Reliability of Ratings for Multiple Judges: Intraclass Correlation and Metric Scales

Robert F. Fagot
University of Oregon

Scale-dependent procedures are presented for assessing the reliability of ratings for multiple judges using intraclass correlation. Scale type is defined in terms of admissible transformations, and standardizing transformations for ratio and interval scales are presented to solve the problem of adjusting ratings for "arbitrary scale factors" (unit and/or origin of the scale). The theory of meaningfulness of numerical statements is introduced and the coefficient of relational agreement (Stine, 1989b) is defined as the degree of agreement among judges, with respect to (scale-dependent) empirically meaningful relationships. Other topics discussed include the treatment of variability due to judges in relation to scale type, and the reliability of magnitude estimates in psychophysics.
Index terms: coefficient of agreement, intraclass correlation, meaningfulness, metric scales, reliability of rating scales.

Intraclass correlation is widely used to estimate the reliability of ratings by multiple judges. Several important related issues have been investigated in recent articles (Algina, 1978; Bartko, 1966, 1976; Lahey, Downey, & Saal, 1983; Shrout & Fleiss, 1979), including the problem of the choice of the appropriate intraclass correlation coefficient (ICR)[1]. Another problem, virtually ignored in past studies, is the implications of scale type for intraclass correlation. Although the necessity is widely recognized of standardizing scores when target comparisons are made, the issue has not been addressed in the context of intraclass correlation. It will be shown that such standardization is required before computing the ICR, and further that the standardization procedure is dependent on scale type.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 1, March 1991, pp. 1-11

0146-6216/91/010001-11$1.80

Zegers and ten Berge (1985), and Zegers (1986) derived a general coefficient of association for two variables of the same metric scale type. Their work was extended by Stine (1989b), who derived coefficients for additional scale types, and more importantly for present purposes, formalized a notion of relational agreement based on meaningfulness concepts. In this paper, procedures are presented for the scale-dependent assessment of relational agreement (as a special case of reliability) for multiple judges through the use of intraclass correlation.

Two points should be made about the term "reliability" as used here. First, the focus is on the assessment of reliability for single ratings (of each target or ratee) by each judge, even though the ICR can be used to assess the reliability of mean ratings, as well (cf. Shrout & Fleiss, 1979), and reliability estimates for multiple ratings can be made through procedures in generalizability theory (Brennan, 1983). Second, not all ICR indices are coefficients of agreement, although various forms of the ICR are used as indices of the reliability of ratings by multiple judges.

This paper briefly reviews statistical models for reliability estimation. Scale type is then defined in terms of admissible transformations, and the discussion is restricted to the more commonly used interval and ratio scales. The concept of meaningfulness of statements about numerical scales is discussed, and the coefficient of agreement is defined as the degree of agreement among judges with respect to (scale-dependent)

[1]The abbreviation "ICC" has generally been used widely to denote the intraclass correlation coefficient. Because ICC has also been used to denote the item characteristic curve, the abbreviation ICR is used for intraclass correlation coefficient, with the "R" for reliability.

empirically meaningful relationships (Stine, 1989b). Properties of a rating matrix are specified in the case of perfect agreement among judges, relative to scale type. The problem of "adjusting" ratings for arbitrary scale factors (unit and/or origin of the scale) is considered, and appropriate standardizing transformations relative to scale type are proposed. The ICR computed from standardized ratings is shown to satisfy key properties of invariance and sensitivity. The treatment of variability due to judges in relation to scale type and the reliability of magnitude estimates in psychophysics are also discussed.

Statistical Models for Reliability Estimation

Shrout and Fleiss (1979) thoroughly analyzed different forms of intraclass correlation for reliability studies. The ICR is based on an analysis of variance (ANOVA) model. The population index is typically defined as the ratio of the variance of interest, divided by the sum of the variance of interest plus error. Shrout and Fleiss considered three cases. Case 1 was a 1-way ANOVA design in which each target was rated by a different random sample of judges. Case 2 (proposed by Bartko, 1966, and Rajaratnam, 1960) and Case 3 were two-way ANOVA designs, both of which assumed that targets were randomly sampled from some population. Judges were treated as a random factor for Case 2 and as a fixed factor for Case 3. The effects due to judges and to the Judge × Target interaction were separable for Cases 2 and 3, unlike Case 1. The linear model was also given for each case, and it was explained how the estimators were derived, based on the expected mean squares for each model. All of the estimators they provided are biased but consistent.

