Psychological Methods
1996, Vol. 1, No. 1, 30-46

1082-989X/96/$3.00

Forming Inferences About Some Intraclass Correlation Coefficients

Kenneth O. McGraw
University of Mississippi

S. P. Wong
University of Memphis

Although intraclass correlation coefficients (ICCs) are commonly used in behavioral measurement, psychometrics, and behavioral genetics, procedures available for forming inferences about ICCs are not widely known. Following a review of the distinction between various forms of the ICC, this article presents procedures available for calculating confidence intervals and conducting tests on ICCs developed using data from one-way and two-way random and mixed-effect analysis of variance models.

To measure the bivariate relation of variables representing different measurement classes, one must use an *inter*class correlation coefficient, of which there is but one in common use, the Pearson *r*. Thus the Pearson *r* is used for measuring the relation of IQ points (a class of measurement representing aptitude) to grade point averages (a class of measurement representing achievement) or the relation of measurements in the length class (e.g., inches) to measurements in the weight class (e.g., pounds). Such measurements share neither their metric nor variance. But when one is interested in the relationship among variables of a common class, which means variables that share both their metric and variance, *intra*class correlation coefficients (ICCs) are alternative statistics for measuring homogeneity, not only for pairs of measurements but for larger sets of measurements as well.[1,2]

The most fundamental interpretation of an ICC is that it is a measure of the proportion of a variance (variously defined) that is attributable to objects of measurement, often called *targets* (e.g., Shrout & Fleiss, 1979). The objects might be gymnastics contestants, litters, twin pairs, or students, and the corresponding measurements might be judges' ratings, IQs of the twins, weights of the littermates, or test scores of the students. Common examples of ICCs in the literature are twin correlations, Cronbach's alpha, heritability coefficients, Kish's rate of homogeneity (Kish, 1965), and measures of reliability that arise from either classical test theory or generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Kenneth O. McGraw, Department of Psychology, University of Mississippi; S. P. Wong, Mathematical Sciences, University of Memphis.

We acknowledge the uncommonly valuable contributions made during the development of this article by the reviewers and by *Psychological Bulletin* Associate Editor Scott E. Maxwell, who directed us to relevant literature, clarified key conceptual issues, and identified errors.

Correspondence concerning this article should be addressed to Kenneth O. McGraw, Department of Psychology, University of Mississippi, University, Mississippi 38677. Electronic mail may be sent to pymcgraw @sunset.backbone.olemiss.edu.

[1] Interpretation of the meaning of the term *class* in this contrast of interclass and intraclass correlations follows Fisher (e.g., 1938, pp. 217–218). Frequently, however, one sees interpretations (e.g., Haggard, 1958, passim; Harris, 1913, passim) in which *class* is used to refer to the test takers, persons, families, or other entities that serve as objects of measurement in a correlational analysis. Fisher's usage is preferable in that it emphasizes that the correlation is among measures constituting a class because they have a common metric and variance, although even Fisher refers ambiguously on occasion to objects of measurement as "classes" (e.g., 1938, p. 227). Genuine confusion exists, therefore, on this semantic issue.

[2] Fagot (1993) has proposed a family of coefficients for measuring the association of variables that offers yet additional alternatives to inter- and intraclass correlations.

Despite the widespread use of ICCs in psychology, textbook coverage of them is so limited that procedures for forming inferences about ICCs are seldom mentioned; when they are, the only procedures given are those based on Fisher's normalizing transformation of ρ, the z' transformation (e.g., Haggard, 1958). These procedures are the best known, but they are not the best available. The primary purpose of this article is to provide a review of options to Fisher's approach for determining confidence intervals and conducting significance tests. The procedures are limited to ICCs developed on data with just one or two sources of systematic variance (one-way and two-way models), because these are the only ICCs for which confidence intervals and test statistics have been derived.

The first step in using any of the procedures described here is to specify an additive variance model appropriate to one's sample data. This is important because the modern method for calculating ICCs, originated by Harris (1913), uses mean squares from an analysis of variance, so one must specify a model for the sample data in order to know which analysis to perform. The five models to be considered here are given as Cases 1, 2, 2A, 3, and 3A in Table 1. The case labels 1, 2, and 3 are taken from Shrout and Fleiss (1979), who did not formally consider Cases 2A and 3A. Readers who are not yet familiar with intraclass correlations and the work of Shrout and Fleiss may safely ignore Table 1 for the moment and proceed with the narrative discussion.

Selecting the Appropriate Model

Common to any model for which an ICC is defined will be a factor that represents randomly selected objects of measurement (test takers, subjects, litters, twinships, etc.) as a source of variance. From a design point of view, this means that data sets used in calculating ICCs require multiple measurements on these objects. Because the objects are randomly selected, they constitute a random factor in the design. A convenient arrangement for the measurements is the one given by Bartko (1976) and reproduced in Table 2, where i is used as the subscript for the randomly chosen objects of measurement (which vary in number from 1 to n) and j the subscript for multiple observations (which vary in number from 1 to k).

One-Way Random Effects Model

For the simplest case, assume that the random row variable in Table 2 represents the only systematic source of variance. This would be the case when data are collected in such a way that their ordering on j is irrelevant. In common parlance, this represents a nested design because unordered observations are nested within objects. As an example, consider behavioral genetics data used in assessing familial resemblance. When looking at sibships of size k, one has k measures (scores on twins, say, where $k = 2$), but there is no way to assign these scores to measurement categories; thus their assignment to j or j' is random. In such cases as this, the one-way analysis of variance model given as Case 1 in Table 1 can be used to represent the data. The same one-way model could be used for data reflecting the measurement of persons in which each observation x_{ij} was made under unique measurement conditions, a situation that creates what are called "unmatched" data in generalizability theory (Cronbach et al., 1972). In each of these examples, r_i represents the random effects of the row variable (twinships in the one case and persons in the other), and w_{ij} represents random residual effects associated with the idiosyncracies of the measurement conditions, their interaction with the row variable, and measurement error. Assumptions concerning these effects are given in Table 1 along with other assumptions relevant to the models.

As shown in Table 3, an analysis of variance on data conforming to the one-way model yields two mean squares, one for object of measurement (the row variable in the Table 2 data matrix) and one for residual sources of variance. By tradition the latter of these is labeled MS_W, for mean square within. We will use the label MS_R (mean square rows) for the former. The expectations for these mean squares are given in Table 3.

