Open access • Journal Article • DOI:10.1177/014662169201600208

Test of the Hypothesis that the Intraclass Reliability Coefficient Is the Same for Two Measurement Procedures. — Source link

Yousef M. Alsawalmeh, Leonard S. Feldt

Institutions: Yarmouk University

Published on: 01 Jun 1992 - Applied Psychological Measurement (SAGE Publications)

Topics: Test statistic, Statistical hypothesis testing, Type I and type II errors, Sampling distribution and Chi-square test

Test of the Hypothesis That the Intraclass Reliability Coefficient is the Same for Two Measurement Procedures

Yousef M. Alsawalmeh, Yarmouk University

Leonard S. Feldt, University of Iowa

An approximate statistical test is derived for the hypothesis that the intraclass reliability coefficients associated with two measurement procedures are equal. Control of Type 1 error is investigated by comparing empirical sampling distributions of the test statistic with its derived theoretical distribution. A numerical illustration of the procedure is also presented. *Index terms: intraclass reliability, reliability, sampling theory, Spearman-Brown extrapolation, statistical test.*

Comparison of two reliability coefficients has been, and will continue to be, a major focus of many measurement studies in education and psychology. Feldt (1980) cited several examples drawn from behavioral research in which a test of the equality of reliability coefficients is required. Investigators who need to compare the values of Cronbach's alpha obtained from two measurement procedures have techniques to make these comparisons in both the independent (Feldt, 1969) and dependent (Feldt, 1980) case. However, at present there is no direct test available for the equality of two intraclass reliability coefficients. This paper presents the derivation of such a test.

The intraclass correlation is a reliability coefficient with a broad range of applications. These include not only the assessment of the reliability of subtests, but also interobserver, test-retest, and equivalent forms reliability (Bartko, 1976; Haggard, 1958). Researchers often have employed the intraclass correlation to estimate the reliability of a rating by a single observer or the score of a single trial by an examinee on a performance measure (Bartko, 1966; Baumgartner & Jackson, 1987; Ebel, 1951; Shrout & Fleiss, 1979). It also can be used to estimate and compare the reliabilities of tests differentially shortened or lengthened to fit within the time period allotted for measurement. The importance of testing time, rather than number of items, as an important measure of test length is noted by Lord and Novick (1968, pp. 118–125) and Feldt (1989, pp. 116–117). Feldt (1990) presented general educational settings for which the intraclass correlation is the relevant reliability coefficient.

The intraclass reliability coefficient can be viewed as a "reverse" Spearman-Brown extrapolation from the reliability of k measurements, $\hat{\alpha}$, to the reliability of a single measurement, $\hat{\rho}$. Specifically, the relationship is

$$\hat{\rho} = \frac{\hat{\alpha}}{k - (k-1)\hat{\alpha}} \quad . \tag{1}$$

Using this relation, Kraemer (1981) extended Feldt's procedures for testing H_o: $\alpha_1 = \alpha_2$ to a test of H_o: $\rho_1 = \rho_2$. However, this extension applies only when the two measurement procedures have the same number of parts (i.e., $k_1 = k_2$). Earlier, Schumann and Bradley (1957, 1959) considered

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 2, June 1992, pp. 195-205

0146-6216/92/020195-11$1.80

essentially the same hypothesis in the context of comparing the sensitivity of two experiments. However, the theoretical solution when $k_1 \neq k_2$ proved exceedingly complex. They provided necessary tables that were limited to $k_1 = k_2$ and $N_1 = N_2$.

For the case of unequal numbers of measurements, an exact test for $\rho_1 = \rho_2$ cannot be obtained (Bross, 1959). Approximate methods have been developed (Donner, 1986; Donner & Bull, 1983), but these methods require an iterative solution, and they are inaccurate when the common intraclass coefficient is large. For more than two measurements, the approximation becomes even poorer (Fisher, 1970, p. 221). The present paper extends the statistical methodology so that the parameter values of two independent intraclass reliability coefficients can be compared. This test, in effect, is a test of the equality of two reliability coefficients after adjustment for the unequal length of the instruments that were employed in the reliability study. It is required in investigations in which test length or testing time should be controlled, but the researcher is unable to do so experimentally.