Shrout and Fleiss (1979) presented the following formulas for these intraclass reliability estimators:

Case 2: Two-way ANOVA model with random factors of targets and judges

$$\mathrm{ICR}(2,1) = \frac{\mathrm{BMS} - \mathrm{EMS}}{\mathrm{BMS} + (k-1)\mathrm{EMS} + k(\mathrm{JMS} - \mathrm{EMS})/n}, \quad (1)$$

Case 3: Two-way ANOVA model with random factor of targets and fixed factor of judges

$$\mathrm{ICR}(3,1) = \frac{\mathrm{BMS} - \mathrm{EMS}}{\mathrm{BMS} + (k-1)\mathrm{EMS}}, \quad (2)$$

where

k = number of judges
n = number of targets
BMS = target mean square
JMS = judge mean square
EMS = judge × target interaction mean square.

These forms of the ICR may be viewed as a special case of a one-facet generalizability (G) study (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Analyzed as a G study, the ICR is defined in terms of various variance component parameters. To compute the ICR coefficient, the variance components are replaced with their estimators. The resulting generalizability coefficient for a one-facet study will agree with the appropriate mean square estimator given by Shrout and Fleiss (1979), except in the case of a negative variance estimate. If the negative variance estimate is set equal to 0 (the usual procedure), then the generalizability coefficient will differ slightly from the Shrout and Fleiss estimate.

Scale Type and Admissible Transformations

Scale type is defined here in terms of admissible transformations. For *ratio* scales, similarity transformations are the admissible transformations of the scale ϕ:

$$\phi' = \beta\phi \quad , \quad (\beta > 0) \quad . \quad (3)$$

For *interval* scales, positive linear (affine) transformations are the admissible transformations of the scale ϕ:

$$\phi' = \beta\phi + \alpha \quad , \quad (\beta > 0) \quad . \quad (4)$$

For both equations, ϕ is the original representing function (scale) and ϕ' is the transformed representing function (Krantz, Luce, Suppes, & Tversky, 1971, chap. 1). A salient characteristic of

a ratio scale is the existence of a unique zero point. Thus, the unit is the only arbitrary scale factor for judges making ratings on a ratio scale, and different units are "admissible" but different zero points are not.

For an interval scale, changes in unit (β) and zero point (α) are admissible. For example, a change from temperature measured in degrees Centigrade (C) to Fahrenheit (F) would be accomplished by setting $\beta = 9/5$ and $\alpha = 32$, which results in the well-known formula $F = (9/5)C + 32$. Hence, if the judges are making ratings on an interval scale, then both the unit and the zero point may vary among judges.

Meaningfulness and the Coefficient of Relational Agreement

The concept of meaningfulness (with roots in Stevens' theory of scale types; Stevens, 1946) now plays an important role in modern measurement theory. A statement about numerical scales is said to be *meaningful* if and only if its truth value is unchanged whenever admissible transformations are applied to the scales (for discussions of meaningfulness, see Adams, Fagot, & Robinson, 1965; Narens, 1985; Roberts, 1979; Stine, 1989a; and Suppes & Zinnes, 1963). For example, a statement about the ratio of the weights of two objects is meaningful in this sense, because the numerical ratio is unchanged with changes in unit (an admissible transform for a ratio scale). Such a ratio statement would not be meaningful for interval scales for which the addition of a constant (change of origin) is an admissible transformation.

Consistent with this interpretation of meaningfulness, the *coefficient of relational agreement* (sometimes referred to as simply "coefficient of agreement") assesses the degree of agreement among judges with respect to empirically meaningful relationships (Stine, 1989b). This is equivalent to the degree to which the judges' ratings are admissible transformations of each other. For example, the coefficient of agreement assesses the degree to which judges' ratings are *proportional* for ratio scales, and the degree to which the judges' ratings are *linear* for interval scales.

It follows that in any application there may be empirically meaningless disagreement among judges that should not influence the coefficient of agreement, and empirically meaningful disagreement (systematic bias) that should influence it. For example, the use of different units and/or origins ("arbitrary scale factors") by judges for interval scales is empirically meaningless disagreement; if this is the only source of disagreement, the judges' ratings are related by admissible transformations.