Two-Way Models

For cases in which the k observations per object of measurement differ in some systematic way, a two-way model can be used to represent the data. The reason is that there is a systematic source of variance associated with columns as well as with rows. For example, if the columns represent items on a mathematics test, the items may differ in difficulty, thus creating a separable source of vari-

Table 1
Analysis of Variance Models Used in Developing Intraclass Correlation Coefficient Definitions

Case label	Model	Assumptions
Case 1: One-way random effects	$x_{ij} = \mu + r_i + w_{ij}$ where $i = 1, \ldots, n$ and $j = 1, \ldots, k$.	μ (the population mean for all observations) is constant; r_i (the row effects) are random, independent, and normally distributed with mean 0 and variance σ_r^2; and w_{ij} (residual effects) are random, independent, and normally distributed with mean 0 and variance σ_w^2. Moreover, the effects r_i and w_{ij} are pairwise independent.
Case 2: Two-way random effects, with interaction	$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$ where $i = 1, \ldots, n$ and $j = 1, \ldots, k$.	μ and r_i are as before; c_j (the column effects) are random, independent, and normally distributed with mean 0 and variance σ_c^2; rc_{ij} (the interaction effects) are random, independent, and normally distributed with mean 0 and σ_{rc}^2; and e_{ij} (residual effects) are random, independent, and normally distributed with mean 0 and variance σ_e^2. Moreover, all the effects are pairwise independent.
Case 2A: Two-way random effects, interaction absent	$x_{ij} = \mu + r_i + c_j + e_{ij}$ where $i = 1, \ldots, n$ and $j = 1, \ldots, k$.	Same as for Case 2 except that there is no interaction effect.
Case 3: Two-way mixed effect model, with interaction	$x_{ij} = \mu + r_i + c_j + rc_{ij} + e_{ij}$ where $i = 1, \ldots, n$ and $j = 1, \ldots, k$.	Same as for Case 2 except that c_j are fixed so that $\Sigma c_j = 0$, $\sum_{j=1}^{k} rc_{ij} = 0$, and the parameter corresponding to σ_c^2 in Case 2 is $\theta_c^2 = \Sigma c_j^2/(k-1)$.
Case 3A: Two-way mixed model, interaction absent	$x_{ij} = \mu + r_i + c_j + e_{ij}$ where $i = 1, \ldots, n$ and $j = 1, \ldots, k$.	Same as for Case 3 except that there is no interaction effect.

ance. The same would be true if the columns represented different judges who might differ in their anchor points. These situations specify a two-way model, and from a design point of view, a randomized blocks design in which the column variable is crossed with the row (blocks) variable.

Whereas there was just one one-way model given in Table 1, there are four given for two-way models. These four models differ in whether c_j represents random or fixed effects and in whether the model includes an interaction component. For Cases 2 and 2A, the effects c_j are random, but for Cases 3 and 3A they are fixed. The *A* extension to case numbers indicates that the interaction component is absent from the model. These distinc-

Table 2
A Convenient Data Matrix and Notational System for Data Used in Calculating Intraclass Correlation Coefficients

Object of measurement	Measurement			
	1	2	... j	... k
1	X_{11}	X_{12}	$\cdots X_{1j}$	$\cdots X_{1k}$
2	X_{21}	X_{22}	$\cdots X_{2j}$	$\cdots X_{2k}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
i	X_{i1}	X_{i2}	$\cdots X_{ij}$	$\cdots X_{1ik}$
.	.	.	.	.
.	.	.	.	.
n	X_{n1}	X_{n2}	$\cdots X_{nj}$	$\cdots X_{nk}$

tions among models are important because they have implications for the ICCs that can be calculated and for their interpretations, as will be explained below.

An analysis of variance on randomized block data yields three mean squares: a mean square for rows (MS_R), a mean square for columns (MS_C), and a residual mean square traditionally referred to as mean square error (MS_E). MS_R in a two-way analysis is the same as that for a one-way analysis but MS_E is smaller than MS_W by the amount of variance attributable to the k-level columns factor (e.g., items or judges). Note that because there is just one observation per cell of a randomized blocks design, the expected mean square for MS_E under Cases 2 and 3 (see Table 3) estimates the combined interaction and error variance. Interaction is absent from the models given as Cases 2A and 3A, thus eliminating the problem of confounded interaction and error variance.

Defining and Calculating ICCs for One-Way and Two-Way Models

ICCs that can be defined for one-way and two-way models are given in Tables 4 and 5. Also given are calculation formulas, designations, and interpretations for each type of ICC. Using Tables 4 and 5 requires making distinctions between (1) ICCs that are for single measurements and those that are for average measures, (2) ICCs that measure the degree of relationship between measurements (whether single measurements or averages) in terms of consistency or of absolute agreement, and (3) ICCs that reflect the degree of relationship between observations made under fixed levels of the column factor or under randomly chosen levels of the column factor.

ICCs for Single Measurements Versus Average Measurements

Tables 4 and 5 differ in that Table 4 gives ICCs that apply to single measurements x_{ij} (e.g., the ratings of judges, individual item scores, or the body weights of individuals), whereas the Table 5 ICCs apply to average measurements (e.g., the average rating for k judges, the average score for a k-item test, or the average weight of k littermates). We refer to the Table 4 ICCs for single measurements as Type 1 ICCs and the Table 5 ICCs for average measurements as Type k ICCs. Even though the interpretation of the two types of coefficients differs, mathematically the Type 1 coefficients are simply a special case of Type k coefficients.

ICCs for Consistency Versus Agreement

Whereas there is only one ICC of each type for one-way data, there are two ICCs of each type for two-way data, as is discussed in detail in a number of excellent sources (e.g., Berk, 1979; Crocker & Algina, 1986; Shavelson & Webb, 1991; Suen & Ary, 1989). The first type of ICC measures correlation using a consistency definition; the second, an absolute agreement definition. Understanding the conceptual difference between them begins by noting their formal distinction, which is in the definition of the ICC denominator. For consistency measures, column variance is excluded from denominator variance, and for absolute agreement measures, it is not.

Column variance is excluded from the denominators of consistency measures because it is deemed to be an irrelevant source of variance. Consider, for example, that measurements are needed to determine the relative standing of job applicants. In this context it does not matter that Judge 1 assigns relatively high scores and Judge 2 low scores. The ratings of the two judges agree to the extent that an additive transformation serves to equate them (e.g., subtracting the mean rating for each judge from their individual ratings). This

Table 3
Mean Square Expectations for Analysis of Variance Models Given in Table 1

Model and source of variation	*df*	*MS*	*EMS*
Case 1: One-way random effects model			
Between rows	$n - 1$	MS_R	$k\sigma_r^2 + \sigma_w^2$
Within rows	$n(k - 1)$	MS_W	σ_w^2
Case 2: Two-way random model with interaction			
Between rows	$n - 1$	MS_R	$k\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2$
Within rows	$n(k - 1)$	MS_W	$\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2$
Between columns	$k - 1$	MS_C	$n\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	MS_E	$\sigma_{rc}^2 + \sigma_e^2$
Case 2A: Two-way random model, interaction absent			
Between rows	$n - 1$	MS_R	$k\sigma_r^2 + \sigma_e^2$
Within rows	$n(k - 1)$	MS_W	$\sigma_c^2 + \sigma_e^2$
Between columns	$k - 1$	MS_C	$n\sigma_c^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	MS_E	σ_e^2
Case 3: Two-way mixed model with interaction			
Between rows	$n - 1$	MS_R	$k\sigma_r^2 + \sigma_e^2$
Within rows	$n(k - 1)$	MS_W	$\theta_c^2 + \frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2$
Between columns	$k - 1$	MS_C	$n\theta_c^2 + \frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	MS_E	$\frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2$
Case 3A: Two-way mixed model, interaction absent			
Between rows	$n - 1$	MS_R	$k\sigma_r^2 + \sigma_e^2$
Within rows	$n(k - 1)$	MS_W	$\theta_c^2 + \sigma_e^2$
Between columns	$k - 1$	MS_C	$n\theta_c^2 + \sigma_e^2$
Error	$(n - 1)(k - 1)$	MS_E	σ_e^2

Note. E(*MS*) = expected mean squares; MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_C = mean square for columns; MS_E = mean square error.

definition of agreement, which is useful in contexts in which comparative judgments are made about the objects of measurement, contrasts with an absolute agreement definition of correlation, which takes total score variance as its denominator. In this case, when measurements differ in absolute value, regardless of the reason, they are viewed as disagreements. Thus paired scores (2,4), (4,6), and (6,8) are in perfect agreement using a consistency definition [ICC(C,1) = 1.00] but not an absolute agreement definition [ICC(A,1) = .67].

