Testing the Equality of Two Related Intraclass Reliability Coefficients

Yousef M. Alsawalmeh, Yarmouk University

Leonard S. Feldt, University of Iowa

An approximate statistical test of the equality of two intraclass reliability coefficients based on the same sample of people is derived. Such a test is needed when a researcher wishes to compare the reliability of two measurement procedures and both procedures can be applied to the performances or products of the same group of individuals. A numerical example is presented. Monte carlo studies indicate that the proposed test effectively controls Type I error with as few as two or three measurements on each of 50 people. *Index terms: equality of related intraclass reliability coefficients, intraclass reliability, sampling theory, Spearman-Brown extrapolation, statistical test.*

In many educational and psychological settings the reliability of a measurement procedure may be investigated with multiple raters, but the operational use of the procedure may be restricted to a single rater. For example, the reliability of a behavior rating scale for children may be studied with two, three, or four raters, but ongoing applications of the instrument may be limited to scores from a single rater. The reliability coefficient applicable to the single-rater situation is extrapolated from multiple-rater data and is referred to as the intraclass coefficient. It is easily obtained using analysis of variance (ANOVA) or generalizability methods and has many applications in the behavioral sciences (Bartko, 1966; Baumgartner & Jackson, 1987; Brennan, 1983; Ebel, 1951; Feldt, 1990; Lindquist, 1953, chap. 16; Shrout & Fleiss, 1979).

The need to compare two intraclass coefficients is encountered in a variety of contexts. For example, an experimenter may wish to compare the reliability of two evaluation approaches—analytical versus holistic—in the assessment of the quality of essays or performances. In the analytical approach, k_1 raters may be used, and in the holistic approach k_2 raters may be used. For an equitable comparison between methods, the number of raters must be equal because reliability is a function of the number of measures or ratings obtained on each person. But the experimental circumstances may not permit an equal number of raters to be used. In this case, a valid comparison requires consideration of the intraclass or single-rater (judge, observer, or scorer) reliabilities for the two approaches.

In some settings, it is necessary or advisable for experimental reasons to investigate the reliabilities with entirely independent groups of raters and people. In others, however, several procedures can be applied using the same sample of raters and people. In some applications, not only the same people but even the same sample of performances or products may be evaluated by the two measurement procedures. For example, a sample of videotaped performances or written compositions could be scored by one panel of raters using analytical techniques and by a second set of raters using the holistic approach. When the same sample of people and/or the same raters are used, the two intraclass estimates are statistically dependent or related. Rigorous comparisons require a statistical test that recognizes that the intraclass estimates, $\hat{\rho}_1$ and $\hat{\rho}_2$, are dependent in the statistical sense.

If the same number of raters is used in the two measurement procedures, their reliabilities may be compared statistically using procedures that have been developed for coefficient alpha (Feldt, 1969, 1980). In this

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 2, June 1994, pp. 183-190

0146-6216/94/020183-08$1.65

special case, rejection of the equality of alpha coefficients is equivalent to rejection of the hypothesis of equality of the intraclass coefficients. However, when the number of raters, k_1 and k_2, differs, no exact test of $\rho_1 = \rho_2$ is possible (Bross, 1959). Alsawalmeh & Feldt (1992) derived an approximate F test for independent values of $\hat{\rho}_1$ and $\hat{\rho}_2$. Their approach controlled Type I error quite precisely even in the limiting case of small numbers of raters or ratings. However, when the estimates are computed from the same sample of people or raters, the Alsawalmeh/Feldt procedure cannot be used validly. The assumptions of independence on which this test is based render it inappropriate for dependent coefficients. The purpose of this paper is to derive an approximate statistical test of the equality of two dependent coefficients and to report the results of sampling studies that investigated the test's control of Type I error rates.

