1)$ and $(N - 1)$. Since the expected value of a central F is $\nu_2/(\nu_2 - 2)$, where ν_2 is the second DF value,

$$E[(1 - \hat{\zeta})/(1 - \zeta)] = (N - 1)/(N - 3) \quad , \tag{11}$$

and hence

$$E(\hat{\zeta}) = 1 - (1 - \zeta)(N - 1)/(N - 3) \quad . \tag{12}$$

It follows that

$$E(\hat{\zeta}) - \zeta = 2(\zeta - 1)/(N - 3) \quad . \tag{13}$$

Since the difference $\zeta - 1$ must be negative, it follows that $\hat{\zeta}$ tends to underestimate ζ. This result was first presented by Kristof (1963).

The negative bias of $\hat{\zeta}$ is generally of little consequence unless N is small. If $N = 50$ and $\zeta = .70$, for example, the expected value of $\hat{\zeta}$ is .687. With $N = 100$, the expected value is .694. Where an unbiased estimate of ζ is required, it may be obtained by the formula

$$\tilde{\zeta} = [(N - 3)\hat{\zeta}/(N - 1)] + 2/(N - 1) \quad . \tag{14}$$

Comparison of Alpha Coefficients Obtained from Independent Samples

Rigorous comparisons of alternative test scoring procedures, test construction techniques, item formats, item selection strategies, modes of test administration, or competing test instruments entail, in part, the comparison of reliabilities. Oaster (1984), for example, needed such a test in his investigation of the comparative reliabilities of various types of Likert scales. Studies of the efficacy of corrections for guessing have frequently employed independent random samples of examinees who took the same multiple-choice test under alternative directions and scoring formulas. The trend toward test administration by microcomputer raises the question of whether paper-and-pencil and computerized administrations produce equally reliable scores.

Comparisons among examinee subpopulations are a significant area of application of statistical procedures for comparing alpha coefficients. For example, it might be of some practical and theoretical concern whether the vocational interests of males are measured as reliably as those of females. Researchers might wish to examine whether aptitude, achievement, or interests of minority groups are assessed as reliably as those of the majority. Answers to these questions would require a test of H_0: $\rho_{\alpha_1} = \rho_{\alpha_2} = \ldots = \rho_{\alpha_m}$. However, this type of application is complicated by differences among populations in their inherent variability. It is widely recognized that test reliability is directly related to the variance of true scores. Rejection of the hypothesis of equality of alpha coefficients may come about through a combination of lower error variance and higher true score variance. Therefore, the outcome of statistical tests based on independent subpopulations must be interpreted with considerable caution.

The null hypothesis testing problem was first addressed by Feldt (1969), who derived a statistical test of the hypothesis H_0: $\zeta_1 = \zeta_2$. The Feldt approach uses the test statistic $W = (1 - \hat{\zeta}_2)/(1 - \hat{\zeta}_1)$. He proved that when the reliability parameters are equal, W is distributed as the product of two independent central F variables. This product, it was shown, could be well approximated by a single F with DF of $N_1 - 1$ and $N_2 - 1$. With modern computing equipment it is relatively simple to determine the probability that a central F will exceed the obtained value of W. If the probability is less than the significance level, the hypothesis of equality can be rejected.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Hakstian and Whalen (1976) extended the methodology to the case of m coefficients. Their test is based on the normalizing transformation of F developed by Paulson (1942) and the fact that the statistic $(1 - \hat{\zeta})/(1 - \zeta)$ is distributed as F with DF of $(N - 1)(n - 1)$ and $(N - 1)$. Paulson proved that

$$z = \frac{[1 - (2/9\nu_2)]\, F^{1/3} - [1 - (2/9\nu_1)]}{[(2/9\nu_2)\, F^{2/3} + (2/9\nu_1)]^{1/2}} , \tag{15}$$

where z is distributed as a unit normal deviate. In the present context, the transformation may be stated as follows:

$$z = \frac{(1 - \hat{\zeta})^{1/3} - \dfrac{[1 - (2/9\nu_1)](1 - \zeta)^{1/3}}{[1 - (2/9\nu_2)]}}{\left[\dfrac{18\nu_2(1 - \hat{\zeta})^{2/3}}{(9\nu_2 - 2)^2} + \dfrac{18\nu_2^2(1 - \zeta)^{2/3}}{\nu_1(9\nu_2 - 2)^2}\right]^{1/2}} . \tag{16}$$