In conformity with the distinction above, the definitions for ICCs based on two-way models differ by virtue of the presence or absence of column variance in the denominator of the variance ratio, as shown in column 1 of Tables 4 and 5. Coefficients labeled as Type C, for consistency coefficients, do not include column variance, whereas those labeled as Type A, for absolute agreement coefficients, do. Thus, we refer to coefficients based on two-way models using the designations (C,1), (C,k), (A,1) and (A,k). For one-way models, there are no C-type coefficients because only absolute agreement is measurable in this context.

Contrasting ICC(C,1) and Pearson's r

The example given above in which paired scores (2,4), (4,6), and (6,8) yield a value of 1.00 for

Table 4
Single Score Intraclass Correlation Coefficients (ICCs) for One-Way and Two-Way Models

Definitions of ICCs ρ	Formulas for calculating $\hat{\rho}$	Designation	Interpretation of ICC
	Row effects random		
One-way model *Case* 1 *model* $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2}$	$\frac{MS_R - MS_W}{MS_R + (k-1)MS_W}$	ICC(1)	The degree of absolute agreement among measurements made on randomly selected objects. It estimates the correlation of any two measurements.
	Column and row effects random (two-way random effects model)		
Two-way models[a] *Case* 2 *model* $\frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)}$ *or Case* 2*A model* $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$	ICC(C,1)	The degree of consistency among measurements. Also known as norm-referenced reliability and as Winer's adjustment for anchor points (Winer, 1971). In generalizability theory, this ICC estimates the squared correlation of individual measurements and universe scores.
Case 2 *model* $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + (\sigma_{rc}^2 + \sigma_e^2)}$ *or Case* 2*A model* $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$	ICC(A,1)	The degree of absolute agreement among measurements. Also known as criterion-referenced reliability. Estimates the Type 1 ICC for one-way, unmatched data (Rajartnam, 1960).
	Column effects fixed, row effects random (two-way mixed effect model)		
Case 3 *model* $\frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)}$ *or Case* 3*A model* $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$	ICC(C,1)	The degree of consistency among measurements made under the fixed levels of the column factor. This ICC estimates the correlation of any two measurements, but when interaction is present, it underestimates reliability.
Case 3 *model* $\frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \theta_c^2 + (\sigma_{rc}^2 + \sigma_e^2)}$ *or Case* 3*A model* $\frac{\sigma_r^2}{\sigma_r^2 + \theta_c^2 + \sigma_e^2}$	$\frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$	ICC(A,1)	The absolute agreement of measurements made under the fixed levels of the column factor.

Note. MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_E = mean square error; MS_C = mean square for columns.

[a] In the event of data with a two-way classification for which the column variance is zero (i.e., $\sigma_c^2 = 0$ or $\theta_c^2 = 0$, depending on the model), a one-way model should be used. Thus even though test scores on k parallel tests can be classified by test and test taker, he column variance by definition is zero, which means that a one-way model applies.

Table 5
Average Score Intraclass Correlation Coefficients (ICCs) for One-Way and Two-Way Models

Definitions of ICCs ρ	Formulas for calculating $\hat{\rho}$	Designation	Interpretation of ICC
	One-way model: Row effects random		
Case 1 model $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2/k}$	$\frac{MS_R - MS_W}{MS_R}$	ICC(k)	The degree of absolute agreement for measurements that are averages of k independent measurements on randomly selected objects.
	Two-way models[a]: Column and row effects random (random effects model)		
Case 2 model $\frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k}$ *or Case 2A model* $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2/k}$	$\frac{MS_R - MS_E}{MS_R}$	ICC(C,k)	The degree of consistency for measurements that are averages of k independent measurements on randomly selected objects. Known as Cronbach's alpha in psychometrics. In generalizability theory, this ICC estimates the squared correlation of average scores and universe scores.
Case 2 model $\frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}$ *or Case 2A model* $\frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_e^2)/k}$	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$	ICC(A,k)	The degree of absolute agreement for measurements that are averages based on k independent measurements on randomly selected objects. Also estimates from two-way data the Type k ICC for one-way data (Rajaratnam, 1960).
	Column effects fixed and row effects random (mixed effects model)		
Case 3 model $\frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k}$	Not estimable		
Case 3A model $\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2/k}$	$\frac{MS_R - MS_E}{MS_R}$	ICC(C,k)	The degree of consistency for averages of k independent measures made under the fixed levels of the column factor.
Case 3 model $\frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\theta_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}$	Not estimable		
Case 3A model $\frac{\sigma_r^2}{\sigma_r^2 + (\theta_c^2 + \sigma_e^2)/k}$	$\frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}}$	ICC(A,k)	The degree of absolute agreement for measurements that are based on k independent measurements made under the fixed levels of the column factor.

Note. MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_E = mean square error; MS_C = mean square for columns.
[a] In the event of data with a two-way classification for which the column variance is zero (i.e., $\sigma_c^2 = 0$ or $\theta_c^2 = 0$, depending on the model), a one-way model should be used. Thus even though test scores on k parallel tests can be classified by test and test taker, the column variance by definition is zero, which means that a one-way model applies.

ICC(C,1) may cause some readers to wonder how ICC(C,1) compares to a Pearson r coefficient, because for the above set of paired scores Pearson's r would also be 1.00. Until recently, an important contrast between ICC(C,1) and a Pearson r was that the latter could be computed only for $k = 2$, but Fagot (1993) has introduced a correlation index (L, for linearity index) that equals the Pearson

r for $k = 2$ and equals the arithmetic mean of all possible pairwise rs for $k \geq 2$, so restrictions on k no longer apply for this type of relational measure. The distinction, therefore, is between what Fagot calls a linearity index and an additivity index. The Pearson r is a linearity index because it measures the degree to which one variable y can be equated to another variable x by a linear transformation ($y = ax + b$). ICC(C,1) on the other hand is an additivity index because—for the case $k = 2$—it measures the degree to which one variable y can be equated to another variable x by adding a constant ($y = x + b$).