The Independent Groups Test

The sampling theory for the test based on independent groups draws on the ANOVA approach to reliability estimation. Under this approach, the intraclass reliability coefficient is viewed as a reverse Spearman-Brown extrapolation from the reliability of k measurements or coefficient alpha ($\hat{\alpha}$) to the reliability of a single measurement ($\hat{\rho}$). The relationship between $\hat{\alpha}$ and the intraclass coefficient is

$$\hat{\rho} = \frac{\hat{\alpha}}{k-(k-1)\hat{\alpha}} = 1 - \frac{k\mathrm{MS}_{\mathrm{m\times p}}}{\mathrm{MS}_{\mathrm{p}} + (k-1)\mathrm{MS}_{\mathrm{m\times p}}}, \tag{1}$$

where $\mathrm{MS}_{\mathrm{m\times p}}$ is the mean square for the measures $\times$ persons interaction, and MS_{p} is the mean square for persons in a two-factor random ANOVA model. In this model, a single score is represented as

$$x_{ih} = \mu + \tau_i + \beta_h + e_{ih}, \tag{2}$$

where $i = 1, \ldots, N$ indexes persons and $h = 1, \ldots, k$ indexes ratings. The parameter μ is the expected value of the mean of all kN measures. The person effect, τ_i, is a random variable equal to the expected value of $(X_{ih} - \mu)$ over an infinite number of measures on person i. The measure effect or relative difficulty of the hth measure, β_h, is the expected value of $(X_{ih} - \mu)$ over an infinite number of people. The quantity e_{ih} is a random error of measurement, that is, the interaction effect of measure h with person i plus random error from all other sources. The quantities τ_i, β_h, and e_{ih} are assumed to be pairwise independent, normally distributed, and homogeneous in variance.

Alsawalmeh & Feldt (1992) showed that if the components of the k ratings on each of N people satisfy the above assumptions, the quantity $1-\hat{\rho}$ is distributed approximately as $(1-\rho)F^*_{(N-1)(k-1)\upsilon}$. In this expression, $F^*_{n,m}$ denotes a random variable equal to the ratio of two nonindependent χ^2 variables, each divided by its degrees of freedom (df). υ is the effective df associated with the synthetic mean square (Satterthwaite, 1941) that approximates the distribution of the denominator of the last term on the right side of Equation 1.

Building on this theory, Alsawalmeh & Feldt (1992) noted that when the null hypothesis $H_0: \rho_1 = \rho_2$ is true, the statistic

$$T = (1-\hat{\rho}_1)/(1-\hat{\rho}_2) \tag{3}$$

is distributed approximately as the ratio $F^*_{c_1,\upsilon_1} \big/ F^*_{c_2,\upsilon_2}$, that is, as the ratio of two independent F^* variables. In the case of independent intraclass coefficients,

$$c_1 = (N_1 - 1)(k_1 - 1) \tag{4}$$

and

$$c_2 = (N_2 - 1)(k_2 - 1). \tag{5}$$

Alsawalmeh & Feldt (1992) further showed that the distribution of the ratio of two such variables is approximately that of a central F with d_1 and d_2 *df*. These *df* were determined in such a way that F_{d_1,d_2} has the same expected value and variance as the test statistic T when H_0 is true. This laid the foundation for using T as the test statistic for $H_0: \rho_1 = \rho_2$. If T is too large or too small to be accepted as a randomly drawn central F with d_1 and d_2 *df*, then the null hypothesis is false.

An Approximate Test of $H_0: \rho_1 = \rho_2$ With Related Samples

When the same sample of people and possibly the same raters are used to estimate ρ_1 and ρ_2, the above theory breaks down. An approximate test that may be used to test the hypothesis $\rho_1 = \rho_2$ with related samples is developed below. It is, in effect, a test of the equality of two extrapolated α coefficients. The extrapolations are required to adjust for the unequal length of the instruments that were used in the reliability study. In the case of ratings, unequal length takes the form of unequal numbers of raters. However, the statistical test is applicable to a wider range of reliability studies. It may be used whenever the reliabilities of two competing measurement procedures, arbitrarily designated 1 and 2, are to be compared and the data are obtained on the same sample of people. It also may be used when the instruments under study are found to require unequal testing time but the researcher cannot equalize testing time experimentally.