If, on the other hand, the judges' ratings are non-linear, this represents meaningful disagreement, which should influence (attenuate) the coefficient of agreement. Similarly, if judges are making ratings on a ratio scale, the use of different units by the judges is empirically meaningless disagreement. However, if the judges' ratings differ systematically by an additive constant, this represents empirically meaningful disagreement (relative to a ratio scale) and should influence (attenuate) the coefficient of agreement.

Scale Type and Intraclass Correlation

Bartko (1976) criticized the Winer (1971, pp. 289–296) "adjustment for anchor points" method of estimating the ICR, in part because Winer's method yields an ICR of unity if the judges display a "constant additive bias." Bartko illustrated his argument with three sets of data (Table 1), each set consisting of two ratings of

Table 1
Three Sets of Ratings on a 1 to 10 Scale

	1a		1b		1c	
Target	J_1	J_2	J_1	J_2	J_1	J_2
1	1	1	1	5	1	2
2	2	2	2	6	2	4
3	3	3	3	7	3	6
4	4	4	4	8	4	8
5	5	5	5	9	5	10

Note. Adapted from Bartko, J. J. (1976). Copyright 1979 by the American Psychological Association. Reprinted by permission.

each of five targets. The two ratings are identical for each target of the first set (1a); they differ by the same additive constant for each of the targets of the second set (1b); and the second rating is double the first of each target for the third set (1c).

Bartko interpreted the data as exhibiting perfect "reliability" for 1a (identicality), "constant additive bias" for 1b, and "multiplicative bias" for 1c. However, although data such as 1b and 1c may indicate systematic bias (empirically meaningful disagreement) in the ratings, such patterns of ratings also may be due entirely to "arbitrary scale factors" (empirically meaningless disagreement). For example, if dataset 1b resulted from ratings made on an interval scale with the two judges selecting their own origin, the ratings would show perfect agreement *relative to an interval scale*, and would exhibit no systematic bias. Similarly, if dataset 1c resulted from ratings made on a ratio scale with the two judges permitted to use their own (arbitrary) unit, the observed proportionality of the ratings would indicate perfect agreement *relative to a ratio scale*, and would show no systematic bias.

These examples illustrate the role of scale type and meaningfulness in assessing reliability, and the need to adjust for arbitrary scale factors before computing the ICR. This adjustment (standardization) should be carried out by applying an appropriate transformation of the ratings, dependent on scale type.

Properties of a Judge × Target Rating Matrix

The assumption of perfect agreement for interval and ratio scales, which are the most commonly used "metric" scales, has implications for the properties of a Judge × Target rating matrix. Consider the hypothetical data of Table 2, for three judges (J_1, J_2, J_3) and three targets (T_1, T_2, T_3). For Table 2, the spacing of the ratings on the assumed interval scale is the same for each judge (i.e., the judges' ratings are perfectly linear). In fact, the ratings were constructed using the well-known interval scales of temperature: J_1 is the C scale, J_2 is the F scale, and J_3 is another hypothetical interval temperature scale. The transformation between scales 3 and 1, for example, is $J_3 = (7/5)J_1 + 10$. Table 2 demonstrates that perfect agreement for interval scales is achieved if judges' ratings are linear (i.e., are admissible transformations for an interval scale), given that both unit and origin are free to vary among judges.

Table 2
Ratings by Three Judges of Three Targets on an Interval Scale With Perfect Agreement

	Target		
Judge	T_1	T_2	T_3
J_1	−5	10	30
J_2	23	50	86
J_3	3	24	52

In the ratio scale example of Table 3, the ratio of ratings for each pair of targets is the same for all judges, (e.g., the rating for T_3 is three times that of T_2 for all judges). These ratings were derived as measures of length using scales with units of feet (J_1), inches (J_2), and meters (J_3). The data illustrate the case of judges making ratings on a ratio scale, with the arbitrary unit selected by each judge. In principle, then, perfect agreement for ratio scales is achieved if the judges' ratings are proportional (i.e., are admissible transformations for a ratio scale), given that the unit is selected by the judges.