This ratio implies that $(1 - \hat{\zeta})^{1/3}$ is approximately normally distributed with non-zero mean (the fractional term in the numerator of Equation 16) and variance approximated by

$$S^2 = \left[\frac{18\nu_2(1 - \hat{\zeta})^{2/3}}{(9\nu_2 - 2)^2}\right]\left(1 + \frac{\nu_2}{\nu_1}\right) = \left[\frac{18(N - 1)(1 - \hat{\zeta})^{2/3}}{(9N - 11)^2}\right]\left(\frac{n}{n - 1}\right) . \tag{17}$$

Hakstian and Whalen (1976) proposed that the weighted average (μ^*) of $(1 - \hat{\zeta}_i)^{1/3}$ be obtained, the weights equaling the reciprocals of the variances. The test statistic is then defined as

$$M = \sum_i^m \left[\frac{(1 - \hat{\zeta}_i)^{1/3} - \mu^*}{S_i}\right]^2 , \tag{18}$$

which is interpreted as a chi-square with $m - 1$ degrees of freedom. The justification for this interpretation is that the sum of m square standardized deviations of normal variables from their weighted mean is so distributed. The test is thus analogous to the test of the equality of m correlation coefficients, wherein Fisher's transformation of the coefficients has achieved normality with variances $1/(N_i - 3)$. (See Hays, 1981, p. 469.)

There are two minor problems with the Hakstian–Whalen test. First, the variance of $(1 - \hat{\zeta}_i)^{1/3}$ is an estimate based on the sample values, $\hat{\zeta}_i$. This is contrary to theory and in contrast to the case of transformed correlations, in which the variances, $1/(N_i - 3)$, do not depend upon sample estimates. Second, even if all ζ_i are equal, the statistics $(1 - \hat{\zeta}_i)^{1/3}$ do not come from the *same* normal distribution unless the fractional term in the numerator of Equation 16 is the same for all tests. This equality demands that $(1 - 2/9\nu_1)/(1 - 2/9\nu_2)$ be constant over all tests.

Fortunately, these problems appear to be of little consequence. The use of the sample statistic, $\hat{\zeta}_i$, to replace the parameter, ζ_i, in the second term in the denominator of Equation 16 seems to have little effect on the distribution of the ratio. The net effect might be likened to that of interpreting a t statistic as a normally distributed variable—an interpretation that involves no serious error when the sample size is larger than 50 (see Marascuilo, 1966). The minimal effect of replacing ζ by $\hat{\zeta}$ in the second term probably results from the fact that this term is of order $2/(9)(n - 1)(N - 1)$. The first term, which properly includes $\hat{\zeta}$, is of order $2/(9)(N - 1)$.

The second problem also proves to be of negligible importance because $1 - (2/9\nu_1)$ and $1 - (2/9\nu_2)$ are both very close to 1, regardless of the variation in n_i and N_i from test to test. For example, if $n_1 = 50$ and $N_1 = 100$, the ratio of these terms is 1.0022. If $n_2 = 10$ and $N_2 = 50$, the ratio is 1.0040. Thus, the hypothesis that $1.0022(1 - \zeta_1)^{1/3} = 1.0040(1 - \zeta_2)^{1/3}$ is essentially a hypothesis that $\zeta_1 = \zeta_2$.

