PSYCHOMETRIKA—VOL. 45, NO. 1.
MARCH, 1980

A TEST OF THE HYPOTHESIS THAT CRONBACH'S ALPHA RELIABILITY COEFFICIENT IS THE SAME FOR TWO TESTS ADMINISTERED TO THE SAME SAMPLE

LEONARD S. FELDT

UNIVERSITY OF IOWA

In measurement studies the researcher may wish to test the hypothesis that Cronbach's alpha reliability coefficient is the same for two measurement procedures. A statistical test exists for independent samples of subjects. In this paper three procedures are developed for the situation in which the coefficients are determined from the same sample. All three procedures are computationally simple and give tight control of Type I error when the sample size is 50 or greater.

Key words: coefficient alpha, reliability coefficients, tests of significance.

Measurement studies in education and psychology occasionally call for a test of the hypothesis that Cronbach's coefficient alpha is the same for two tests or measurement procedures. For example, a researcher might wish to determine if the reliability of one approach to the assessment of a particular trait differs from that of an alternative approach. An investigator might want to evaluate the effect on reliability of a training program for evaluators or study the reliability implications of variations in test directions. Such situations demand a statistical test of the hypothesis that the population values of the reliability coefficients are equal.

If the values of coefficient alpha are obtained from independent random samples, the technique proposed by Feldt [1969], and the extention to the k-sample situation by Hakstian and Whalen [1976], may be used. This approach has been found to control Type I error quite precisely even in the limiting case of dichotomously scored items, in which coefficient alpha reduces to Kuder-Richardson Formula 20. However, when the coefficients are determined from the *same* sample of examinees, the Feldt statistical test cannot be validly employed. If applied to coefficients obtained from the same sample, the test would be unduly conservative. The purpose of this paper is to derive several approximate tests that may be used in this situation and to report the results of sampling studies which bear on the Type I error control of these tests.

The sampling theory for independent groups draws on the analysis of variance approach to computation of coefficient alpha. Regardless of the length of the score scale for each test part or item, r_α may be computed as $1 - (ms_{PS}/ms_S)$, where ms_{PS} is the mean square for the parts (items) by subjects interaction and ms_S is the mean square for subjects. Kristof [1963] and Feldt [1965] demonstrated that if the scores on k parallel parts of a test conform to the assumptions of the two-factor random model, $(1 - \rho_\alpha)ms_S/ms_{PS}$ is distributed as a central F with $N - 1$ and $(N - 1)(k - 1)$ degrees of freedom. In this expression N is the number of examinees and ρ_α is the population value of the coefficient alpha. Since ms_S/ms_{PS} equals $1/(1 - r_\alpha)$, it follows that $(1 - \rho_\alpha)/(1 - r_\alpha)$ is distributed as $F_{N-1,(k-1)(N-1)}$ and $(1 - r_\alpha)/(1 - \rho_\alpha)$ is distributed as $F_{(k-1)(N-1),N-1}$.

Building on this theory, Feldt [1969] noted that

The author is indebted to Jerry S. Gilmer for development of the computer programs used in this study.

Requests for reprints should be sent to Leonard S. Feldt, College of Education, Division of Educational Psychology, Measurement & Statistics, The University of Iowa, Iowa City, Iowa 52242.

0033-3123/80/0300-2834 $00.75/0

$$\text{(1)} \qquad \frac{1-\rho_1}{1-r_1} \cdot \frac{1-r_2}{1-\rho_2} \sim [F_{N_1-1,(k_1-1)(N_1-1)}][F_{(k_2-1)(N_2-1),N_2-1}]$$

where r_1 and r_2 are coefficients based on independent samples. He further showed that if $(k_1 - 1)(N_1 - 1)$ and $(k_2 - 1)(N_2 - 1)$ are fairly large, as they would be with at least 21 items and 51 subjects, the distribution of the product of the two independent F variables is practically identical to that of F_{N_1-1,N_2-1}. This makes possible a relatively simple test of H_0: $\rho_1 = \rho_2$, since

$$\text{(2)} \qquad W = \frac{1-r_2}{1-r_1} \sim \left[\frac{1-\rho_2}{1-\rho_1}\right] F_{N_1-1,N_2-1}.$$

If $\rho_1 = \rho_2$, W is distributed as a central F with $N_1 - 1$ and $N_2 - 1$ degrees of freedom. If the population coefficients are not equal, W is distributed as a constant greater than or less than 1.0 times a central F. If W is too large or too small to be accepted as a randomly drawn central F, the implication is that the constant $(1 - \rho_2)/(1 - \rho_1)$ does not equal 1.0 and hence the hypothesis is false.