Table 3
Ratings by Three Judges of Three Targets on a Ratio Scale With Perfect Agreement

	Target		
Judge	T_1	T_2	T_3
J_1	1	2	6
J_2	12	24	72
J_3	.305	.610	1.830

Standardization

Given that the data of Table 2 exhibit perfect agreement for an interval scale, what would the

reliability estimates be using ICR(2,1) and ICR(3,1) that treat the data as sample ratings? Depending on the assumed statistical model, the estimates are ICR(2,1) = .55 and ICR(3,1) = .92, which obviously are less than perfect reliability. Given that the ratings in Table 2 are measured on an interval scale with perfect agreement, it is clear that the attenuated reliability estimates are due entirely to arbitrary scale factors, the effects of which must be removed before computing the ICR. Because the "judges" are temperature scales, the problem can be solved simply by transforming to a single temperature scale, in which case the ratings for each target would be identical, and all ICR = +1. However, a standard scale is not typically known in psychological research, and a more general standardization procedure is required.

Zegers and ten Berge (1985) and Zegers (1986) derived a family of association coefficients for metric scales that could be used to estimate reliability (relative to scale type) in the special case of just two judges. Zegers and ten Berge based their family of coefficients on a "standardized version" of the variables relative to scale type, and it is their concept of standardized variables that is adapted here as a solution to the standardization problem for intraclass correlation. (Zegers and ten Berge use the terms "uniformed version," and "uniforming" transformation.)

The standardizing transformation must be an admissible transformation relative to scale type (i.e., it must satisfy Equation 3 for ratio scales and Equation 4 for interval scales). Standardizing transformations satisfying this condition for interval scales (Z) and ratio scales (P) are

$$Z_{ij} = (Y_{ij} - M_i)/S_i \tag{5}$$

for interval scales, and

$$P_{ij} = Y_{ij}/K_i \quad , \tag{6}$$

where

$$K_i = \left(\sum_{j=1}^{n} Y_{ij}^2/n\right)^{1/2} \tag{7}$$

for ratio scales. The Y_{ij} are the observed ratings ($i = 1, 2, \ldots, k$ judges; $j = 1, 2, \ldots, n$ targets), and M_i and S_i are the mean and standard deviation (SD), respectively, for judge i.

Note that for both Equations 5 and 6, the standardizing transformation is applied separately to each judge. Equation 5 is the well-known linear Z transformation (with mean = 0 and SD = 1), but P appears to have been proposed first by Zegers and ten Berge (1985) in the context of bivariate correlation. P rescales the variables to obtain a mean squared value of 1. P is not unique, but Zegers (1986) has shown that it maximizes the value of the correlation coefficient for ratio scales in the bivariate case.

It was pointed out above that the data of Table 2 exhibit perfect agreement relative to an interval scale, yet the estimated reliability was less than 1, due to the attenuating effect of arbitrary scale factors. However, the appropriate standardization (Equation 5) on Table 2 results in the transformed ratings of Table 4, which shows that the transformed ratings are identical for each target, resulting in ICR(2,1) = ICR(3,1) = +1.

Table 4
Standardized Ratings for Table 2 Data (Table 2 Ratings Adjusted for Arbitrary Unit and Origin by Equation 5)

Judge	T_1	T_2	T_3
J_1	−.95	−.10	1.05
J_2	−.95	−.10	1.05
J_3	−.95	−.10	1.05

Table 3 illustrates perfect agreement for ratio scales, based on the proportionality of the judges' ratings. The reliability estimates, treating these data as sample ratings, are ICR(2,1) = .06 and ICR(3,1) = .11. Applying the appropriate standardizing transformation to the ratings (Equation 6), $K_1 = 3.7$, $K_2 = 44.36$, and $K_3 = 1.13$. Dividing the ratings by the appropriate K_i resulted in the transformed ratings in Table 5. Note that the transformed ratings are identical for each target, as in Table 4, resulting in ICR(2,1) = ICR (3,1) = +1.

Thus, a precise pattern for judges' ratings is

Table 5
Standardized Ratings for Table 3 Data (Table 3 Ratings Adjusted for Arbitrary Unit by Equation 6)

Judge	T_1	T_2	T_3
J_1	.27	.54	1.62
J_2	.27	.54	1.62
J_3	.27	.54	1.62

implied by perfect agreement, dependent on scale type. The judges' ratings are linear for interval scales, and the ratings are proportional for ratio scales. Applying the appropriate standardizing transformation resulted in transformed ratings that are identical for each target. Hence the coefficient of agreement may be defined alternatively as the degree to which the standardized versions are identical.

The identicality of transformed ratings results in a mean square of 0 for the main effect of judges, and for the interaction of judges and targets. Provided that the between-targets mean square is greater than 0, all ICR estimates are unity. Standardization avoids underestimating ICR due to disagreement about empirically meaningless relationships in the ratings.