Woodruff and Feldt (in press) treated the case of m independent coefficients as a special case of m

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

dependent or related coefficients, and arrived at a similar hypothesis testing procedure. They adopted the transformation $1/(1 - \hat{\zeta})^{1/3}$ rather than $(1 - \hat{\zeta})^{1/3}$. A critical point in the subsequent derivation is the identification of a chi-square distribution (DF to be determined) for which the variable χ^2/DF has nearly the same mean, variance, skewness, and kurtosis as $(1 - \zeta)/(1 - \hat{\zeta})$. The latter variable is distributed as F with DF of $N - 1$ and $(N - 1)(n - 1)$. The chi-square distribution which best satisfies this requirement takes $\text{DF} = \tilde{N}_i - 1$, where $\tilde{N}_i = (n_i - 1)N_i/(n_i + 1)$. Woodruff and Feldt approximated the variance of $1/(1 - \hat{\zeta})^{1/3}$ by the Wilson–Hilferty (1931) normalizing transformation of a chi-square variable. This leads to the following estimate of the variance of $1/(1 - \hat{\zeta}_i)^{1/3}$:

$$S_i^2 = \frac{2}{9(\tilde{N}_i - 1)(1 - \hat{\zeta}_i)^{2/3}} \quad . \tag{19}$$

Unlike Hakstian and Whalen, Woodruff and Feldt used the arithmetic unweighted mean of the transformed coefficients,

$$\bar{\mu} = \sum_i^m (1 - \hat{\zeta}_i)^{-1/3}/m \quad . \tag{20}$$

Their test statistic, under the condition of independent samples, is

$$UX_1 = \sum_i^m [(1 - \hat{\zeta}_i)^{-1/3} - \bar{\mu}]^2/\bar{S}^2 \quad , \tag{21}$$

where $\bar{S}^2$ is the arithmetic mean of the several variances, S_i^2. Under H_0 it is approximately distributed as χ^2 with DF of $m - 1$.

These two approaches may be illustrated by the following data:

Group 1 and Test 1: $\hat{\zeta}_1 = .784$; $n_1 = 5$, $N_1 = 51$; $(1 - \hat{\zeta}_1)^{1/3} = .6$
Group 2 and Test 2: $\hat{\zeta}_2 = .875$; $n_2 = 5$, $N_2 = 101$; $(1 - \hat{\zeta}_2)^{1/3} = .5$
Group 3 and Test 3: $\hat{\zeta}_3 = .936$; $n_3 = 5$, $N_3 = 151$; $(1 - \hat{\zeta}_3)^{1/3} = .4$

The Hakstian and Whalen variances are .0020179, .00068754, and .00029718. The weighted average, μ^*, is .4458. The test statistic is 23.053.

The same data, analyzed according to the Woodruff and Feldt approach, yield variances of .0187056, .0134003, and .0139353, and $\bar{\mu}$ of 2.05556. The test statistic, also interpreted as a chi-square with DF of 2, is 22.926. The tests might be expected to give highly consistent results, as they did in this instance.

If tests of pairwise contrasts among the coefficients are warranted on the basis of a significant outcome of the omnibus test, the pairs can be considered by Feldt's (1969) test for two coefficients. In the present instance, all pairs lead to rejection of the null hypothesis.

Comparison of Alpha Coefficients Obtained From the Same Sample

In some settings it is possible to administer all instruments or to apply all procedures to the *same* sample of N examinees. In such instances the coefficients are statistically dependent, and the test of the null hypothesis must recognize this dependence. To ignore it is tantamount in most applications to adoption of a significance level far more stringent than the nominal level.

One important area of application includes studies designed to compare alternative instruments or measurement techniques (multimethod experimentation). It is important to note that in educational and psychological settings, testing time—not the number of discrete exercises—is almost always the most crucial dimension of test length. Hence, in studies of this type the instruments should be highly similar in their time requirements. When tests are equated in time, they likely will differ in the number of items

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

they contain. For example, a study designed to compare the cloze technique and the question-and-answer technique of measuring reading comprehension might well involve instruments with unequal numbers of exercises, since these types of items cannot be answered at the same rate. Analogously, reliability comparisons among various types of verbal fluency measures would probably involve tests containing different numbers of items. Items constructed in an analogy format, for instance, require more testing time per item than exercises in the "choose the best synonym" format. Comparisons among measures which require different amounts of testing time would be reasonable only in those situations in which existing standardized instruments are compared. In such circumstances the researcher must administer the instruments at the lengths and within the time limits for which the norms apply.