For person i, let $X_{i1}, X_{i2}, \ldots, X_{ik_1}$, represent k_1 observable continuous assessments of a particular trait, as derived from k_1 raters or, more generally, using Procedure 1. Analogously, let $Y_{i1}, Y_{i2}, \ldots, Y_{ik_2}$, denote k_2 continuous scores on person i of the same trait arising from k_2 raters or from Procedure 2. Measures X_{ih} and Y_{ih} are correlated, over the population of persons, to the extent permitted by their errors of measurement. However, the units of measurement of X_h and Y_h are not directly comparable because the X and Y distributions have different means and variances. Person i is selected randomly from a large population of persons. The k_1 measures of X are presumed to be selected randomly from the population of scores for person i, and the k_2 measures of Y are analogously randomly selected for this person. Each score is assumed to conform to the constraints of the two-factor random ANOVA model. In addition, the measurement errors within and across procedures are assumed to be independent. The equality of the intraclass reliabilities for the two measurement procedures is to be tested.

Let $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ denote the unbiased estimates of the variances over persons of total scores X and Y resulting from Procedures 1 and 2. Let $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$ represent the variances of measurement errors for single scores within persons for these procedures. The following expectations, symbolized by E, hold (Feldt, 1980; Lindquist, 1953, p. 360):

$$E\left({}_1MS_p\right) = E\left(\hat{\sigma}_x^2/k_1\right) = \sigma_x^2/k_1 = k_1\left(\sigma_{r_1}^2\right) + \sigma_{e_1}^2, \tag{6}$$

$$E\left({}_2MS_p\right) = E\left(\hat{\sigma}_y^2/k_2\right) = \sigma_y^2/k_2 = k_2\left(\sigma_{r_2}^2\right) + \sigma_{e_2}^2, \tag{7}$$

$$E\left({}_1MS_{m\times p}\right) = \sigma_{e_1}^2, \tag{8}$$

and

$$E\left({}_2MS_{m\times p}\right) = \sigma_{e_2}^2. \tag{9}$$

In these expectations ${}_jMS_p$ is the mean square for persons and ${}_jMS_{m\times p}$ is the mean square for the measures × persons interaction derived from a persons × measures ANOVA with one observation per cell for measurement procedure j ($j = 1, 2$). Because the measurement procedures include k_1 and k_2 measures, respectively, and because the errors are independent, the error variances for the total scores on X and Y are $(k_1)\sigma_{e_1}^2$ and $(k_2)\sigma_{e_2}^2$.

Now consider the test statistic used by Alsawalmeh & Feldt (1992) for independent groups:

$$T = \frac{1-\hat{\rho}_1}{1-\hat{\rho}_2} = \left[\frac{(k_1)({}_1\mathrm{MS}_{\mathrm{m\times p}})}{(k_2)({}_2\mathrm{MS}_{\mathrm{m\times p}})}\right]\left[\frac{{}_2\mathrm{MS}_{\mathrm{p}}+(k_2-1)({}_2\mathrm{MS}_{\mathrm{m\times p}})}{{}_1\mathrm{MS}_{\mathrm{p}}+(k_1-1)({}_1\mathrm{MS}_{\mathrm{m\times p}})}\right]. \tag{10}$$

Given the expected values of ${}_j\mathrm{MS}_{\mathrm{p}}$ and ${}_j\mathrm{MS}_{\mathrm{m\times p}}$ defined above, the expected value of ${}_j\mathrm{MS}_{\mathrm{p}}+(k_j-1)({}_j\mathrm{MS}_{\mathrm{m\times p}})$ is:

$$k_j\sigma_{\tau_j}^2+\sigma_{e_j}^2+(k_j-1)\sigma_{e_j}^2 = k_j(\sigma_{\tau_j}^2+\sigma_{e_j}^2). \tag{11}$$

The quantity $\sigma_{\tau_j}^2+\sigma_{e_j}^2$ equals the variance of a single observed score under procedure j, which is symbolized by σ_j^2. Thus, the expected value of the linear combination of mean squares equals $k_j\sigma_j^2$. Dividing both the numerator and denominator of each ratio of Equation 10 by their expected values and multiplying by these expected values to preserve the equality gives

$$T = \frac{\left[\frac{(k_1)({}_1\mathrm{MS}_{\mathrm{m\times p}})}{k_1\sigma_{e_1}^2}\right]\left[\frac{{}_2\mathrm{MS}_{\mathrm{p}}(k_2-1)({}_2\mathrm{MS}_{\mathrm{m\times p}})}{k_2\sigma_2^2}\right](k_1\sigma_{e_1}^2)(k_2\sigma_2^2)}{\left[\frac{(k_2)({}_2\mathrm{MS}_{\mathrm{m\times p}})}{k_2\sigma_{e_2}^2}\right]\left[\frac{{}_1\mathrm{MS}_{\mathrm{p}}(k_1-1)({}_1\mathrm{MS}_{\mathrm{m\times p}})}{k_1\sigma_1^2}\right](k_2\sigma_{e_2}^2)(k_1\sigma_1^2)}. \tag{12}$$