With repeated use of the same sample for both instruments, the foregoing theory breaks down. Crucial to its derivation is the stipulation that ms_S/ms_{PS} for the first instrument be independent of ms_{PS}/ms_S for the second. This condition cannot be met if the scores on the two instruments under study are correlated. Statistical tests for which such dependence is assumed are developed in the following section.

An Approximate Test of the Hypothesis $\rho_1 = \rho_2$ with Related Samples

Assume that Tests 1 and 2, whose reliabilities are to be compared, have been taken by the same sample of N examinees. Test 1 is composed of k_1 scoreable units and Test 2 of k_2 scoreable units. The scores and score distributions for the two tests may not be directly comparable, since the instruments may be of different lengths and may contain different kinds of exercises.

For each test the following assumptions are made:

(i) The examinees are a random sample from the population of interest.

(ii) The k_1 and k_2 units are random samples from the populations of units included in the domains covered by the tests.

(iii) In the entire population of examinees, true scores are normally distributed on the two tests. For Tests 1 and 2 the correlation between true scores is greater than or equal to zero.

(iv) Throughout the entire examinees-by-units matrix for Test j (j = 1,2), the errors of measurement associated with the part-test scores are homogeneous in variance and normally distributed. However, the variance of these errors is not necessarily the same for Tests 1 and 2.

(v) Errors of measurement on the parts of the tests are independent of each other and of the true scores, both within and across tests.

These are the usual assumptions associated with the two-factor, random model of analysis of variance. Independence of errors of measurement, within and across tests, is generally assumed within classical test theory.

We first note that the following expected values hold for the various mean squares:

$$E[ms_{S_1}] = E\left[\frac{\tilde{\sigma}^2_{X_1}}{\mathrm{k}_1}\right] = \frac{\sigma^2_{X_1}}{k_1}$$

$$E[ms_{S_2}] = E\left[\frac{\tilde{\sigma}^2_{X_2}}{k_2}\right] = \frac{\sigma^2_{X_2}}{k_2}$$

$$E[ms_{PS_1}] = \sigma^2_{e_1}$$

$$E[ms_{PS_2}] = \sigma^2_{e_2}.$$

In the expressions for the expected values of ms_{S_1} and ms_{S_2}, $\tilde{\sigma}^2_{X_1}$ and $\tilde{\sigma}^2_{X_2}$ are the unbiased estimates of the variances of observed total scores. The error variances $\sigma^2_{e_1}$ and $\sigma^2_{e_2}$ pertain to measurement errors on the individual parts of Tests 1 and 2. Because the tests include k_1 and k_2 parts, respectively, and because the errors are independent, the total error variances for the two tests are $k_1\sigma^2_{e_1}$ and $k_2\sigma^2_{e_2}$

We now consider the test statistic employed by Feldt [1969] for independent groups:

$$W = \frac{1 - r_2}{1 - r_1} = \frac{ms_{PS_2}}{ms_{S_2}} \cdot \frac{ms_{S_1}}{ms_{PS_1}} = \frac{ms_{PS_2}}{ms_{PS_1}} \cdot \frac{ms_{S_1}}{ms_{S_2}} . \tag{3}$$

Substituting $(\tilde{\sigma}^2_{X_j}/k_j)$ for ms_{S_j}, dividing the several mean squares by their expected values and multiplying by these same expected values to preserve the equality, we obtain

$$W = \frac{1 - r_2}{1 - r_1} = \frac{\dfrac{ms_{PS_2}}{\sigma^2_{e_2}}\left(\dfrac{\tilde{\sigma}^2_{X_1}/k_1}{\sigma^2_{X_1}/k_1}\right)\sigma^2_{e_2}\left(\dfrac{\sigma^2_{X_1}}{k_1}\right)}{\dfrac{ms_{PS_1}}{\sigma^2_{e_1}}\left(\dfrac{\tilde{\sigma}^2_{X_2}/k_2}{\sigma^2_{X_2}/k_2}\right)\sigma^2_{e_1}\left(\dfrac{\sigma^2_{X_2}}{k_2}\right)} .$$