A consequence of the linearity–additivity distinction is that differences in the sample variances for the variables x and y will attenuate ICC(C,1) relative to r. Recall that ICCs are constructed using models that assume equal variance (see Table 1). Variance differences between columns in the sample data, therefore, indicate lack of agreement among the observations. ICC(C,1) is appropriately sensitive to this source of disagreement. Thus, whereas a Pearson r—using its linear scale definition of agreement—judges paired scores (0,4), (5,5), and (10,6) to be in perfect agreement ($r = 1.00$), ICC(C,1) judges them to be in imperfect agreement [ICC(C,1) = .38]. A point to remember, therefore, when choosing between an index of linear agreement and ICC(C,1) is that ICC(C,1) is an appropriate measure of agreement only when there is a common population variance for all measurement conditions. Where this assumption is not met, it would be meaningless to calculate ICC(C,1) or any other ICC. This fact harks back to the original justification for the term *intraclass,* which is that measurements must be of a single class.

Column Variables Representing Fixed Versus Random Effects

A third matter of importance for distinguishing among the different ICCs in a two-way model concerns whether the column variable represents a random effect or a fixed effect. This distinction was introduced above as the basis for distinguishing Cases 2 and 2A from Cases 3 and 3A in Table 1. In technical terms, a factor is random when its levels are selected by random sampling from a larger set of equally usable levels; it is fixed when its levels are dictated by the research question. In practical terms, one knows that the levels of a variable are random when a change in the levels of the variable would have no effect on the question being asked. Jackson and Brashers (1994) call this the "replaceability test" for determining when a factor is random. Subjects constitute a random factor, for example, because the particular subjects selected for any study are always replaceable by others from the same population. In contrast, changing the levels of a variable with fixed effects substantially alters the research question. An example of a fixed effect variable is the biological relation variable in a study that has levels of mother and child. Changing these levels to uncle and nephew would imply a totally different research interest.

The importance of the random–fixed effects distinction is in its effect on the interpretation, but not calculation, of an ICC. Namely, when levels of the column factor are randomly sampled, one can generalize beyond one's data, but not when they are fixed. In either case, however, the value of the ICC is the same, though one should keep in mind that the population ICCs are defined differently in the two cases (see Tables 4 and 5).

Not only are ICC calculation formulas the same for random and mixed effect models, so too are the confidence intervals and test statistics, as shown in Tables 7 and 8 and as demonstrated in Appendix A. Case 3—the two-way model that includes a fixed column factor in conjunction with interaction variance—provides the only complication. In this case, ICCs (C,k) and (A,k) cannot to our knowledge be estimated, as indicated in Table 5.

ICCs Not Defined by Shrout and Fleiss (1979)

Readers familiar with Shrout and Fleiss's (1979) classic paper on intraclass correlation coefficients will note that ICC(A,1) for mixed effects models (Cases 3 and 3A) and ICCs (C,1) and (C,k) for random effects models (Cases 2 and 2A) were not among the ICCs that Shrout and Fleiss defined. These ICCs were omitted because they are not correlations in the strict sense of being ratios of covariance to total variance. Nonetheless, these ICCs are of considerable practical value for measuring degree of relationship. The practical value of ICC(C,1) and ICC(C,k) coefficients for random effects models is well documented in measurement

theory. Hartmann (1982), for example, suggested the random effects model ICC(C,1) as a measure of interobserver reliability, a suggestion seconded by Suen (1988), who nonetheless prefers to call the ICC a measure of intraobserver reliability. Type (C,k) coefficients for random effects models are very widely used as the generalizability theory analog to reliability coefficients in classical test theory (e.g., Shavelson & Webb, 1991, p. 92). They are equal to Cronbach's alpha (Cronbach, 1951), the most widely used measure in psychometrics for estimating the internal consistency of multi-item tests. To our knowledge, however, the (A,1) mixed model coefficient is novel. An example will illustrate its practical value.

Use of ICC(A,1) for Mixed Effects Designs

When adoption study data are analyzed, there are two effects of major interest. One is measured in mean differences; the other in correlations. The interest in mean differences is for determining whether adopted-away children have trait values that differ on average from those of their biological parents. The interest in correlations is for determining whether the degree of correlation for genetically paired individuals who do not live together is different from the correlation for genetically unrelated individuals who do live together.

A consistent finding from adoption research using IQ as a trait measure is that while adopted children have higher IQs on average than do their biological parents, their scores nonetheless correlate better with those of their biological parents than with those of their adoptive parents (Horn, Loehlin, & Willerman, 1979; Skodak & Skeels, 1949). These findings have been difficult to reconcile (Turkheimer, 1991) because on the basis of the difference in correlations, which are typically measured with Pearson *r*s, one is inclined to view the adoption data as evidence of the importance of genetic differences in creating individual differences in IQ. On the basis of the mean differences, one is inclined to see evidence of the role of environmental differences in creating individual differences in IQ. Historically, therefore, analysts have treated adoption data on IQ as a figure–ground illusion: They have focused first on one effect and then the other. Although both effects need to be highlighted, there is an additional need to reconcile the two findings. ICC(A,1) is ideal for the purpose.

Using ICC(A,1) to measure the correlation between parents and children (fixed effects in the model) serves to resolve the different foci of traditional analyses—absolute differences on the one hand and rank order similarities on the other—into a single measure, one that can be clearly contrasted with a linearity (e.g., *r*) or additivity (e.g., ICC(C,1)) index. When the mean differences between groups are small, there will be little difference between ICC(A,1) and ICC(C,1). As mean differences increase, however, ICC(A,1) will diminish in value relative to ICC(C,1), thereby emphasizing even for the most casual reader that agreement in an additive or linear sense must not be interpreted as agreement in an absolute sense. This is demonstrated with the data in Table 6.

For all three data sets in Table 6, ICC(C,1) equals the Pearson *r* coefficient because the mother and child variances in IQ are equal ($SD = 15$ in each). ICC(A,1), however, differs from the other two measures. As the mother–child difference—indexed here by Cohen's *d*—increases from $d = 0.2$ to $d = 0.6$ to $d = 1.0$, ICC(A,1) declines from .68 to .58 to .46. As differences between the fixed groups increase, therefore, ICC(A,1) is attenuated in value. For this reason, reporting parent–child correlations in this metric would enable those who work with adoption data to report their findings in a way that considers simultaneously the two highly prominent group and individual difference effects that are present when children are adopted away into environments more favorable to their development.

A Flow Chart for Selecting an ICC

As an aid to readers in selecting the ICC in Tables 4 and 5 that is appropriate for their data and conceptual purpose, we have included the flow chart given as Figure 1. The series of decisions terminate with the designation of an ICC.

Forming Inferences About ICCs

With the foregoing as background that will aid in selecting and interpreting a calculation formula for an ICC, we turn now to the issues of forming inferences about ICCs.