The intraclass reliability for procedure j is

$$\rho_j = (\sigma_j^2-\sigma_{e_j}^2)/\sigma_j^2. \tag{13}$$

The distribution of ${}_j\mathrm{MS}_{\mathrm{p}}+(k_j-1)({}_j\mathrm{MS}_{\mathrm{m\times p}})$ is approximated by the synthetic mean square approach of Satterthwaite (1941). Under this approach, the distribution of a linear combination of mean squares is approximated by a single weighted χ^2 variable divided by its *df*. The *df* and the weight are determined so that the weighted χ^2 variable has the same mean and variance as the linear combination of mean squares. In the present application,

$${}_j\mathrm{MS}_{\mathrm{p}}+(k_j-1)({}_j\mathrm{MS}_{\mathrm{m\times p}}) \sim k_j(\sigma_j^2)(\chi_\upsilon^2)/\upsilon. \tag{14}$$

This leads to the following approximation of the T distribution:

$$T = \frac{1-\hat{\rho}_1}{1-\hat{\rho}_2} \sim \left[\frac{{}_1\mathrm{MS}_{\mathrm{m\times p}}/\sigma_{e_1}^2}{{}_2\mathrm{MS}_{\mathrm{m\times p}}/\sigma_{e_2}^2}\right]\left[\frac{\chi_{\upsilon_2}^2/\upsilon_2}{\chi_{\upsilon_1}^2/\upsilon_1}\right]\left[\frac{1-\rho_1}{1-\rho_2}\right]. \tag{15}$$

The effective *df* υ_j is approximately equal to

$$\upsilon_j \approx (N-1)k_j/\left[1+(k_j-1)\hat{\rho}_j^2\right]. \tag{16}$$

The first factor on the right side of Equation 15 may be recognized as the ratio of two independent χ^2 variables divided by their *df*. By the assumed independence of errors, this ratio is distributed as a central F with $c_1 = (N-1)(k_1-1)$ and $c_2 = (N-1)(k_2-1)$ *df*. With even moderate numbers of measures (raters) and a reasonably large sample size, these *df* will be quite large (1,000 or more). For all practical purposes, this F distribution may be considered to be almost totally concentrated at the point $F = 1.0$ (Hogg & Craig, 1978, pp. 196–198). Hence, this factor has negligible influence on the distribution of T.

Thus,

$$T = \frac{1-\hat{\rho}_1}{1-\hat{\rho}_2} \simeq \frac{1-\rho_1}{1-\rho_2} F^*_{\upsilon_2,\upsilon_1} . \tag{17}$$

In this expression, F^* is the ratio of two related variance estimates, each with an expected value of 1.0. One approximation to the distribution of F^* draws on the Bose (1935) demonstration that the general form of the distribution is that of a central F with modified *df*. This result is the starting point for a better and more accurate approximation to the distribution of T—one that is not limited to $(k_j)(N) > 1{,}000$. The new approximation is represented by that central F for which, under H_0, $E(F) = E(T)$ and $Var(F) = Var(T)$. Under H_0, T is distributed as the ratio $F^*_{c_1,\upsilon_1}/F^*_{c_2,\upsilon_2}$. This ratio, denoted F_1/F_2 for simplicity, involves related F^* variables. The covariance of these two variables, $\mathrm{Cov}(F_1, F_2)$, can be approximated using the Δ method described by Kendall & Stuart (1977, pp. 246–262). It is equivalent to the covariance between the two synthetic mean squares involved in Equation 12. That is,

$$\mathrm{Cov}(F_1, F_2) = \mathrm{Cov}\left[\frac{{}_1\mathrm{MS}_p + (k_1 - 1){}_1\mathrm{MS}_{m\times p}}{k_1\sigma_1^2}, \frac{{}_2\mathrm{MS}_p + (k_2 - 1){}_2\mathrm{MS}_{m\times p}}{k_2\sigma_2^2}\right]. \tag{18}$$