A Test of the Hypothesis H_0: $\rho_1 = \rho_2$

Let $X_{i1}, X_{i2}, \ldots, X_{ik}$ denote the observable continuous scores on k measures, all supposedly measuring the same trait for person i. The k measures are presumed to be randomly selected from the population of scores for person i; person i is randomly selected from a large population of persons. The k measures may differ in their means for the population of persons but are homogeneous in variance. The measures are administered to a random sample of size N. The test score for person i on measure j can be written as a linear model,

$$X_{ij} = \mu + \tau_i + \beta_j + e_{ij} \quad , \qquad (i = 1, \ldots, N;\ j = 1, \ldots, k) \quad . \tag{2}$$

In this model, μ is the expected value of the overall mean of all kN measures; τ_i is a random variable equal to the expected value of $(X_{ij} - \mu)$ over an infinite number of measures on person i, and it indicates person i's true trait level (examinee effect); β_j is the expected value of $(X_{ij} - \mu)$ over an infinite number of examinees, and indicates the relative difficulty level of the jth measure (measure effect); and e_{ij} is a random error score (interaction effect of measure j with person i plus random error from all other sources).

It is assumed that the kN scores conform to the assumptions of a two-factor random model analysis of variance (ANOVA). The quantities τ_i, β_j, and e_{ij} are assumed to be pairwise independent and

$$\tau_i \sim \text{NID}(0, \sigma_\tau^2) \quad , \tag{3}$$

$$\beta_j \sim \text{NID}(0, \sigma_\beta^2) \quad , \tag{4}$$

and

$$e_{ij} \sim \text{NID}(0, \sigma_e^2) \quad . \tag{5}$$

The notation $\text{NID}(\mu, \sigma^2)$ signifies a normally and independently distributed random variable with mean μ and variance σ^2. Under these assumptions the following expectations (E) hold:

$$\text{E}(\text{MS}_p) = \sigma_e^2 + k\sigma_\tau^2 \tag{6}$$

and

$$\text{E}(\text{MS}_{m \times p}) = \sigma_e^2 \quad . \tag{7}$$

In these expectations, MS_p is the mean square for persons, and $\text{MS}_{m \times p}$ is the mean square for the

measures × persons interaction derived from a persons × measures ANOVA with one observation per cell.

From Equation 1 and the ANOVA expression for coefficient alpha, the estimate of the intraclass reliability, $\hat{\rho}$, is given by

$$\hat{\rho} = \frac{MS_p - MS_{m\times p}}{MS_p + (k-1)MS_{m\times p}} = 1 - \frac{kMS_{m\times p}}{MS_p + (k-1)MS_{m\times p}} \quad . \tag{8}$$

The population value of the intraclass reliability is defined by

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} \quad . \tag{9}$$

The sampling distribution of $\hat{\rho}$ can be defined by derivation of the sampling distribution of the second term of the right-hand expression in Equation 8. The denominator of this term is a linear combination of independent mean squares. Following Satterthwaite (1941, 1946), the distribution of this denominator can be approximated by the following transformation of a χ^2 distribution with effective v degrees of freedom (*df*):

$$MS_p + (k-1)MS_{m\times p} \doteq \frac{E[MS_p + (k-1)MS_{m\times p}]}{v}\chi_v^2 \doteq \frac{k(\sigma_\tau^2 + \sigma_e^2)}{v}\chi_v^2 \quad , \tag{10}$$

where

$$v = \frac{(N-1)k}{1 + (k-1)\rho^2} \quad . \tag{11}$$

In practice, the population value of the intraclass correlation, ρ, is unknown and must be estimated by the sample intraclass correlation, $\hat{\rho}$. Substitution of the sample value of the intraclass reliability results in the following estimate of v:

$$\hat{v} = \frac{(N-1)k}{1 + (k-1)\hat{\rho}^2} \quad . \tag{12}$$

This approximation adds to the approximate character of the distribution defined by Equation 10 and makes empirical validation of the ultimate statistical test more necessary.