Invariance and Sensitivity

The standardized ICR coefficient satisfies two important properties: (1) invariance in value under admissible transformations of the ratings, and (2) sensitivity to non-admissible transformations (Zegers & ten Berge, 1985)—that is, the value of the ICR coefficient will change if a non-admissible transformation of the ratings is performed. As an example from bivariate correlation, the Pearson r (a member of the family of coefficients appropriate for interval scales of Zegers and ten Berge, 1985) is invariant under linear (affine) transformations, which are the admissible transformations of an interval scale, and it is sensitive to non-admissible transformations such as a log transformation.

To illustrate the properties of invariance and sensitivity for the ICR, consider Tables 3 and 5, the former consisting of ratings made on a ratio scale, and the latter of transformed ratings using Equation 6. Multiplying the ratings for J_2 in Table 3 by 2 would be an admissible transformation (change of unit), and hence the transformed ratings for J_2 in Table 5 would not change (invariance), and the ICR value of unity would be preserved.

Transforming the ratings for J_2 in Table 3 by $J_2 + 2$ would be a non-admissible transformation for a ratio scale, and the ratings of J_2 would no longer be proportional to J_1 and J_3. The transformed ratings for J_2 in Table 5 would change to .304, .565, and 1.609 for T_1, T_2, and T_3, respectively, with a consequent change in ICR to less than 1, thus satisfying sensitivity.

This example also explains why the Z transformation (Equation 5) would not be appropriate for ratio scales. Z is a non-admissible transformation because it permits an additive constant, whereas the zero point for a ratio scale is unique. As a result, any transformation satisfying Equation 4 but not Equation 3 will preserve the value of the ICR, thereby satisfying invariance but not sensitivity. Using the Z transformation for ratio scales will generally overestimate ICR by removing variability due to a possible systematic (additive) bias (meaningful disagreement) reflected in the additive constant.

Rating Scales, Scale Type, and Standardization

The analysis of scale type and standardization has been made with respect to "idealized" rating scales for which the scale type is known, whereas the concept of scale type is fuzzy at best in most applications of rating scales. Nevertheless, whenever a numerical rating scale is constructed, it is assumed implicitly that the judges are capable of making direct *quantitative* ratings of stimuli relative to some attribute of interest, and it may be inferred from the instructions to the judges what the class of admissible transformations is assumed to be. A similar point was made by Krantz (1972).

The central idea is that although formal theory justifying scale types for numerical rating

scales (including magnitude estimation in psychophysics) is lacking, it is implicitly assumed that the judges are making direct quantitative estimates, which implies an underlying *metric*, and that the use of arbitrary assignments requires adjustment of the ratings. However, the scale-dependent standardization is also determined by the allocation of arbitrary assignments between the experimenter and the judges.

For ratio scales, there are two cases to consider:

R1. If the experimenter for R1 sets the unit of the ratio scale by assigning the same positive number for all judges to any one of the targets, there are no scale factors free to vary among judges. Therefore, there can be no meaningless disagreement among judges. Perfect agreement in this case would require identical ratings for all judges, and the agreement coefficient would assess the degree to which the untransformed ratings were identical.

R2. If judges are allowed to set the unit, the coefficient of agreement assesses the degree to which the ratings are proportional prior to standardization, and the degree to which the transformed ratings are identical after the *P* transformation (Equation 6) is applied.

For interval scales, there are four cases to consider:

I1. If both arbitrary assignments are made by the experimenter, no standardization is required, and the coefficient of agreement assesses the degree to which the untransformed ratings are identical.

I2. If both assignments are made by the judges, unit and origin are free to vary among judges, and the *Z* transformation adjusts for the empirically meaningless disagreement of unit and origin.

I3. If the experimenter sets the same origin for all judges by assigning 0 to one of the targets, perfect agreement requires that the ratings for each pair of judges be related by a *similarity* transformation (proportionality), and hence the *P* transformation (Equation 6) is the appropriate transformation. Note that even though judges are making ratings on an interval scale, *Z* is not the appropriate transformation. This is because once the zero point of the scale is fixed by the experimenter, judges must make ratings relative to the common zero point; thus, only the proportionality of the judges' ratings is now required for perfect agreement. If the judges' ratings are perfectly proportional relative to the fixed origin, both *Z* and *P* transformations would result in an ICR of unity. However, the *Z* transformation would lead to an *inflated* estimate of ICR in the presence of additive bias (meaningful disagreement), which would be treated as empirically meaningless disagreement. The *P* transformation would correctly treat the additive bias as empirically meaningful disagreement that would be allowed to influence (attenuate) the ICR.