The covariance of one sum with another equals the sum of the covariances of all pairs of separate terms. Consistent with the assumption of independence of measurement errors, the independence of ${}_1\mathrm{MS}_p$ from ${}_1\mathrm{MS}_{m\times p}$, ${}_2\mathrm{MS}_p$ from ${}_1\mathrm{MS}_{m\times p}$, and ${}_1\mathrm{MS}_{m\times p}$ from ${}_2\mathrm{MS}_{m\times p}$ is assumed. Therefore, $\mathrm{Cov}(F_1, F_2)$ can be evaluated by approximating $\mathrm{Cov}({}_1\mathrm{MS}_p, {}_2\mathrm{MS}_p)$ or

$$\mathrm{Cov}(F_1, F_2) = \frac{1}{k_1 k_2 \sigma_1^2 \sigma_2^2} \mathrm{Cov}({}_1\mathrm{MS}_p, {}_2\mathrm{MS}_p) . \tag{19}$$

Because

$${}_1\mathrm{MS}_p = \hat{\sigma}_X^2/k_1 \tag{20}$$

and

$${}_2\mathrm{MS}_s = \hat{\sigma}_Y^2/k_2 , \tag{21}$$

where $X = X_1 + X_2 + \ldots + X_{k_1}$ and $Y = Y_1 + Y_2 + \ldots + Y_{k_2}$, it is possible to approximate $\mathrm{Cov}({}_1\mathrm{MS}_p, {}_2\mathrm{MS}_p)$ from a relationship cited by Kendall & Stuart (1977, pp. 331–332):

$$\mathrm{Cov}(\hat{\sigma}_X^2, \hat{\sigma}_Y^2) = 2\sigma_{XY}^2/(N-1) . \tag{22}$$

Accordingly,

$$\mathrm{Cov}(F_1, F_2) = \frac{1}{k_1 k_2 \sigma_1^2 \sigma_2^2} \mathrm{Cov}\left(\hat{\sigma}_X^2/k_1, \hat{\sigma}_Y^2/k_2\right) = \frac{1}{k_1^2 k_2^2 \sigma_1^2 \sigma_2^2} \mathrm{Cov}\left(\hat{\sigma}_X^2, \hat{\sigma}_Y^2\right) = \left(\frac{1}{k_1^2 k_2^2 \sigma_1^2 \sigma_2^2}\right)\left(\frac{2\sigma_{XY}^2}{N-1}\right). \tag{23}$$

However,

$$\sigma_{XY}^2 = \sigma^2_{(X_1+\ldots+X_{k_1})(Y_1+\ldots+Y_{k_2})} = \left[k_1 k_2 \sigma_{12}\right]^2 = k_1^2 k_2^2 \sigma_{12}^2 , \tag{24}$$

where σ_{12} is the covariance between one measure obtained under Procedure 1 and another measure obtained under Procedure 2. Therefore, the covariance is:

$$\mathrm{Cov}(F_1, F_2) = \left(\frac{2}{N-1}\right)\rho_{12}^2 , \tag{25}$$

where ρ_{12} is the correlation between two individual measures, X_h and Y_h, from Procedures 1 and 2.

Now the mean and variance of T must be obtained. The exact mean and variance of T are unknown under H_0, but estimates of these moments are sufficient to identify the desired F distribution. The estimates of these moments also may be obtained using the Δ method. Using this method, the mean (M) and sampling variance (Var) of T are given by the following two formulas:

$$\mathrm{M} = \frac{\mathrm{E}(F_1)}{\mathrm{E}(F_2)} + \frac{\mathrm{E}(F_1)\,\mathrm{Var}(F_2)}{\mathrm{E}^3(F_2)} - \frac{\mathrm{Cov}(F_1, F_2)}{\mathrm{E}^2(F_2)} \tag{26}$$

and

$$\mathrm{Var} = \left[\frac{\mathrm{E}(F_1)}{\mathrm{E}(F_2)}\right]^2 \left[\frac{\mathrm{Var}(F_1)}{\mathrm{E}^2(F_1)} + \frac{\mathrm{Var}(F_2)}{\mathrm{E}(F_2)} - \frac{2\mathrm{Cov}(F_1, F_2)}{\mathrm{E}(F_1)\mathrm{E}(F_2)}\right], \tag{27}$$