The methodology for the case of dependent statistics, like that for independent statistics, was first developed for H_0: $\zeta_1 = \zeta_2$. Feldt (1980) derived three procedures for testing this hypothesis. Simulation studies indicated that all three procedures control type I error rates satisfactorily. Feldt recommended the following test statistic:

$$t = \frac{(\hat{\zeta}_1 - \hat{\zeta}_2)(N - 2)^{1/2}}{[4(1 - \hat{\zeta}_1)(1 - \hat{\zeta}_2)(1 - \hat{\rho}^2)]^{1/2}} \qquad (\text{DF} = N - 2) \quad . \tag{22}$$

The squared correlation in the denominator refers to the squared coefficient between the two total-test scores for the sample.

The derivation of this test rests on the fact that if $\zeta_1 = \zeta_2$, then $(1 - \hat{\zeta}_2)/(1 - \hat{\zeta}_1)$ is distributed identically as the ratio of two dependent sample variances, each with an expected value of 1.0. Pitman (1939) proved that the following function of such a ratio is distributed as

$$t = \frac{[(\hat{\sigma}_2^2/\hat{\sigma}_1^2) - 1](N - 2)^{1/2}}{[(4\hat{\sigma}_2^2/\hat{\sigma}_1^2)(1 - \hat{\rho}^2)]^{1/2}} \qquad (\text{DF} = N - 2) \quad . \tag{23}$$

Thus, the same function of $(1 - \hat{\zeta}_2)/(1 - \hat{\zeta}_1)$ must be distributed as t with DF of $N - 2$. Substitution of $(1 - \hat{\zeta}_2)/(1 - \hat{\zeta}_1)$ for $\hat{\sigma}_2^2/\hat{\sigma}_1^2$ in this expression ultimately leads, after algebraic simplification, to Equation 22.

Woodruff and Feldt (in press) extended the methodology to the case of m dependent coefficients. They considered 11 possible test statistics. Extensive monte carlo simulation led to three procedures that showed excellent control of type I error and superior power, compared to the others. Of these three techniques, the procedure identified as UX_1 was the simplest computationally and is summarized here.

As in the case of independent coefficients, Woodruff and Feldt (in press) approximated the variance of $1/(1 - \hat{\zeta}_i)^{1/3}$ by the quantity

$$S_i^2 = \frac{2}{9(\tilde{N}_i - 1)(1 - \hat{\zeta}_i)^{2/3}} \quad . \tag{24}$$

However, the test for dependent alphas also demands an approximation of the covariance between $1/(1 - \hat{\zeta}_i)^{1/3}$ and $1/(1 - \hat{\zeta}_j)^{1/3}$. Using the delta method of Kendall and Stuart (1977), Woodruff and Feldt derived the following estimate:

$$S_{ij} = \frac{2\hat{\rho}_{ij}^2}{9(\tilde{N} - 1)(1 - \hat{\zeta}_i)^{1/3}(1 - \hat{\zeta}_j)^{1/3}} \quad . \tag{25}$$

As in the case of two coefficients, $\hat{\rho}_{ij}^2$ is the square of the sample correlation between the scores on tests i and j. When the numbers of items differ, then $\tilde{N} = N(\tilde{n} - 1)/(\tilde{n} + 1)$, where $\tilde{n}$ is the harmonic mean of all test lengths. On the assumption that the variables $1/(1 - \zeta_i)^{1/3}$ have a multivariate normal distribution, a matrix function of $\hat{\zeta}_i$, ρ, σ_i^2, σ_{ij}, and $\tilde{N}$ is shown to be distributed approximately as χ^2 with DF of $m - 1$. Woodruff and Feldt (in press) further showed that an approximation of this function serves satis-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

factorily as a test statistic. It is

$$UX_1 = \sum_{i}^{m} [(1 - \hat{\zeta}_1)^{-1/3} - \bar{\mu}]^2 / (\bar{S}^2 - \bar{C}) \quad , \tag{26}$$

where $\bar{S}^2$ is the average of the variances S_i^2 (Equation 24),
$\bar{C}$ is the average of the covariances S_{ij} (Equation 25), and
$\bar{\mu}$ is the average of the transformed coefficients, $1/(1 - \hat{\zeta}_i)^{1/3}$.
UX_1 is distributed approximately as chi-square with $m - 1$ DF when H_0 is true.