Note that the assignment of 0 need not imply "no amount of" the attribute being rated for an interval scale. Therefore, the theory can be applied to bipolar scales for which 0 may be assigned to "neutral," and both positive and negative ratings are admissible.

I4. If the experimenter assigns a positive number to one of the targets, this does not fix the unit of the scale. If it did, only the origin could vary among judges, and the ratings of two judges could differ only by an additive constant in the case of perfect agreement, which obviously is not possible because one target is assigned the same number for all judges. Perhaps this is more readily apparent when it is considered that the origin is a *point* on the scale, but the unit is a *distance*. Thus the assignment of a positive number to a target has no measurement-theoretic significance in terms of restrictions placed on the class of admissible transformations—as does the assignment of 0, which fixes the origin and restricts the

class of admissible transformations to similarity transformations. Although the assignment of the origin for an interval scale is separable from the assignment of the unit, the converse is not true: The unit may be assigned by the experimenter only conjointly with the origin (i.e., requiring two assignments).

The allocation of arbitrary assignments also has implications for Case 1 in Shrout and Fleiss (1979), in which each judge rates only one target in a one-way ANOVA design, ruling out the possibility of standardization of ratings for these designs. If judges use different units or origins, a one-way design is therefore not defensible for the estimation of reliability using intraclass correlation, because the reliability coefficient would be attenuated by empirically meaningless disagreement. However, a simple solution is for the experimenter to make the same arbitrary assignments for all judges (e.g., unit and origin for interval scales), which avoids the problem of standardization of ratings, and renders Case 1 appropriate for estimating rater reliability using intraclass correlation. In this case, of course, perfect reliability would require that the judges' untransformed ratings be identical within each target (within mean square = 0).

Variability Due to Judges

Shrout and Fleiss (1979) discuss whether the effect due to judges may be ignored in the reliability index. In their analysis, the treatment of judge variability is based on statistical assumptions: If judges are a fixed factor (their Case 3), then ICR(3,1) is the appropriate index and the effect of judges is ignored. If judges are a random factor, then ICR(2,1) is the appropriate index and the effect of judges influences reliability.

The effect of judges is apparent from examination of the estimators (Equations 1 and 2). JMS (mean square for judges) appears in the formula for ICR(2,1), but not in the formula for ICR(3,1). For ICR(3,1), however, it is only the *main* effect of judges—an *additive* effect—that is not allowed to influence reliability. A possible *multiplicative* effect may influence the reliability index through the judge × target interaction (EMS), which appears in the formulas for both estimators. One example is dataset 1b in Table 1, in which the ratings differ by a constant and EMS = 0, so that ICR(3,1) = +1. A main effect of judges is present (JMS > 0) and is an influence for Case 2, with ICR(2,1) = .24. For dataset 1c, for which the ratings are proportional (indicating perfect reliability for ratings made on a ratio scale), the proportionality is captured by the interaction (EMS > 0), and both estimators are less than 1.

Measurement-Theoretic Issues in Relation to Judge Variability

The main effect of judges could reflect either empirically meaningful (additive) disagreement or differences among judges due to the use of different origins of the scale (empirically meaningless disagreement). If ratings were made on a ratio scale, however, origin differences among judges would not be admissible, and the use of ICR(3,1) on the (untransformed) ratings would erroneously ignore the effect of empirically meaningful (additive) disagreement on reliability. Conversely, if ratings were made on an interval scale, origin differences would be admissible, and the main (additive) effect of judges would be correctly ignored in the reliability index using ICR(3,1). However, the multiplicative effect of judges, as estimated by the judge × target interaction, may be due to admissible differences in unit. Although ICR(3,1) can be used to adjust for arbitrary zero points, its use with untransformed interval scale ratings would treat unit differences as empirically meaningful disagreement that would erroneously influence (attenuate) the reliability index.

Consider datasets 1b and 1c in Table 1. If the judges made their ratings on a ratio scale, and were free to select their own unit, for both datasets there would be no basis for inferring (as Bartko did) that dataset 1c showed "multiplicative bias," because the proportionality of the ratings could be accounted for by the use of different units by the two judges. The ratings of

dataset 1b, though, showed a constant *difference* in ratings for all targets that cannot be accounted for by the arbitrary unit of the ratio scale, and thus could be interpreted as exhibiting "additive bias," although ICR(3,1) = +1 for the untransformed ratings.