Table 6
A Comparison of ICC(A,1), (C,1) and Pearson r Correlation Coefficients for Measuring the Correlation Between IQ Scores That Differ in Their Means

Mean difference = 3 pts (d = 0.20)		Mean difference = 9 pts (d = 0.60)		Mean difference = 15 pts (d = 1.00)	
Mother's	Child's	Mother's	Child's	Mother's	Child's
103	119	97	119	91	119
82	65	76	65	70	65
116	106	110	106	104	106
102	102	96	102	90	102
99	105	93	105	87	105
98	100	92	100	86	100
104	107	98	107	92	107
62	85	56	85	50	85
97	101	91	101	85	101
107	110	101	110	95	110
M = 97	100	91	100	85	100
SD = 15	15	15	15	15	15
r = 0.670		0.670		0.670	
ICC(C,1) = 0.670		0.670		0.670	
ICC(A,1) = 0.679		0.584		0.457	

Note. ICC = intraclass correlation coefficient; pts = points.

Confidence Intervals for ICC Population Values (ρ)

Confidence intervals on the population value of ICCs (ρ) for one-way and two-way models are given in Shrout and Fleiss (1979) and were developed by Haggard (1958) and Fleiss and Shrout (1978). The formulas for the upper and lower limits to the $1 - \alpha$ confidence intervals are given in Table 7. One notes that the procedure is more complicated for Type A ICCs that contain column variance in their denominators than for Type C ICCs that ignore this source of variance.

F Tests

In addition to using one's sample data to compute confidence intervals, researchers frequently use their data to test hypotheses about the population value of ρ. The most common is to test the hypothesis that $\rho = 0$ against the alternative that $\rho > 0$. The row effects F statistic (MS_R/MS_W for one-way designs and MS_R/MS_E for two-way designs) serves this purpose. That is, the test for the significance of differences among the row means also serves to test the hypothesis that ρ is zero.

Although tests of the hypothesis $\rho = 0$ are common, they are not particularly informative. In studies of test score reliability, twin resemblance, and rater agreement, nonzero correlations are assumed. In these contexts, it is more useful to determine whether the obtained value $\hat{\rho}$ permits the inference that ρ exceeds some nonzero value. One might, for example, determine whether to accept the hypothesis that ρ exceeds the small, medium, or large effect size criteria set up by Cohen (Cohen, 1988, p. 83) or whether a parent–child ICC of .65 is sufficiently greater than the theoretical limit of .50 to conclude that assortative mating has occurred. Test statistics to conduct these tests on ICCs defined using one-way and two-way models are given in Table 8 (see Appendix A for their development). For the Type C ICCs, the Table 8 statistics are the product of the row effect F statistic and a quantity that equals 1.00 when the null hypothesis value, ρ_0, equals 0 and that equals an ever smaller fractional value as ρ_0 approaches 1.00. Thus the test statistic is a fractional value of the original row effect F value. For Type A ICCs, the appropriate F for H_0: $\rho = \rho_0$, $\rho_0 > 0$, is obtained by multiplying MS_C and MS_E by factors that are dependent on n, the value of ρ_0 and, in the case of ICC(A,1), k. Schönemann (1991) gave a test statistic for Type (C,1) coefficients equivalent to the one given here, but only for $k = 2$.

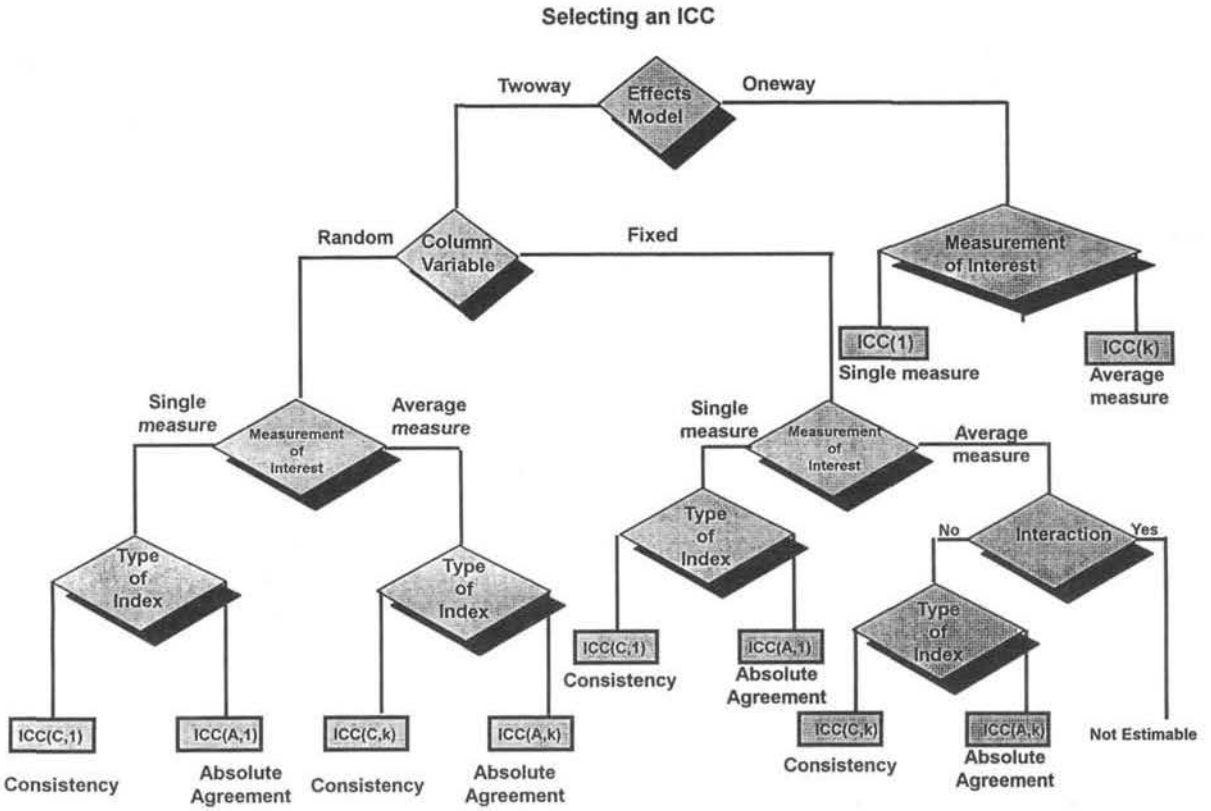

Figure 1. Flow chart for selecting an appropriate intraclass correlation coefficient (ICC).

The circumstances under which the Table 8 F statistics lead to a rejection of a one-tailed or two-tailed H_0 are the same as for any F test. For one-tailed tests (H_1: $\rho > \rho_0$), H_0 is rejected when the p value, $P(F \geq F_{\text{Observed}})$, is less than or equal to α. For two-tailed tests, one must consider $P(F \leq F_{\text{Observed}})$ as well as $P(F \geq F_{\text{Observed}})$. The smaller of these must be less than or equal to $\alpha/2$ to reject H_0. These probabilities may be easily obtained using any statistical package that includes a routine for calculating probabilities using the F distribution (e.g., SAS, 1989).