This procedure may be illustrated by data for a four-test situation with $N = 100$. To make the situation more concrete, imagine that each of these tests is composed of a specific type of item, with the numbers of items in each test determined so that all tests have the same time requirements. Test data are as follows:

Test 1: $n_1 = 50$ $\hat{\zeta}_1 = .857$ $(1 - \hat{\zeta}_1)^{-1/3} = 1.91229$
Test 2: $n_2 = 40$ $\hat{\zeta}_2 = .875$ $(1 - \hat{\zeta}_2)^{-1/3} = 2.00000$
Test 3: $n_3 = 35$ $\hat{\zeta}_3 = .800$ $(1 - \hat{\zeta}_3)^{-1/3} = 1.70998$
Test 4: $n_4 = 25$ $\hat{\zeta}_4 = .833$ $(1 - \hat{\zeta}_4)^{-1/3} = 1.81591$

For all tests combined, $\tilde{n} = 35.22013$, $\tilde{N} = 94.47820$, and $\bar{\mu} = 1.85955$. Test intercorrelations are

$$\begin{bmatrix} 1.00 & .80 & .60 & .75 \\ & 1.00 & .65 & .70 \\ & & 1.00 & .55 \\ & & & 1.00 \end{bmatrix}$$

S_i^2 and S_{ij} are

$$\begin{bmatrix} .0086934 & .0058189 & .0027985 & .0046435 \\ & .0095091 & .003450 & .0042305 \\ & & .0069514 & .0022330 \\ & & & .0078390 \end{bmatrix}$$

and $\bar{S}^2 = .0082482$, $\bar{C} = .0038599$, $\bar{S}^2 - \bar{C} = .0043883$, $UX_1 = 10.661$, and $P[\chi^2_{(3)} > 10.66] = .014$. Analogous to the situation involving independent coefficients, follow-up tests of pairwise contrasts can be made using the t test presented earlier for two coefficients.

Cruciality of the Statistical Assumptions

The most fundamental distributional assumption required by these inferential procedures is that the quantity $(1 - \zeta)/(1 - \hat{\zeta})$ be distributed as F. As previously noted, this assumption will be met if the scores conform to the dictates of the two-way random effects model (Type II ANOVA) with one observation per cell. These requirements will be met if the item or part scores are normally distributed with homogeneous error variances. However, these assumptions will almost surely be violated if each part of the instrument gives rise to a restricted range of scores. Therefore, the question arises of how well the procedures may be expected to perform with actual data.

Feldt (1965), in deriving the F distribution for the transformed alpha coefficient, gave a detailed discussion of the assumptions required under the random effects model and how they might be violated with dichotomously scored items. He also reported on the results of a simulation study based on real test data with dichotomously scored items. The results indicate that the F distribution holds up well with such data.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

In an experimental design context, Seeger and Gabrielsson (1968) simulated the distribution of mean square ratios under the mixed model ANOVA when applied to dichotomous data. They considered the situation involving several observations per cell and focused attention on the F ratio pertaining to treatment effects. Though this ratio is not the one used in reliability studies, their simulations offer further indirect evidence that $(1 - \zeta)/(1 - \hat{\zeta})$ is distributed approximately as F even if the items are dichotomously scored.

Inference for several alpha coefficients based on independent samples requires, in addition to distributional assumptions, that the sample sizes be large enough to justify the asymptotic chi-square distribution for the Hakstian–Whalen test statistic M and the Woodruff–Feldt statistic UX_1. Hakstian and Whalen used monte carlo methods to investigate the sampling distribution of test statistic M when computed from dichotomous part scores. Their results indicate good control of Type I error rates with as few as 20 examinees per test, even for this gross departure from normality and homogeneity of variance.