If the *Z* transformation (inappropriate for assumed ratio scale) were used to obtain an estimate of ICR for set 1b, an estimate of unity would be obtained, regardless of which ICR formula was used. This is because the additive "bias" was implicitly treated as an admissible transformation of the scale; therefore, the variability due to this source was removed and not allowed to influence (attenuate) the reliability estimate. If, on the other hand, the *P* transformation (Equation 6) were used, variability due to the additive bias would be allowed to influence reliability, and the estimates would be ICR(2,1) = .78 and ICR(3,1) = .76.

If the ratings for dataset 1b were made on an interval scale, Bartko's (1976) interpretation of additive bias is not justifiable, because the constant difference in ratings may be accounted for by arbitrary zero points. In this case, the *Z* transformation would correctly remove variability due to the constant difference in ratings (an admissible transformation between judges for an interval scale), and all ICR estimates would be unity for the transformed ratings.

Finally, if the ratings for dataset 1c were made on an interval scale, Bartko's (1976) interpretation of multiplicative bias is not justifiable, because the proportionality of the ratings (as in the case of the ratio scale) may be accounted for by admissible differences in unit. The *Z* transformation would correctly remove variability due to the proportionality of the ratings, and all ICR estimates based on *Z* would be unity. However, computed on the untransformed ratings, ICR(3,1) = .80, this result is attenuated because admissible differences in unit were allowed to influence reliability.

The treatment of judge variability is also related to the question of whether ICR(2,1) and ICR(3,1) are coefficients of relational agreement. Three factors are at issue: (1) ICR(3,1) treats the judge factor as a fixed factor and ignores variability due to judges; (2) ICR(2,1) treats the judge factor as a random factor and allows variability due to judges to influence reliability, so by virtue of the statistical models, these two forms of ICR treat the main effect of judges differently; and (3) this differential treatment interacts with meaningfulness issues and standardization, because the judge main effect confounds two possible sources of variability that are not always separable—meaningful (additive) disagreement and meaningless origin differences among judges. These considerations result in the following conclusions:

1. For both ratio and interval scales, ICR(3,1) is *not* a coefficient of relational agreement. By removing the main effect of judges (through statistical assumptions), ICR(3,1) does not allow meaningful (additive) disagreement to influence the coefficient. Shrout and Fleiss (1979) interpret ICR(3,1) as a measure of *consistency*, although they do not consider issues of meaningfulness.
2. ICR(2,1) is a coefficient of relational agreement for ratio scales, because standardization (using *P*) does not remove the main effect of judges, which allows meaningful (additive) disagreement to influence the coefficient.
3. For interval scales, the *Z* transformation correctly removes variability due to origin differences, but also removes (if present) meaningful additive disagreement, because both are confounded in the main effect of judges. If ICR(2,1) is to be interpreted as a coefficient of relational agreement for interval scales, it is necessary to separate these two sources of disagreement. This may be done by using interval scale Case I1 (both assignments made by the experimenter), or I3 (origin only fixed by the experimenter). In both cases, the main effect of judges will be allowed to influence the coefficient, and the resultant variability of judges will be due only to additive disagreement (because the

same origin is used by all judges). Thus ICR(2,1) is a coefficient of relational agreement for Cases I1 and I3.

With respect to the choice between ICR(2,1) and ICR(3,1) as an index of reliability, Bartko (1976) preferred to limit the meaning of rater reliability to "agreement," but Algina (1978) objected to this limitation, pointing out that generalizability theory (Cronbach et al., 1972) includes the case of fixed raters. Shrout and Fleiss (1979, Decision 2, pp. 424–425) discussed issues relating to a choice between these two indices, and gave examples of reliability studies in response to Bartko (1976) for which ICR(3,1) would be the appropriate index. Typically, ICR(3,1) results in higher values than ICR(2,1), because ICR(3,1) ignores meaningful additive disagreement among judges. For example, ICR(2,1) = .29 and ICR (3,1) = .71 for data from Shrout and Fleiss (1979, Table 4, p. 424), indicating substantial meaningful disagreement among judges that is being ignored through the use of ICR(3,1). It should be noted that Algina, Bartko, and Shrout and Fleiss do not discuss "agreement" explicitly in terms of meaningfulness concepts, but the coefficient of relational agreement may be viewed as one formalization of "rater reliability as agreement" in the context of meaningfulness theory.