We use the fact that $k_j\sigma^2_{e_j}$ equals the total error variance for Test j and that $\rho_j = 1 - (k_j\sigma^2_{e_j})/\sigma^2_{X_j}$. This leads to

$$W = \frac{1 - r_2}{1 - r_1} = \left[\frac{ms_{PS_2}/\sigma^2_{e_2}}{ms_{PS_1}/\sigma^2_{e_1}}\right]\left[\frac{\tilde{\sigma}^2_{X_1}/\sigma^2_{X_1}}{\tilde{\sigma}^2_{X_2}/\sigma^2_{X_2}}\right]\left[\frac{1 - \rho_2}{1 - \rho_1}\right] . \tag{4}$$

The first factor on the right may be recognized as the ratio of two chi-square variables divided by their degrees of freedom. By the assumed independence of errors, this ratio is distributed as a central F with degrees of freedom $(k_2 - 1)$ $(N - 1)$ and $(k_1 - 1)$ $(N - 1)$. With even moderate numbers of parts and a reasonably large sample size, these degrees of freedom will be quite large, 1000 or larger. For all practical purposes, such an F distribution may be considered to be almost totally concentrated at the point $F = 1.0$ [Hogg & Craig, 1970, 187–188], and hence this factor has negligible influence on the distribution of W.

Thus, we have

$$W = \frac{1 - r_2}{1 - r_1} \doteq \frac{1 - \rho_2}{1 - \rho_1} F^*_{N-1,N-1} \tag{5}$$

where F^* is the ratio of *related* variance estimates, each with an expected value of 1.0. In essence, we have arrived at the same point reached by Feldt [1969], except that the statistic W is now seen to involve the ratio of related sample variances. If an observed W is too large or too small to be accepted as a value drawn at random from $F^*_{N-1,N-1}$, one must conclude at the designated significance level that $\rho_1 \neq \rho_2$.

At least two solutions to the problem of evaluating probabilities associated with F^* can be found in the statistical literature. To these we shall add a third. The first solution was published by Bose [1935], and the derivation was later simplified by Finney [1938]. They demonstrated that the probability of F^* exceeding a specific numerical value, F^*_0, may be determined by first transforming F^*_0 as follows:

$$F^{**}_0 = \frac{([F^*_0 + 1]^2 - 4\rho^2 F^*_0)^{1/2} + (F^*_0 - 1)}{([F^*_0 + 1]^2 - 4\rho^2 F^*_0)^{1/2} - (F^*_0 - 1)} . \tag{6}$$

In this expression ρ represents the correlation between the total scores on Tests 1 and 2. Then

$$P[F^*_{N-1,N-1} > F^*_0] = P[F_{N-1,N-1} > F^{**}_0], \tag{7}$$

where $F_{N-1,N-1}$ is the central F with $N - 1$ and $N - 1$ degrees of freedom. In the present context, F^*_0 equals $W = (1 - r_2)/(1 - r_1)$, not the ratio of the variances of the tests under study. The transformation is applied to $(1 - r_2)/(1 - r_1)$.

Strictly speaking, this approach demands knowledge of ρ, a value which is almost always unknown. Substitution of the sample correlation is an obvious possibility, though Finney suggested a more laborious alternative procedure. Empirical simulation data bearing on the use of r in (6) are presented in the next section.

The second solution was derived by Pitman [1939]. It may be shown that the product-moment correlation between $X_1 - X_2$ and $X_1 + X_2$ equals

$$\frac{\sigma^2_{X_1} - \sigma^2_{X_2}}{([\sigma^2_{X_1} + \sigma^2_{X_2}]^2 - 4\sigma^2_{X_1X_2})^{1/2}} = \frac{\dfrac{\sigma^2_{X_1}}{\sigma^2_{X_2}} - 1}{\left(\left[\dfrac{\sigma^2_{X_1}}{\sigma^2_{X_2}} + 1\right]^2 - 4\rho^2 \dfrac{\sigma^2_{X_1}}{\sigma^2_{X_2}}\right)^{1/2}} . \tag{8}$$