If the same sample or matched samples are used for testing the equality of several alpha coefficients, two additional assumptions are required. The first is that the $1/(1 - \hat{\zeta}_i)^{1/3}$ have a joint multivariate normal distribution. The second is that the correlations between total scores Y_i and Y_j are identical (homogeneous) for all pairs of tests. If the $(1 - \zeta_i)/(1 - \hat{\zeta}_i)$ have approximate F distributions, then the $1/(1 - \hat{\zeta}_i)^{1/3}$ have marginal distributions approximately normal in form. Given these marginal normal distributions, it is reasonable to assume multivariate normality. However, multivariate normality does not automatically follow from the condition of marginal normality.

Woodruff and Feldt (in press) investigated the power and Type I error control of UX_1 using monte carlo methods. They found that for a sample size as small as 50 and with moderately heterogeneous, positive inter-test correlations (range of ρ equal to .30), control of Type I error rates was quite good. However, these simulations were based on continuous, normally distributed scores. They did not provide evidence as to the cruciality of the normality assumption for total scores, nor did they document the effects of dichotomous item scoring.

The results of subsequent monte carlo investigations of these issues are summarized in Tables 1 and 2. In the first of these studies, dichotomous item scores were generated using a computer simulation technique described by Nitko (1968). Two true null hypotheses were considered. In the first, $\zeta_i = .80$ for each of four tests with 30 items in each test. In the second, $\zeta_i = .65$ for each of three tests with 30, 30, and 60 items, respectively. Each 30-item test exhibited a range of item difficulties from .30 to .80; the 60-item test had a range of item difficulties of .35 to .73. The item difficulty distribution for each test was unimodal and symmetrical around the value .55. The resultant distributions of total test scores were slightly skewed negatively ($\gamma_1 = -.13$) and platykurtic ($\gamma_2 = -.53$), generally similar to the distributions for many standardized tests. The inter-test correlations were homogeneous and equal to their shared reliability (.65 or .80).

For each null hypothesis, 2,200 simulations of random sample data were produced, based on $N = 50$ and $N = 100$. Test UX_1 was performed on each replication, and the percent of test statistics exceeding the upper 10%, 5%, and 1% points of $\chi^2_{(m-1)}$ was tabulated. These data are summarized in Table 1.

It may be observed that the UX_1 test showed no gross effects from dichotomous item scoring. There is a tendency toward liberality if $N = 50$ and a .10 or .05 level is employed, but this deviation from the nominal significance level would not disturb most researchers.

The second empirical study used actual test data: scores of Iowa students in Grades 9 and 11 on various subtests of the Iowa Tests of Educational Development (ITED), Form X–7. The subtests for Grade 9 were selectively shortened by the deletion of items so that all tests had $\zeta = .75$. The subtests for Grade 11 were differentially shortened so that all tests had $\zeta = .87$. From the pool of 16,443 records for Grade 9 and 16,760 records for Grade 11, 2,000 random samples of $N = 50$ and 2,000 samples of $N = 100$ were chosen by sampling examinees randomly with replacement. The UX_1 test was then executed on each examinee sample, using $m = 2$, 3, 4, or 5 ITED subtests. With $\zeta = .75$, the test lengths were

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Estimated Probability of Type I Error Based on 2200 Replications of the UX_1 Test: Simulated Dichotomous Item Scores

	ζ = .8	m = 4	n_i = 30	ζ = .65	m = 3	n_i = 60,30,30
	10%	5%	1%	10%	5%	1%
N=50	10.7	5.4	1.4	10.0	5.1	0.9
N=100	10.9	5.6	1.1	9.8	5.0	1.1

11, 13, 16, 21, and 21 items. With ζ = .87, the test lengths were 24, 28, 31, 36, and 46 items. The results of this study are summarized in Table 2.