where

$$\mathrm{E}(F_1) = \frac{\upsilon_1}{\upsilon_1 - 2} - \frac{2(1-\rho_1)}{k_1(N-1)}, \tag{28}$$

$$\mathrm{E}(F_2) = \frac{\upsilon_2}{\upsilon_2 - 2} - \frac{2(1-\rho_2)}{k_2(N-1)}, \tag{29}$$

$$\mathrm{Var}(F_1) = \frac{2\upsilon_1^2(c_1 + \upsilon_1 - 2)}{c_1(\upsilon_1 - 2)^2(\upsilon_1 - 4)} - \frac{4(1-\rho_1)}{k_1(N-1)}, \tag{30}$$

and

$$\mathrm{Var}(F_2) = \frac{2\upsilon_2^2(c_2 + \upsilon_2 - 2)}{c_2(\upsilon_2 - 2)^2(\upsilon_2 - 4)} - \frac{4(1-\rho_2)}{k_2(N-1)} \tag{31}$$

(Alsawalmeh & Feldt, 1992, p. 199). Substitution of these quantities and the covariance quantity indicated in Equation 25 into Equations 26 and 27 yields estimates of the mean and variance of T.

Given the values of M and Var, the *df* for the numerator (d_1) and denominator (d_2) of the desired F distribution are

$$d_1 = \frac{2d_2^3 - 4d_2^2}{\mathrm{Var}(d_2 - 2)^2(d_2 - 4) - 2d_2^2} \tag{32}$$

and

$$d_2 = \frac{2\mathrm{M}}{\mathrm{M} - 1}. \tag{33}$$

Thus, under $H_0: \rho_1 = \rho_2$,

$$T = \frac{1 - \hat{\rho}_1}{1 - \hat{\rho}_2} \doteq F_{d_1, d_2}. \tag{34}$$

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*, the conclusion at the designated significance level is that $\rho_1 \neq \rho_2$.

In practice, the parameters ρ_1, ρ_2, and ρ_{12} are unknown and are estimated from sample statistics. This adds to the approximate character of the distribution of T and makes empirical validation of the statistical test

necessary.

A numerical illustration of the steps involved in identifying the *df* for the appropriate *F* distribution is as follows. Suppose that $N = 101$, $\hat{\rho}_1 = .3$, $k_1 = 5$, $\hat{\rho}_2 = .4$, $k_2 = 7$, and $\rho_{12} = .2$. For these data, the following are obtained:

1. $T = (1 - .30)/(1 - .40) = 1.17$, $c_1 = 400$, and $c_2 = 600$ (Equations 3–5).
2. From Equation 16, $\upsilon_1 = 368$ and $\upsilon_2 = 357$.
3. From Equations 25 and 28–31, $\text{Cov}(F_1, F_2) = .0008$, $\text{E}(F_1) = 1.0027$, $\text{E}(F_2) = 1.00392$, $\text{Var}(F_1) = .00504$, and $\text{Var}(F_2) = .00569$.
4. From Equations 26 and 27, M = 1.00363 and Var = .00905.
5. Finally, from Equations 32 and 33, $d_2 = 553$ and $d_1 = 376$. $P(F_{376,553} > 1.17) = .0471$ and H_0 is rejected at the 5% level.

Computer Simulation of the Approximate Test

The technique developed by Odell & Feiveson (1966) was used to generate 4,000 joint sample covariance matrices for two hypothetical measurement procedures, one with k_1 measurements and the other with k_2 measurements. Three levels of intraclass reliability were simulated: .5, .4, and .3. Within each of these levels, monte carlo experiments were generated for two values of N (100 or 200), two combinations of k_1 (5 or 7) and k_2 (5 or 10), and two levels of ρ_{12} (.2 and .4 with $\rho = .5$; .2 and .3 with $\rho = .4$; .1 and .2 with $\rho = .3$). Thus, under each level of intraclass reliability there were eight combinations of N, k_j, and ρ_{12}.