Under normal theory, it can be shown (e.g., Scheffé, 1959, p. 243) that

$$(N-1)(k-1)k\,MS_{m\times p} \sim k\sigma_e^2\chi^2_{(N-1)(k-1)} \quad . \tag{13}$$

Thus, from Equations 10 and 13,

$$\frac{k\,MS_{m\times p}}{MS_p + (k-1)MS_{m\times p}} \doteq \frac{[k\sigma_e^2\chi^2_{(N-1)(k-1)}]/(N-1)(k-1)}{[k(\sigma_\tau^2 + \sigma_e^2)]\chi_v^2/v} \doteq [\sigma_e^2/(\sigma_\tau^2 + \sigma_e^2)]F^*_{(N-1)(k-1),v} \quad . \tag{14}$$

An equivalent statement is

$$1 - \hat{\rho} \doteq (1-\rho)F^*_{(N-1)(k-1),v} \quad . \tag{15}$$

In these expressions, $F^*_{n,m}$ denotes a random variable equal to the ratio of two nonindependent χ^2 variables, each divided by its *df*.

Suppose there are now two measurement procedures. The first involves k_1 measures, all measuring the same trait, administered to a random sample of size N_1. The second involves k_2 measures, again all measuring the same trait, administered to an independent random sample of size N_2. The units of measurement of the two sets of scores obtained from the two procedures may not be directly comparable—the score distributions may have different means and variances. If the scores obtained from each procedure conform to the assumptions of a two-factor random model ANOVA, then the relationship in Equation 15 applies to both $\hat{\rho}_1$ and $\hat{\rho}_2$, which represent the intraclass reliability estimates for the first and second procedures, respectively. Therefore,

$$1 - \hat{\rho}_1 \doteq (1 - \rho_1)F^*_{c_1,\nu_1} \tag{16}$$

and

$$1 - \hat{\rho}_2 \doteq (1 - \rho_2)F^*_{c_2,\nu_2} \quad , \tag{17}$$

where

$$c_1 = (N_1 - 1)(k_1 - 1) \quad , \tag{18}$$

$$c_2 = (N_2 - 1)(k_2 - 1) \quad , \tag{19}$$

$$\nu_1 = \frac{(N_1 - 1)k_1}{1 + (k_1 - 1)\hat{\rho}_1^2} \quad , \tag{20}$$

and

$$\nu_2 = \frac{(N_2 - 1)k_2}{1 + (k_2 - 1)\hat{\rho}_2^2} \quad . \tag{21}$$

A promising test statistic for H_o: $\rho_1 = \rho_2$ is

$$T = \frac{1 - \hat{\rho}_1}{1 - \hat{\rho}_2} \doteq \frac{(1 - \rho_1)F^*_{c_1\nu_1}}{(1 - \rho_2)F^*_{c_2\nu_2}} \quad , \tag{22}$$

because under H_o, T equals the ratio $F^*_{c_1,\nu_1}/F^*_{c_2,\nu_2}$ (i.e., the ratio of two independent F^* variables). An accurate approximation to the distribution of this ratio could be used to conduct the test of the null hypothesis H_o.

It can be shown that the ratio of these two independent F^* variables is equivalent to the product $(F^*_{c_1,c_2})$ $(F^*_{\nu_2,\nu_1})$, where the first F^* variable is not independent of the second F^* variable. However, each F^* variable is the ratio of independent χ^2 variables. If the sample sizes and numbers of measures are reasonably large, the *df*, c_1, and c_2 will be quite large—1,000 or more. For very large *df*, the value of F_{c_1,c_2} will deviate only slightly from its expected value, which is very close to 1.0 (Hogg & Craig, 1978, pp. 196–198). Hence, F_{c_1,c_2} has negligible influence on the distribution of T, and the distribution of T under H_o will be dictated almost entirely by F_{ν_2,ν_1}.