Heretofore, the common method of conducting significance tests concerning nonzero values of ρ_0 has been to use Fisher's r to z' transformation (see Appendix B). As an alternative to Fisher's z test, the F tests given in Table 8 are clearly preferable. For all but the Type A coefficients, they provide exact probabilities, not approximate ones as does Fisher's test. More important, the Table 8 tests differentiate among forms of intraclass correlation, whereas Fisher's test does not. In addition, the tests have an advantage in computational ease because the terms needed to produce the F statistics come directly from the analyses of variance used to calculate $\hat{\rho}$, and the tests are somewhat more powerful than Fisher's z' test, at least for those formulas that permit an exact determination of power.[3] This small but consistent power advantage is reflected also in the confidence interval procedures because confidence intervals obtained using the formulas of Table 7 will be smaller than those obtained using Fisher's alternative procedure. A word of caution, however, is that the magnitude of correlations may be difficult to interpret when nonnormality is present; so when normality

[3] Because the denominator degrees of freedom for the F statistics calculated on Type (A, 1) and Type (A, k) coefficients are sample dependent, the power of these statistics for detecting a true H_1 cannot be determined exactly.

Table 7
Upper and Lower Limits on $1 - \alpha$ Confidence Intervals on ρ for One-Way and Two-Way Models

ICC type	Confidence interval limits: Lower limit	Confidence interval limits: Upper limit
	One-way model	
ICC(1) for Case 1	$\frac{F_L - 1}{F_L + (k-1)}$ where $F_L = F_{obs}/F_{tabled}$ F_{obs} is the row effects F from the ANOVA. F_{tabled} denotes the $(1 - \frac{1}{2}\alpha) \times 100$th percentile of the F distribution with $n - 1$ numerator degrees of freedom and $n(k-1)$ denominator degrees of freedom.	$\frac{F_U - 1}{F_U + (k-1)}$ where $F_U = F_{obs} \times F_{tabled}$ F_{obs} is the row effects F from the ANOVA. F_{tabled} denotes the $(1 - \frac{1}{2}\alpha) \times 100$th percentile of the F distribution with $n(k-1)$ numerator degrees of freedom and $n - 1$ denominator degrees of freedom.
ICC(k) for Case 1	$1 - \frac{1}{F_L}$ where F_L is defined as above.	$1 - \frac{1}{F_U}$ where F_U is defined as above.
	Two-way models	
ICC(C,1) for Cases 2, 2A, 3, and 3A	$\frac{F_L - 1}{F_L + (k-1)}$ where $F_L = F_{obs}/F_{tabled}$ F_{obs} is the row effects F from the two-way ANOVA. F_{tabled} denotes the $(1 - \frac{1}{2}\alpha) \times 100$th percentile of the F distribution with $n - 1$ numerator degrees of freedom and $(n-1)(k-1)$ denominator degrees of freedom.	$\frac{F_U - 1}{F_U + (k-1)}$ where $F_U = F_{obs} \times F_{tabled}$ F_{obs} is the row effects F from the two-way ANOVA. F_{tabled} denotes the $(1 - \frac{1}{2}\alpha) \times 100$th percentile of the F distribution with $(n-1)(k-1)$ numerator degrees of freedom and $n - 1$ denominator degrees of freedom.
ICC(C,k) for Cases 2, 2A, and 3A, but not for 3.	$1 - \frac{1}{F_L}$ where F_L is defined as for Type (C,1) above.	$1 - \frac{1}{F_U}$ where F_U is defined as for Type (C,1) above.
ICC(A,1) for Cases 2, 2A, 3, and 3A	$\frac{n(MS_R - F_* MS_E)}{F_*[kMS_C + (kn - k - n)MS_E] + nMS_R}$ F_* denotes the $(1 - \frac{1}{2}\alpha) \times 100$th percentile of the F distribution with $n - 1$ numerator degrees of freedom and v denominator degrees of freedom.	$\frac{n(F^* MS_R - MS_E)}{kMS_C + (kn - k - n)MS_E + nF^* MS_R}$ F^* denotes the $(1 - \frac{1}{2}\alpha) \times 100$th percentile of the F distribution with v numerator degrees of freedom and $n - 1$ denominator degrees of freedom.
	where $v = \frac{(aMS_C + bMS_E)^2}{\frac{(aMS_C)^2}{k-1} + \frac{(bMS_E)^2}{(n-1)(k-1)}}$ and $a = \frac{k(\hat{\rho})}{n(1-\hat{\rho})}, b = 1 + \frac{k\hat{\rho}(n-1)}{n(1-\hat{\rho})}$	
ICC(A,k) for Cases 2, 2A, and 3A, but not 3	$\frac{n(MS_R - F_* MS_E)}{F_*(MS_C - MS_E) + n(MS_R)}$ where F_* is defined as for ICC(A,1) above.	$\frac{n(F^* MS_R - MS_E)}{MS_C - MS_E + nF^* MS_R}$ where F^* is defined as for ICC(A,1) above.

Note. ICC = intraclass correlation coefficient; ANOVA = analysis of variance; obs = observed; MS_R = mean square for rows; MS_E = mean square error; MS_C = mean square for columns.

Table 8
Test Statistics for Testing the Null Hypothesis H_0: $\rho = \rho_0$

Model and ICC	Formulas for F statistics: Type 1 ICCs	Formulas for F statistics: Type k ICCs	df_1	df_2
Case 1: One-way random	$\frac{MS_R}{MS_W} \times \frac{1-\rho_0}{1+(k-1)\rho_0}$	$\frac{MS_R}{MS_W}(1-\rho_0)$ or $\frac{1-\rho_0}{1-\hat{\rho}}$	$n-1$	$n(k-1)$
Cases 2 and 2A, 3 and 3A: Two-way random and mixed effects				
Type C ICCs	$\frac{MS_R}{MS_E} \times \frac{1-\rho_0}{1+(k-1)\rho_0}$	$\frac{MS_R}{MS_W}(1-\rho_0)$ or $\frac{1-\rho_0}{1-\hat{\rho}}$	$n-1$	$(n-1)(k-1)$
Type A ICCs	$\frac{MS_R}{aMS_C + bMS_E}$ where $a = \frac{k(\rho_0)}{n(1-\rho_0)}$, $b = 1 + \frac{k\rho_0(n-1)}{n(1-\rho_0)}$	$\frac{MS_R}{cMS_C + dMS_E}$ where $c = \frac{\rho_0}{n(1-\rho_0)}$, $d = 1 + \frac{\rho_0(n-1)}{n(1-\rho_0)}$	$n-1$	where for Type$(A,1)$ $v = \frac{(aMS_C + bMS_E)^2}{\frac{(aMS_C)^2}{k-1} + \frac{(bMS_E)^2}{(n-1)(k-1)}}$ or for Type(A,k) $v = \frac{(cMS_C + dMS_E)^2}{\frac{(cMS_C)^2}{k-1} + \frac{(dMS_E)^2}{(n-1)(k-1)}}$

Note. ρ_0 is the hypothesized value of ρ, and $\hat{\rho}$ is the intraclass correlation coefficient (ICC). MS_R = mean square for rows; MS_W = mean square for residual sources of variance; MS_E = mean square error; MS_C = mean square for columns.

is in question, one might consider transforming one's data before applying a test statistic. For one-way random effects data, Wilcox (1994) suggests using Winsorized ICCs.