Reliability of Magnitude Estimates

To demonstrate the effect on reliability of ignoring arbitrary scale factors, the standardization procedure was used to estimate the reliability index for psychophysical data from Fagot and Pokorny (1989). Twelve raters made magnitude estimations of loudness and heaviness for nine log-spaced stimuli. No standard was provided by the experimenter, which allowed raters to select their own units on the assumed ratio scale. The raters made five estimates of each stimulus, but only the last response was used in the reliability study.

The basic data consisted of a 12 × 9 [judge (rater) × target (physical stimulus)] matrix of ratings. Because judges were allowed to select their own unit on an underlying ratio scale, the *P* transformation (Equation 6) was applied to the ratings.

Table 6 shows the reliability estimates for loudness and heaviness, for ratings (*R*) and transformed ratings (*P*) for each of the ICR estimators. The reliability indices are very low for the ratings (.20 and .33 for loudness, .06 and .14 for heaviness), but the estimates (ranging from .85 to .89) show impressive elevation when adjustment was made for the arbitrary unit, with no practical difference between ICR(2,1) and ICR(3,1). This example illustrates the very powerful influence of empirically meaningless disagreement on the reliability index.

Table 6
Reliability Estimates for Magnitude Estimates of Loudness and Heaviness (*R* = Ratings; *P* = Transformed Ratings, Adjusted for Arbitrary Unit)

Reliability Estimate	Loudness		Heaviness	
	R	*P*	*R*	*P*
ICR(2,1)	.20	.85	.06	.88
ICR(3,1)	.33	.85	.14	.89

Directions for Future Research

One goal of future research, suggested by the fact that ICR(3,1) is not a coefficient of relational agreement, could be to develop models for which meaningfulness criteria and the goals of the reliability study are compatible. This would certainly include research on meaningfulness, because there are other possible definitions of meaningful relations (see Narens, 1985, chap. 2.14; Stine, 1989a, 1989b), leading to possibly different definitions of relational agreement. An analysis of the mutual implications of meaningfulness criteria and generalizability theory (Brennan, 1983) would also be of interest.

Another useful direction for future research is the development of indices other than ICR for assessing reliability for multiple judges. As previously mentioned, Stine (1989b) developed procedures for assessing interobserver relational agreement for two judges for various metric scales, based on a family of association coeffi-

cients developed by Zegers and ten Berge (1985). One possibility is the generalization of these results to the case of multiple judges as an alternative to the ICR.

References

Adams, E. W., Fagot, R. F., & Robinson, R. E. (1965). A theory of appropriate statistics. *Psychometrika, 30,* 99–127.

Algina, J. (1978). Comment on Bartko's "On various intraclass correlation reliability coefficients." *Psychological Bulletin, 85,* 135–138.

Bartko, J. J. (1966). The intraclass correlation coefficient 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.

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

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

Fagot, R. F., & Pokorny, R. (1989). Bias effects on magnitude and ratio estimation power function exponents. *Perception and Psychophysics, 45,* 221–230.

Krantz, D. H. (1972). Measurement structures and psychological laws. *Science, 175,* 1427–1435.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I.* New York: Academic Press.

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. *Psychological Bulletin, 93,* 586–595.

Narens, L. (1985). *Abstract measurement theory.* Cambridge MA: MIT Press.

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

Roberts, F. S. (1979). *Measurement theory.* Reading MA: MIT Press.

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Stine, W. W. (1989a). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin, 105,* 147–155.

Stine, W. W. (1989b). Interobserver relational agreement. *Psychological Bulletin, 106,* 341–347.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Vol. II.* New York: Wiley.

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

Zegers, F. E. (1986). A family of chance-corrected association coefficients for metric scales. *Psychometrika, 51,* 559–562.

Zegers, F. E., & ten Berge, J. M. F. (1985). A family of association coefficients for metric scales. *Psychometrika, 50,* 17–24.

Acknowledgments

The author thanks Wm. Wren Stine for helpful comments on an earlier draft of this paper.

Author's Address

Send requests for reprints or further information to Robert Fagot, Department of Psychology, University of Oregon, Eugene OR 97403-1227, U.S.A.