Table 2
Estimated Probability of Type I Error Based on 2000 Replications of the UX_1 Test: Actual Dichotomous Item Scores

		ζ = .75			ζ = .87		
Sample Size	m = 2	n_i = 11, 13 and 11, 16			n_i = 24, 31 and 28, 46		
		10%	5%	1%	10%	5%	1%
N = 50		10.4	5.6	1.2	10.4	5.5	1.4
N = 100		9.8	4.7	0.9	9.7	5.0	1.2
	m = 3	n_i = 11, 16, 21 and n_i = 11, 13, 16			n_i = 24, 36, 46 and n_i = 24, 31, 46		
N = 50		10.2	5.2	1.1	10.2	5.2	1.4
N = 100		9.8	4.1	0.9	9.9	5.1	1.0
	m = 4	n_i = 11, 16, 21, 21 and n_i = 11, 13, 16, 21			n_i = 24, 28, 36, 46 and n_i = 24, 28, 31, 46		
N = 50		10.5	5.9	1.1	11.3	5.8	1.4
N = 100		9.8	4.8	0.9	10.3	5.6	1.4
	m = 5	n_i = 11, 13, 16, 21, 21			n_i = 24, 28, 31, 36, 46		
N = 50		11.9	6.0	1.1	10.2	5.2	1.4
N = 100		10.3	5.2	0.8	9.7	5.5	1.7

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

With actual test data, the control of Type I error was about as tight as with simulated dichotomous item scores. The deviations from the nominal significance level were more pronounced with $N = 50$ than with $N = 100$. It must be borne in mind, of course, that the standard error of a percent in the vicinity of 10% is about .67% with about 2,000 trials; near 5% the standard error is about .49%.

A crude summary over 14 situations resulted in average estimated probabilities at the 10%, 5%, and 1% levels respectively for $N = 50$ of 10.6%, 5.5%, and 1.3%; and for $N = 100$ of 9.9%, 5.0%, and 1.1%. These means are very close to the averages for simulated data (Table 1). Together, they support the conclusion that the UX_1 test works quite well with $N = 100$, but it errs on the side of liberality with $N = 50$. The degree of liberality is not great, and most researchers would probably be willing to accept a test that controls Type I error within one-half of one percent. But there is a need for an improved test for use with sample sizes of 50 or less.

It is pertinent to note that almost all of the test instruments used in this study gave rise to negatively skewed, platykurtic score distributions. The γ_1 index of skewness ranged between $-.597$ and .165, with eight of the ten indices negative. The γ_2 index of kurtosis ranged between $-.015$ and $-.948$. The average value of γ_2 for all 10 tests (5 in each of two grades) was $-.676$. Quite possibly this characteristic of the score distributions accounts for the liberality of the UX_1 test with the smaller sample size.

References

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

Ebel, R. L. (1951). Estimation of reliability of ratings. *Psychometrika, 16,* 407–424.

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.

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

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart, and Winston.

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

Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. I, 4th ed.). London: Griffin.

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.

Marascuilo, L. A. (1966). Large-sample multiple comparisons. *Psychological Bulletin, 65,* 280–290.

Nitko, A. J. (1968). The power functions of some proposed tests for the significance of coefficient alpha in the one-sample and two-sample cases (Doctoral dissertation, University of Iowa, 1968). *Dissertation Abstracts, 29,* 2234B.

Oaster, R. R. F. (1984, April). *Even numbered alternative scale reliability*. Paper presented at the meeting of the American Educational Research Association, New Orleans.

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

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika, 31,* 9–12.

Seeger, P., & Gabrielsson, A. (1968). Applicability of the Cochran Q test and the *F* test for statistical analysis of dichotomous data for dependent samples. *Psychological Bulletin, 69,* 269–277.

Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences, USA, 17,* 684.

Woodruff, D. J., & Feldt, L. S. (in press). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika.*

Acknowledgments

The authors acknowledge the assistance of Mayuree Srichai, who programmed the simulations of dichotomous item data and the monte carlo portions of this research.

Author's Address

Send requests for reprints or further information to Leonard S. Feldt, 334 Lindquist Center, The University of Iowa, Iowa City IA 52242, U.S.A.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/