A slight adjustment in the *df* will result in a closer approximation of the distribution of T. This closer approximation is represented by that central F for which, under H_o, $E(F) = E(T)$ and $\text{Var}(F) = \text{Var}(T)$. The exact mean and variance of T are unknown under H_o, but estimates of these moments can be obtained by using the Δ method described by Kendall and Stuart (1977, pp. 246–262). Note that T in this case is given by the ratio $F^*_{c_1,\nu_1}/F^*_{c_2,\nu_2}$.

To obtain the expected value and the sampling variance of any ratio, the following two formulas

given by Kendall and Stuart (1977, pp. 247–260) can be applied:

$$\mathrm{E}\left[\frac{X_1}{X_2}\right] = \frac{\mathrm{E}(X_1)}{\mathrm{E}(X_2)} + \frac{\mathrm{E}(X_1)\mathrm{Var}(X_2)}{\mathrm{E}^3(X_2)} - \frac{\mathrm{Cov}(X_1,X_2)}{\mathrm{E}^2(X_2)} \tag{23}$$

and

$$\mathrm{Var}\left[\frac{X_1}{X_2}\right] = \left[\frac{\mathrm{E}(X_1)}{\mathrm{E}(X_2)}\right]^2\left[\frac{\mathrm{Var}(X_1)}{\mathrm{E}^2(X_1)} + \frac{\mathrm{Var}(X_2)}{\mathrm{E}^2(X_2)} - \frac{2\mathrm{Cov}(X_1,X_2)}{\mathrm{E}(X_1)\mathrm{E}(X_2)}\right] \quad . \tag{24}$$

Through the use of the Δ method, the mean and the sampling variance of $F_1 = F^*_{c_1,\nu_1}$ can be estimated to the order N^{-1}. These estimates are

$$\mathrm{E}(F_1) = 1 + \frac{2}{\nu_1} - \frac{2(1-\rho_1)}{k_1(N_1-1)} \tag{25}$$

and

$$\mathrm{Var}(F_1) = \frac{2}{c_1} + \frac{2}{\nu_1} - \frac{4(1-\rho_1)}{k_1(N_1-1)} \quad . \tag{26}$$

The fact that these estimates are correct only to the order N^{-1} implies that the more precise expressions are of the form

$$\mathrm{E}(F_1) = 1 + \frac{2}{\nu_1} - \frac{2(1-\rho_1)}{k_1(N_1-1)} + L_1 \tag{27}$$

and

$$\mathrm{Var}(F_1) = \frac{2}{c_1} + \frac{2}{\nu_1} - \frac{4(1-\rho_1)}{k_1(N_1-1)} + L_2 \quad , \tag{28}$$

where L_1 and L_2 are functions of the *df*. When the two χ^2s are independent, $F^*_{c_1,\nu_1}$ is distributed as a central F distribution with c_1 and ν_1 *df*. The mean of this F is $\nu_1/(\nu_1 - 2)$ with variance

$$2\nu_1^2(c_1 + \nu_1 - 2)/[c_1(\nu_1 - 2)^2(\nu_1 - 4)] \quad , \tag{29}$$

where $\nu_1 > 4$. This suggests that more precise expressions of the mean and variance of F_1 are

$$\mathrm{E}(F_1) = \frac{\nu_1}{\nu_1 - 2} - \frac{2(1-\rho_1)}{k_1(N_1-1)} \tag{30}$$

and

$$\mathrm{Var}(F_1) = \frac{2\nu_1^2(c_1 + \nu_1 - 2)}{c_1(\nu_1 - 2)^2(\nu_1 - 4)} - \frac{4(1-\rho_1)}{k_1(N_1-1)} \quad . \tag{31}$$