Conclusion

In summary, when investigators are concerned with the consistency or absolute agreement among multiple (k) observations made on randomly selected objects of measurement and when the error variance for measures is uniform across the conditions of measurement, ICCs provide the appropriate measure. Confidence intervals and test statistics exist for each of the ICCs that can be defined for one-way and two-way models. These are exact for the most part, and therefore they are preferable to confidence intervals and one-sample tests conducted using Fisher's z' transformations of ρ.

References

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

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83,* 460–472.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth: Holt, Rinehart and Winston.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Fagot, R. F. (1993). A generalized family of coefficients of relational agreement for numerical scales. *Psychometrika, 58,* 357–370.

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.). Edinburgh: Oliver and Boyd.

Fleiss, J. L., & Shrout, P. E. (1978). Approximate interval estimation for a certain intraclass coefficient. *Psychometrika, 43,* 259–262.

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

Harris, J. A. (1913). On the calculation of intraclass and interclass coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika, 9,* 446–472.

Hartmann, D. P. (1982). Assessing the dependability of observational data. In D. P. Hartmann (Ed.), *Using observers to study behavior* (pp. 51–65). San Francisco: Jossey-Bass.

Horn, J. M., Loehlin, J. C., & Willerman, L. (1979). Intellectual resemblance among adoptive and biological relatives: The Texas Adoption Project. *Behavior Genetics, 9,* 177–207.

Jackson, S. E., & Brashers, D. E. (1994). *Random factors in ANOVA.* Thousand Oaks, CA: Sage Publications.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Rajaratnam, N. (1960). Reliability formulas for independent decision data when reliability data are matched. *Psychometrika, 25,* 261–271.

SAS Institute, Inc. (1989). SAS/STAT, Version 6. Cary, NE: Author.

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

Schönemann, P. H. (1991). On non-null tests of intraclass correlations. *Chinese Journal of Psychology, 33,* 3–10.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

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

Skodak, M., & Skeels, H. M. (1949). A final follow-up study of one hundred adopted children. *The Journal of Genetic Psychology, 75,* 85–125.

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10,* 343–366.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Erlbaum.

Turkheimer, E. (1991). Individual and group differences in adoption studies of IQ. *Psychological Bulletin, 110,* 392–405.

Wilcox, R. R. (1994). Estimating Winsorized correlations in a univariate or bivariate random effects model. *British Journal of Mathematical and Statistical Psychology, 47,* 167–183.

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

(*Appendix follows on next page*)

Appendix A

Derivation of Test Statistics

The model labels and ICC definitions in this appendix are from Tables 1, 4, and 5. The expected mean squares are from Table 3. The label ICCP denotes the population intraclass correlation. F statistics for testing H_0: ICCP = ρ_0 are derived below,

For Case 1: One-Way Random Effects Model

$$H_0\text{: ICCP(1)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2} = \rho_0. \tag{A1}$$

$$\text{Under } H_0, \quad \frac{k\sigma_r^2 + \sigma_w^2}{\sigma_w^2} = \frac{1 + (k-1)\rho_0}{1 - \rho_0}.$$

From the table of expected mean squares, $E(MS_R) = k\sigma_r^2 + \sigma_w^2$, and $E(MS_W) = \sigma_w^2$.

$$\text{Then} \quad F = \frac{MS_R}{k\sigma_r^2 + \sigma_w^2} \Big/ \frac{MS_W}{\sigma_w^2} = \frac{MS_R}{MS_W} \times \frac{\sigma_w^2}{k\sigma_r^2 + \sigma_w^2} = \frac{MS_R}{MS_W} \times \frac{1 - \rho_0}{1 + (k-1)\rho_0},$$

which under H_0 has an F distribution with $df = n - 1$, $n(k - 1)$.

$$H_0\text{: ICCP(k)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2/k} = \rho_0. \tag{A2}$$

$$\text{Under } H_0, \quad \frac{k\sigma_r^2 + \sigma_w^2}{\sigma_w^2} = \frac{1}{1 - \rho_0}.$$

$$\text{Then} \quad F = \frac{MS_R}{k\sigma_r^2 + \sigma_w^2} \Big/ \frac{MS_W}{\sigma_w^2} = \frac{MS_R}{MS_W} \times \frac{\sigma_w^2}{k\sigma_r^2 + \sigma_w^2} = \frac{MS_R}{MS_W} \times (1 - \rho_0) = \frac{1 - \rho_0}{1 - \hat{\rho}}, \text{ where } \hat{\rho} = \frac{MS_R - MS_E}{MS_R},$$

which under H_0 has an F distribution with $df = n - 1$, $n(k - 1)$.

Case 2: Two-Way Random Effects Model With Interaction

$$H_0\text{: ICCP(C,1)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2} = \rho_0. \tag{A3}$$

Replacing σ_w^2 and MS_W in Equation A1 with σ_e^2 and MS_E, respectively, one can easily show the test statistic is

$$F = \frac{MS_R}{MS_E} \times \frac{1 - \rho_0}{1 + (k-1)\rho_0},$$

which under H_0 has an F distribution with $df = n - 1$, $(n - 1)(k - 1)$.

$$H_0\text{: ICCP(C,k)} = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k} = \rho_0. \tag{A4}$$

Replacing σ_w^2 and MS_W in Equation A2 with σ_e^2 and MS_E, respectively, one can easily show the test statistic is

$$F = \frac{MS_R}{MS_E} \times (1 - \rho_0) = \frac{1 - \rho_0}{1 - \hat{\rho}}, \text{ where } \hat{\rho} = \frac{MS_R - MS_E}{MS_R},$$

which under H_0 has an F distribution with $df = n - 1$, $(n - 1)(k - 1)$.

$$H_0\text{: ICCP(A,1)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2} = \rho_0. \tag{A5}$$

Under H_0,

$$k\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2 = \left[1 + \frac{k\rho_0}{1 - \rho_0} - \frac{k\rho_0}{n(1 - \rho_0)}\right](\sigma_{rc}^2 + \sigma_e^2) + \frac{k\rho_0(n\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)}{n(1 - \rho_0)}.$$

From the table of expected mean squares for Case 2,

$$E(MS_R) = k\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2,\ E(aMS_C) = a(n\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2), \text{ and } E(bMS_E) = b(\sigma_{rc}^2 + \sigma_e^2), \text{ where}$$

$$a = \frac{k\rho_0}{n(1 - \rho_0)},\ b = 1 + \frac{k\rho_0}{1 - \rho_0} - \frac{k\rho_0}{n(1 - \rho_0)}.$$

Let $F = \dfrac{MS_R}{k\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2} \Big/ \dfrac{aMS_C + bMS_E}{a(n\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2) + b(\sigma_{rc}^2 + \sigma_e^2)}$.