Clearly, this correlation will equal zero if and only if the variances are equal. Thus, Pitman proposed that the sample correlation between $X_1 - X_2$ and $X_1 + X_2$, either computed directly from the sums and differences or identically determined by

$$\frac{\left(\dfrac{S^2_{X_1}}{S^2_{X_2}} - 1\right)}{\left(\left[\dfrac{S^2_{X_1}}{S^2_{X_2}} + 1\right]^2 - 4r^2_{X_1X_2} \dfrac{S^2_{X_1}}{S^2_{X_2}}\right)^{1/2}} , \tag{9}$$

could be tested for significance via the usual t-test. As in the previous test, the ratio of variances symbolized in (9) is represented in the present context by $W = (1 - r_2)/(1 - r_1)$. Substitution of the previous expression for the correlation between sums and differences into the formula for testing the significance of a correlation simplifies to

$$t_{N-2} = \frac{(W - 1)(N - 2)^{1/2}}{(4W[1 - r^2_{X_1,X_2}])^{1/2}} . \tag{10}$$

The final derivation of the distribution of F^* draws on the Bose demonstration that the general form of the distribution is that of a central F. Through the use of the Δ method described by Kendall and Stuart [1969, pp. 231–2], the sampling variance of F^*, to order N^{-1}, is found to equal $4(1 - \rho^2_{X_1X_2})/(N - 1)$. The fact that this result is correct only to order N^{-1} implies that a more precise expression is of the form $4(1 - \rho^2_{X_1X_2})/(N - c)$, where c may be taken as a numerical constant. When $\rho_{X_1X_2} = 0$, F^* is distributed as $F_{N-1,N-1}$. A central F with $df_1 = df_2$ has a variance almost exactly equal to $4/(df - 7)$, or $4/(N - 8)$ in this case. This suggests that a more precise expression for the variance of F^* is $4(1 - \rho^2_{X_1X_2})/(N - 8)$.

All central F distributions with large values for the second degrees of freedom have means about equal to 1.0. Therefore, the central $F_{\nu,\nu}$ which has approximately the same mean and variance as F^* can be determined by solving the following equation for ν:

$$\frac{4}{\nu - 7} = \frac{4(1 - \rho^2_{X_1X_2})}{N - 8}$$

$$\nu = \frac{N - 8}{1 - \rho^2_{X_1X_2}} + 7 = \frac{N - 1 - 7\rho^2_{X_1X_2}}{1 - \rho^2_{X_1X_2}} . \tag{11}$$

Using $r_{X_1X_2}$ to estimate $\rho_{X_1X_2}$, we may estimate $P[F^* > W]$ by

$$(12) \qquad P[F_{\nu,\nu} > W],\ \nu = \frac{N - 1 - 7r^2_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

with ν rounded down to a whole number.

To illustrate the three procedures for testing $H_0: \rho_1 = \rho_2$, we take $r_1 = .80$, $r_2 = .72$, $N = 100$, and $r_{X_1X_2} = .50$. For these data $W = (1 - .72)/(1 - .80) = 1.4$. Under the Bose/Finney procedure,

$$F_0^{**} = \frac{[(1.4 + 1)^2 - 4(.25)\,(1.4)]^{\frac{1}{2}} + (1.4 - 1)}{[(1.4 + 1)^2 - 4(.25)\,(1.4)]^{\frac{1}{2}} - (1.4 - 1)} = 1.4739$$

$$P[F^*_{99,99} > 1.4] = P[F_{99,99} > 1.4739] = .0275.$$

If the alternatives to the null hypothesis are $\rho_1 \neq \rho_2$, one would reject H_0 at the 5% level if $P \leq .025$ or $P \geq .975$. If one decides in advance to place the larger of $(1 - r_1)$ and $(1 - r_2)$ in the numerator of W, one would reject if $P < .05$. Presuming the foregoing test was made under this procedure, the experimenter could reject H_0 at the 5% level.

Under the Pitman procedure, the test statistic, in terms of W and $r_{X_1X_2}$, equals

$$t = \frac{(1.4 - 1)\,(98)^{1/2}}{[4(1.4)\,(1 - .25)]^{1/2}} = 1.9322$$

$$P[t_{98} > 1.9322] = .0281.$$

If the alternatives are $\rho_1 \neq \rho_2$, $P \leq .025$ or $P \geq .975$ leads to rejection. If W is defined so that $W > 1.0$, the probability must be less than .05 for rejection to occur.

The final procedure involves $\nu = (N - 1 - 7r^2_{X_1X_2})/(1 - r^2_{X_1X_2}) = 97.25/.75 = 129.7$. The extreme area associated with $W = 1.4$ is

$$P[F_{129,129} > 1.40] = .0285.$$

Again, if one defines W so that $W \geq 1.0$, one must obtain $P \leq .05$ for rejection at the .05 level.