Application of the same methodology to $F_2 = F^*_{c_2,\nu_2}$ provides estimates of its mean and variance:

$$\mathrm{E}(F_2) = \frac{\nu_2}{\nu_2 - 2} - \frac{2(1-\rho_2)}{k_2(N_2-1)} \tag{32}$$

and

$$\mathrm{Var}(F_2) = \frac{2\nu_2^2(c_2 + \nu_2 - 2)}{c_2(\nu_2 - 2)^2(\nu_2 - 4)} - \frac{4(1 - \rho_2)}{k_2(N_2 - 1)} \quad . \tag{33}$$

Thus, from Equations 20 and 21, the estimates M and V of $\mathrm{E}[T]$ and $\mathrm{Var}[T]$, respectively, are given by

$$M = \frac{\mathrm{E}(F_1)}{\mathrm{E}(F_2)} + \frac{\mathrm{E}(F_1)\mathrm{Var}(F_2)}{\mathrm{E}^3(F_2)} - \frac{\mathrm{Cov}(F_1,F_2)}{\mathrm{E}^2(F_2)} \tag{34}$$

and

$$V = \left[\frac{\mathrm{E}(F_1)}{\mathrm{E}(F_2)}\right]^2\left[\frac{\mathrm{Var}(F_1)}{\mathrm{E}^2(F_1)} + \frac{\mathrm{Var}(F_2)}{\mathrm{E}^2(F_2)} - \frac{2\mathrm{Cov}(F_1,F_2)}{\mathrm{E}(F_1)\mathrm{E}(F_2)}\right] \quad . \tag{35}$$

In the present situation, F_1 and F_2 are derived from independent samples. Therefore, $\mathrm{Cov}(F_1,F_2)$ is equal to 0. Consequently, M and V are completely determined.

Given approximations of the expected value (M) and the variance (V), the *df* for the numerator (d_1) and denominator (d_2) of the desired F distribution can be obtained by solving the following pair of equations:

$$M = \frac{d_2}{d_2 - 2} \tag{36}$$

and

$$V = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)} \quad . \tag{37}$$

The resultant values are

$$d_2 = \frac{2M}{M - 1} \tag{38}$$

and

$$d_1 = \frac{2d_2^3 - 4d_2^2}{V(d_2 - 2)^2(d_2 - 4) - 2d_2^2} \quad , \tag{39}$$

which completes the solution.

Thus,

$$T = \frac{1 - \hat{\rho}_1}{1 - \hat{\rho}_2} \doteq \left(\frac{1 - \rho_1}{1 - \rho_2}\right)E_{d_1,d_2} \quad . \tag{40}$$

If an observed T is too large or too small to be accepted as a value drawn at random from the F distribution with d_1 and d_2 *df*, then $\rho_1 \neq \rho_2$ is the conclusion at the designated significance level.

Empirical Investigation of the Procedure by Computer Simulation

Method

The test statistic derived above is approximate because several approximations were used in its

derivation. For this reason, monte carlo simulations were necessary in order to assess its accuracy in controlling Type 1 error. Type 1 error refers to the proportion of rejections when the null hypothesis is true.

One way to generate values of T under H_o is to directly generate a sample covariance matrix for each measurement procedure, without generating responses at the level of single observations. Coefficient alpha and the intraclass reliability can be computed from each sample covariance matrix as usual. Then, T can be computed from each pair of intraclass reliabilities.

Odell and Feiveson (1966) developed a numerical procedure to directly generate a sample covariance matrix. Implementation of the procedure was simplified by Browne (1968). Because of its efficiency, the Odell/Feiveson/Browne method was used here. Programming was done in IBM VS FORTRAN, using double precision arithmetic. The following subroutines from the International Mathematical and Statistical Library (IMSL, 1987) were used: FDF, LFTDS, MXTYF, MXYTF, RNCHI, RNGET, RNNOR, RNSET.