For each configuration of parameter values, each of the 4,000 replications began with the generation of a $(k_1 + k_2) \times (k_1 + k_2)$ joint sample covariance matrix. This matrix was used to compute $\hat{\rho}_1$ and $\hat{\rho}_2$, and the test statistic T was obtained. The empirical distribution of T then was compared to the approximate theoretical model (F_{d_1,d_2}) described above. The proportions of rejections of H_0 that occurred at the 10%, 5%, and 1% levels were obtained.

The results for the three levels of intraclass reliability are summarized in Table 1. Inspection of these data indicates that the proposed test statistic T provides accurate control of Type I error rates at the three significance levels. The average of the 24 empirical error rates was 10.2% at the 10% level, 5.2% at the 5% level, and 1.1% at the 1% level.

Additional analyses were carried out to investigate how the test would perform in situations with small numbers of simulated measurements. In these analyses, one level of intraclass reliability was simulated (.4) with $k_1 = 2$ and $k_2 = 3$. These values were paired with four values of N (50, 100, 200, 400) and two values of ρ_{12} (.2, .3). The results of these analyses suggest that the test statistic T provides accurate control of Type I error rates

Table 1
Empirical Estimates of Type I Error Rates Under a True Null Hypothesis With $\rho_1 = \rho_2 = .30$, .40, or .50 for Nominal Significance Levels of .10, .05, and .01 (the Standard Errors Associated With These Nominal Significance Levels Were .0047, .0034, and .0016, Respectively)

			ρ = .50				ρ = .40				ρ = .30			
N	k_1	k_2	ρ_{12}	.10	.05	.01	ρ_{12}	.10	.05	.01	ρ_{12}	.10	.05	.01
100	5	7	.2	.111	.056	.012	.2	.093	.048	.009	.1	.098	.045	.011
100	5	7	.4	.100	.049	.010	.3	.102	.049	.011	.2	.122	.062	.013
100	5	10	.2	.106	.057	.012	.2	.099	.053	.014	.1	.109	.054	.012
100	5	10	.4	.096	.049	.007	.3	.101	.054	.013	.2	.102	.054	.012
200	5	7	.2	.095	.050	.009	.2	.101	.054	.013	.1	.103	.053	.010
200	5	7	.4	.098	.050	.010	.3	.111	.052	.010	.2	.097	.047	.010
200	5	10	.2	.107	.050	.011	.2	.098	.046	.011	.1	.100	.050	.010
200	5	10	.4	.106	.055	.012	.3	.101	.055	.014	.2	.097	.051	.011

even with kN as small as 100. The averages of the eight empirical estimates of Type I error rates were 10.3% at the 10% level, 5.1% at the 5% level, and 1.2% at the 1% level. Thus, the T statistic appeared to perform satisfactorily in testing the equality of two dependent intraclass reliability coefficients computed from as few as two or three normally distributed scores.

In the theoretical development of the proposed test it was necessary to assume that errors of measurement were independent across procedures. This assumption seems more likely to be satisfied if independent samples of raters and independent samples of behavior are used under the several procedures. If rater h reacts negatively to some aspect of the performances of person i, this reaction will probably be reflected under both measurement procedures. Similarly, if person i performed poorly on occasion h, both X_{ih} and Y_{ih} would fall below the person's respective true scores for the two procedures. The net effect of such carryover from one procedure to the other would be to induce a correlation between errors, violating the assumed independence.

The assumption of normality also cannot be ignored in the proposed test. Several investigators have observed that for test scores with reliabilities above .75, population distributions tend to deviate from normality in the direction of platykurtosis (Brandenburg & Forsyth, 1974; Cook, 1959; Lord, 1955). Feldt (1969) found that platykurtosis tended to lower the probability of Type I error in the test of the hypothesis that $\alpha_1 = \alpha_2$. It seems likely that this same tendency will hold for the present test of $\rho_1 = \rho_2$, but this expectation must be confirmed through further studies based on actual measurement data.

References

Alsawalmeh, Y. M., & Feldt, L. S. (1992). A test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement, 16,* 195–205.

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

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

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

Brandenburg, D. C., & Forsyth, R. A. (1974). Approximating standardized achievement test norms with a theoretical model. *Educational and Psychological Measurement, 34,* 3–9.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City IA: American College Testing Program.

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

Cook, D. L. (1959). A replication of Lord's study of skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement, 19,* 81–87.

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

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

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

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

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

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

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

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

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.