The agreement among the probabilities calculated via the three procedures is very close for this example. The range equals .001, with the two procedures based on F agreeing almost exactly.

Each of these procedures is approximate for one or another reason. Therefore, it seemed important to compare them with regard to control of Type I error. The essential question, beyond those issues already considered by Feldt [1969], is which procedure provides the most accurate probability statements about F^*. To investigate this issue, a series of computer simulation studies was undertaken. These are summarized in the next section.

Computer Simulation of the Distribution of F^*

The technique developed by Odell and Feiveson [1966] was used to generate sample ratios (W) of correlated variances. The simulations included three sample sizes ($N = 50$, 100, and 200) coupled with three levels of population correlation between X_1 and X_2 (.4, .6, and .8). Fifteen thousand ratios were generated for the combination $N = 50$, $\rho = .8$; five thousand ratios were produced for all other combinations N and ρ. The probability $P[F^* > W_0]$, where W_0 is any specific ratio value, was determined and tabulated for each of the three procedures for evaluation of P. For brevity, the results summarized in Table 1 combine the probabilities from both tails of the empirical distributions. They represent empirical estimates of the percent of Type I errors which would have occurred had the

TABLE 1

Empirical Estimates of Type I Error Rates (in Percent)

Nominal Alpha Level	N = 50			N = 100			N = 200		
	ρ=.4	ρ=.6	ρ=.8	ρ=.4	ρ=.6	ρ=.8	ρ=.4	ρ=.6	ρ=.8
Bose/Finney									
10%	11.34	10.04	10.10	10.58	10.54	10.58	9.80	9.86	10.40
5%	5.62	5.10	5.18	5.36	5.26	5.08	5.16	5.04	5.00
1%	0.94	0.98	1.15	1.06	1.00	1.08	1.28	1.06	0.90
Pitman									
10%	11.00	9.76	9.78	10.50	10.32	10.40	9.72	9.86	10.34
5%	5.38	4.90	4.95	5.22	5.14	5.02	5.08	5.00	4.96
1%	0.90	0.90	1.07	0.98	0.96	1.02	1.26	1.06	0.90
Δ Method									
10%	10.82	9.38	8.82	10.44	10.04	10.04	9.72	9.84	9.98
5%	5.42	4.64	4.46	5.22	5.04	4.72	5.10	4.96	4.82
1%	0.92	0.90	0.95	1.02	0.96	0.98	1.28	1.06	0.86

10%, 5%, or 1% level been used to test the null hypothesis against a non-directional alternative. It should be noted that this simulation bears on the adequacy of probability statements about F^*. Therefore, for simplicity the simulation program produced ratios of correlated variances rather than pairs of alpha coefficients.

The summary data suggest that all three methods give tight control of Type I error. The mean of the differences between nominal and empirical significance levels for the 27 values were as follows: 0.20% for the Bose/Finney method, 0.09% for the Pitman method, and −.06% for the Δ method. The mean absolute differences are 0.24% for the Bose/Finney method, 0.19% for the Pitman method, and 0.24% for the Δ method. Since none of the methods is computationally onerous and all may be implemented via available tables or calculators that compute t or F distribution probabilities, there is little basis for choice.

The control of Type I error may be a shade tighter via the Δ and Pitman techniques. On these grounds, these procedures, as summarized by (10) and (12) above, are recommended.

REFERENCES

Bose, S. The distribution of the ratio of variances of two samples drawn from a given normal bivariate correlated population. *Sankhyā: The Indian Journal of Statistics*, 1935, *1*, 65–72.

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

Feldt, L. S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, 1969, *34*, 363-373.

Finney, D. J. The distribution of the ratio estimates of the two variances in a sample from a normal bivariate population. *Biometrika*, 1938, *30*, 190–192.

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

Hogg, R. V. & Craig, A. T. *Introduction to mathematical statistics* (3rd ed.). New York: Macmillan, 1970, 187–188.

Kendall, M. G. & Stuart, A. *The advanced theory of statistics* (Vol. 1, 3rd ed.). London: Charles Griffin, 1969.

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

Odell, P. L. & Feiveson, A. H. A. A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association*, 1966, *61*, 199–203.

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

Manuscript received 2/22/79
Final version received 8/31/79