The first step in generating simulation data is the adoption of arbitrary but reasonable parameter values. Table 1 presents 12 combinations of values of N and k that were used. Note that the first five conditions (primary conditions) are more consistent with the constraints imposed by the derivation of the test statistics. However, the remaining seven conditions (secondary conditions) are more consistent with the practical realities of measurement in educational and psychological situations.

Simulation also requires parameter values for the intraclass correlations and their corresponding population covariance matrices. Type 1 error rates were estimated under four different intraclass values: .2, .3, .4, and .5. These intraclass correlations and k_j values cover a wide range of population alpha coefficients (.56 to .91) that include most practical situations in which comparison of test reliabilities is of interest.

These parameter values were used to generate simulation data for evaluation of Type 1 error rates. Each simulation consisted of a $k_1 \times k_1$ matrix and a $k_2 \times k_2$ matrix from which the test statistic T was computed. For every combination of parameters (ρ, k, N), the simulation process was replicated 4,000 times. Each of the resultant empirical sampling distributions for T was examined for evidence of control of Type 1 error at the three most widely adopted significance levels (.10, .05, and .01).

Results

Table 1 indicates that T offers accurate control of Type 1 error rates for all the primary conditions and situations involved in the study. In the secondary situations, T tends to be a bit liberal, with as much as 6.5% (rather than the nominal 5%) rejections and 1.9% (rather than the nominal 1%) rejections of true null hypotheses. T controls Type 1 error most accurately with equal sample sizes or relatively small differences between N_1k_1 and N_2k_2. For example, the average of the 12 absolute deviations across the four reliability levels of Condition 12 ($N_1 = 100$, $N_2 = 100$, $k_1 = 5$, $k_2 = 7$) was .0024. This result suggests that T might perform quite well in cases with even smaller numbers of observations (e.g., $k_1 = 2$, $k_2 = 3$).

Recall that in the derivation of the sampling distribution of T, if $(N - 1)(k - 1) > 1{,}000$ for both measurement procedures, then the distribution of T is dictated almost entirely by F_{ν_2,ν_1}. Supplemental monte carlo studies verified this conclusion. When $(N)(k)$ is large, F_{ν_2,ν_1} provides a very adequate model interpretation of T. However, when N and k are relatively small, as they typically are in studies of raters or observers, the more complex determination of the df is recommended.

The above results led to an investigation of how the procedure would perform in cases with even smaller numbers of observations. The study was expanded to include cases with $k_1 = 2$ and $k_2 = 3$. This investigation was carried out for two true null hypothesis situations ($\rho = .4$ and $\rho = .5$). The results are presented in Table 2. They indicate that T does achieve sufficient accuracy in these two

Table 1
Empirical Proportions of Rejections for Four True Null Hypothesis Situations Under 12 Different Combinations of N and k Values (N_1, N_2, k_1, k_2) and Nominal Levels of .10, .05, and .01, With Standard Errors of .0047, .0034, and .0016, Respectively

Condition, N_1, N_2, k_1, k_2	ρ = .2			ρ = .3			ρ = .4			ρ = .5		
	.10	.05	.01	.10	.05	.01	.10	.05	.01	.10	.05	.01
Condition 1												
100, 200, 10, 5	.095	.049	.009	.098	.047	.011	.108	.057	.013	.095	.050	.010
Condition 2												
100, 200, 10, 7	.101	.054	.010	.107	.058	.016	.106	.053	.013	.109	.060	.013
Condition 3												
200, 200, 5, 7	.096	.046	.008	.100	.051	.011	.099	.051	.010	.109	.053	.013
Condition 4												
200, 200, 5, 10	.096	.045	.010	.100	.054	.010	.104	.051	.010	.112	.058	.013
Condition 5												
200, 200, 7, 10	.096	.046	.008	.099	.048	.010	.109	.055	.012	.103	.054	.014
Condition 6												
100, 200, 7, 10	.099	.056	.015	.111	.061	.016	.115	.065	.016	.109	.061	.016
Condition 7												
100, 200, 5, 10	.101	.056	.014	.111	.061	.016	.115	.065	.018	.109	.064	.019
Condition 8												
100, 200, 5, 7	.106	.053	.014	.108	.057	.013	.105	.060	.014	.106	.059	.015
Condition 9												
100, 100, 5, 10	.104	.057	.010	.103	.053	.012	.104	.055	.013	.101	.052	.011
Condition 10												
100, 100, 7, 10	.103	.052	.006	.103	.057	.016	.104	.053	.011	.095	.048	.008
Condition 11												
100, 200, 7, 5	.095	.053	.010	.108	.061	.014	.107	.060	.013	.104	.056	.014
Condition 12												
100, 100, 5, 7	.102	.053	.009	.108	.050	.009	.102	.054	.013	.102	.053	.010