Then under H_0

$$F = \frac{MS_R}{aMS_C + bMS_E}$$

has an approximate F distribution (see Satterthwaite, 1946) with $df = n - 1$ and

$$m = (aMS_C + bMS_E)^2 \Big/ \left[\frac{(aMS_C)^2}{k-1} + \frac{(bMS_E)^2}{(n-1)(k-1)}\right].$$

$$H_0\text{: ICCP(A,k)} = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k} = \rho_0. \quad \text{(A6)}$$

$$\text{Under } H_0,\ k\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2 = \frac{\rho_0}{n(1-\rho_0)}(n\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2) + \left(1 + \frac{\rho_0}{1-\rho_0} - \frac{\rho_0}{n(1-\rho_0)}\right)(\sigma_{rc}^2 + \sigma_e^2).$$

Following the same steps as in Equation A5, it can be shown that a statistic for testing H_0 is

$F = \dfrac{MS_R}{cMS_C + dMS_E}$, which has an approximate F distribution under H_0, with $df = n - 1$ and

$$m' = (cMS_C + dMS_E)^2 \Big/ \left[\frac{(cMS_C)^2}{k-1} + \frac{(dMS_E)^2}{(n-1)(k-1)}\right],$$

where $c = \dfrac{\rho_0}{n(1-\rho_0)}$ and

$$d = 1 + \frac{\rho_0}{1-\rho_0} - \frac{\rho_0}{n(1-\rho_0)} = 1 + \frac{\rho_0(n-1)}{n(1-\rho_0)}.$$

Case 2A: Two-Way Random Effects Model, Interaction Absent

$$H_0\text{: ICCP(C,1)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} = \rho_0, \text{ same as in Equation A3.} \quad \text{(A7)}$$

$$H_0\text{: ICCP(C,k)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2/k} = \rho_0, \text{ same as in Equation A4.} \quad \text{(A8)}$$

$$H_0\text{: ICCP(A,1)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2} = \rho_0, \text{ same as Equation A5.} \quad \text{(A9)}$$

$$H_0\text{: ICCP(A,k)} = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_e^2)/k} = \rho_0, \text{ same as Equation A6.} \quad \text{(A10)}$$

Case 3: Two-Way Mixed Model With Interaction

$$H_0\text{: ICCP(C,1)} = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2} = \rho_0. \quad \text{(A11)}$$

$$\text{Under } H_0,\ k\sigma_r^2 + \sigma_e^2 = \frac{1 + (k-1)\rho_0}{(1-\rho_0)} \times \left(\frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2\right).$$

From the table of expected mean squares

$$E(MS_R) = k\sigma_r^2 + \sigma_e^2 \text{ and } E(MS_E) = \frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2.$$

Let $F = \dfrac{MS_R}{k\sigma_r^2 + \sigma_e^2} \Big/ \dfrac{MS_E}{\frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2}$.

Then under H_0, $F = \dfrac{MS_R}{MS_E} \times \dfrac{1-\rho_0}{1 + (k-1)\rho_0}$, which is the same as the F in Equation A3.

$$H_0\text{: ICCP(C,k)} = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k}, \text{ which cannot be tested.} \quad \text{(A12)}$$

$$H_0\text{: ICCP(A,1)} = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \theta_c^2 + \sigma_{rc}^2 + \sigma_e^2} = \rho_0. \quad \text{(A13)}$$

Under H_0,

$$k\sigma_r^2 + \sigma_e^2 = \frac{k\rho_0}{n(1-\rho_0)} \times \left(n\theta_c^2 + \frac{k}{(k-1)}\sigma_{rc}^2 + \sigma_e^2\right) + \left(1 + \frac{k\rho_0}{1-\rho_0} - \frac{k\rho_0}{n(1-\rho_0)}\right)\left(\frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2\right).$$

From the table of expected mean squares

$$E(MS_R) = k\sigma_r^2 + \sigma_e^2,\ E(MS_C) = n\theta_c^2 + \frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2, \text{ and } E(MS_E) = \frac{k}{k-1}\sigma_{rc}^2 + \sigma_e^2.$$

Following the same steps as in Equation A5, one can easily see the test statistic to be the same as in Equation A5.

$$H_0\text{: ICCP(A,k)} = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\theta_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}, \quad \text{(A14)}$$

which cannot be tested.

Case 3A: Two-Way Mixed Model, Interaction Absent

$$H_0\text{: ICCP(C,1)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} = \rho_0, \text{ same as in Equation A3.} \quad \text{(A15)}$$

$$H_0\text{: ICCP(C,k)} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2/k} = \rho_0, \text{ same as in Equation A4.} \quad \text{(A16)}$$

$$H_0\text{: ICCP(A,1)} = \frac{\sigma_r^2}{\sigma_r^2 + \theta_c^2 + \sigma_e^2} = \rho_0, \text{ same as in Equation A5.} \quad \text{(A17)}$$

$$H_0\text{: ICCP(A,k)} = \frac{\sigma_r^2}{\sigma_r^2 + (\theta_c^2 + \sigma_e^2)/k} = \rho_0, \text{ same as in Equation A6.} \quad \text{(A18)}$$

Appendix B

Fisher's r to z Transformation for ICCs

Because textbooks generally give only the formula for converting interclass rs to z', it is important to note that the formula is different for converting ICCs, which Fisher also designates as r (Fisher, 1938, p. 225). Rather than using the interclass formula

$$z' = \frac{1}{2}\log\frac{1+r}{1-r},$$

one uses the formula

$$z_i = \frac{1}{2}\log\frac{1+(k-1)r}{1-r},$$

where k is the number of observations made on each object of measurement. The variance of the above statistic is

$$\sigma^2 = \frac{k}{2(n-2)(k-1)},$$

where n is the number of objects of measurement and k is, again, the number of observations.

Received July 1, 1995
Revision received September 4, 1995
Accepted October 26, 1995 ■

Correction to McGraw and Wong (1996)

The article "Forming Inferences About Some Intraclass Correlations Coefficients" by Kenneth O. McGraw and S. P. Wong (*Psychological Methods, 1996,* Vol. 1, No. 1, pp. 30–46) contained three errors. The intraclass correlation coefficient (ICC) and r values given in Table 6 (p. 39) of the article should be changed to $r = .714$ for each data set, ICC(C,1) = .714 for each data set, and ICC(A,1) = .720, .620, and .485 for the data in Columns 1, 2, and 3 of the table, respectively.

In Table 7 (p. 41), which is used to determine confidence intervals on population values of the ICC, the procedures for obtaining the confidence intervals on ICC(A,k) needs to be amended slightly. The definitions of F_* and F^* are said to be the same as for ICC(A,1); however, the degrees of freedom v need to be calculated using

$$c = \frac{\hat{\rho}}{n(1 - \hat{\rho})}$$

in place of a and

$$d = 1 + \frac{\hat{\rho}(n - 1)}{n(1 - \hat{\rho})}$$

in place of b.

On pages 44–46, references to Equations A3, A4, and so forth in the Appendix should be to Sections A3, A4, and so forth. We regret any inconvenience or confusion these errors may have caused.