Table 2
Empirical Proportions of Rejections for Two True Null Hypotheses ($k_1 = 2$, $k_2 = 3$) for 4,000 Replications At Nominal Levels of .10, .05, and .01, With Standard Errors of .0047, .0034, and .0016, Respectively

				Nominal Levels		
Condition	ρ	N_1	N_2	.10	.05	.01
1	.40	100	100	.107	.049	.014
2	.40	100	200	.094	.049	.011
3	.40	200	100	.094	.043	.008
4	.40	200	200	.100	.053	.010
5	.50	100	100	.103	.055	.010
6	.50	100	200	.103	.053	.016
7	.50	200	100	.104	.053	.014
8	.50	200	200	.104	.053	.011

hypothesis situations. The performance of the statistical test tends to be better with equal rather than unequal sample sizes, and is quite adequate even in the cases in which $N_1 = N_2 = 100$.

These empirical analyses on the proposed statistical test were carried out on normally distributed data. The effects of platykurtosis should be investigated, however, because this condition commonly characterizes reliable test scores (Lord, 1955). Feldt (1969) showed that platykurtosis tended to lower the probability of Type 1 error for his test of $\alpha_1 = \alpha_2$ based on independent samples. This tendency toward conservatism probably holds as well for the test of $\rho_1 = \rho_2$, but this expectation must be confirmed by future monte carlo studies. It is also pertinent to observe that dichotomously-scored test items almost certainly fail to meet the assumption of homogeneity of variance assumed by the proposed test. Therefore, this test is not recommended for comparing the reliabilities of single, dichotomously-scored exercises.

Summary and Numerical Illustration of the Proposed Test

The test of the equality of two independent intraclass reliability coefficients was derived under the conditions that the measurements conform to the assumptions of a two-way random model ANOVA. This model specifies that the errors are independent of the true score and of each other within the measurement procedure and across the procedures. The derived test statistic has an approximate F distribution with d_1 and d_2 df.

The values d_1 and d_2 are calculated using the following four-step process:

1. Compute v_1 and v_2 from Equations 20 and 21.
2. Compute $E(F_1)$ and $Var(F_1)$ from Equations 30 and 31, $E(F_2)$ and $Var(F_2)$ from Equations 32 and 33, and c_1 and c_2 using Equations 18 and 19.
3. Compute the mean and the variance using Equations 34 and 35, based on independent samples with $Cov(F_1,F_2 = 0)$.
4. Compute d_1 and d_2 from Equations 38 and 39.
5. Compute $T = (1 - \hat{\rho}_1)/(1 - \hat{\rho}_2)$ and determine $P[F_{d_1,d_2} > T]$. If this probability is greater than $1 - \alpha/2$ or less than $\alpha/2$, reject H_o: $\rho_1 = \rho_2$ at the α level of significance.

A numerical illustration of the steps involved in identifying the appropriate F distribution follows. Suppose that $\hat{\rho}_1 = .30$, $N_1 = 101$, $k_1 = 5$, $\hat{\rho}_2 = .50$, $N_2 = 101$, and $k_2 = 3$. Then,

$$T = \frac{.70}{.50} = 1.40,\ c_1 = 400,\ c_2 = 200 \quad , \tag{41}$$

$$v_1 = \frac{(100)(5)}{1 + 4(.30)^2} = 368 \quad , \tag{42}$$

$$v_2 = \frac{(100)(3)}{1 + 2(.50)^2} = 200 \quad , \tag{43}$$

$$\mathrm{E}(F_1) = \frac{368}{366} - \frac{2(1 - .30)}{5(100)} = 1.0027 \quad , \tag{44}$$

$$\mathrm{E}(F_2) = \frac{200}{198} - \frac{2(1 - .50)}{3(100)} = 1.00677 \quad , \tag{45}$$

$$\mathrm{Var}(F_1) = \frac{2(368)^2(368 + 400 - 2)}{400(366)^2(364)} - \frac{4(1 - .30)}{5(100)} = .00504 \quad , \tag{46}$$

$$\mathrm{Var}(F_2) = \frac{2(200)^2(200 + 200 - 2)}{200(198)^2(196)} - \frac{4(1 - .50)}{3(100)} = .01405 \quad , \tag{47}$$

$$M = \frac{1.0027}{1.00677} + \frac{(1.0027)(.01405)}{(1.00677)^3} = 1.00976 \quad , \tag{48}$$

$$V = \left[\frac{1.0027}{1.00677}\right]^2 \left[\frac{.01405}{(1.00677)^2} + \frac{.00504}{(1.0027)^2}\right] = .01872 \quad , \tag{49}$$

$$d_2 = 2(1.00976)/.00976 = 207 \quad , \tag{50}$$

$$d_1 = \frac{2(207)^3 - 4(207)^2}{(.01872)(205)^2(203) - 2(207)^2} = 237 \quad , \tag{51}$$

and

$$P(F_{237,207} > 1.40) = .007 \quad . \tag{52}$$

References

Bartko, J. J. (1966). The intraclass correlation coefficients as a measure of reliability. *Psychological Reports, 19,* 3–11.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83,* 762–765.

Baumgartner, T. A., & Jackson, A. S. (1987). *Measurement for evaluation in physical education and exercise science* (3rd ed.). Dubuque IA: W. C. Brown.

Bross, I. D. J. (1959). Note on an application of the Schumann-Bradley table. *Annals of Mathematical Statistics, 30,* 220–238.

Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika, 33,* 267–334.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review, 54,* 67–82.

Donner, A., & Bull, S. B. (1983). Inferences concerning a common intraclass correlation coefficient. *Biometrics, 39,* 771–776.

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

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, 49,* 99–105.

Feldt, L. S. (1989). Reliability. In R. L. Linn (Ed.),

Educational Measurement (3rd ed.) (pp. 116–117). New York: Macmillan.

Feldt, L. S. (1990). The sampling theory for the intraclass reliability coefficient. *Applied Measurement in Education, 3,* 361–367.

Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). Darien CT: Hafner Publishing Company.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance.* New York: Dryden Press.

Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed.). New York: Macmillan.

IMSL. (1987). *International mathematical and statistical libraries* (10th ed.). Houston: Author.

Kendall, M. G., & Stuart A. (1977). *The advanced theory of statistics* (Vol. 1, 4th ed.). New York: MacMillan.

Kraemer, H. C. (1981). Extensions of Feldt's approach to testing homogeneity of coefficients of reliability. *Psychometrika, 45,* 41–45.

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15,* 383–389.

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

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

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

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Schumann, D. E. W., & Bradley, R. A. (1957). The comparison of the sensitivities of similar experiments: Theory. *Annals of Mathematical Statistics, 28,* 902–920.

Schumann, D. E. W., & Bradley, R. A. (1959). The comparison of the sensitivities of similar experiments: Model II of the analysis of variance. *Biometrics, 15,* 405–416.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Author's Address

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