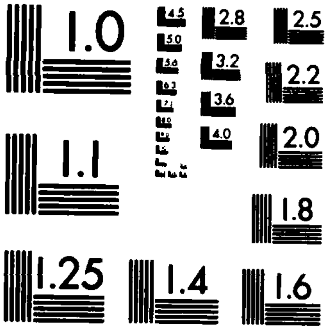
AD-A136 709 ESTIMATING WITHIN-GROUP INTERRATER RELIABILITY WITH AND WITHOUT RESPONSE..(U) GEORGIA INST OF TECH ATLANTA SCHOOL OF PSYCHOLOGY L R JAMES ET AL. 09 DEC 83 1/1

UNCLASSIFIED GT-ONR-4 N00014-80-C-0315 F/G 5/10 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER GT-ONR-4	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Estimating Within-Group Interrater Reliability With and Without Response Bias		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER GT-ONR-4
7. AUTHOR(s) Lawrence R. James, Robert G. Demaree, and Gerrit Wolf		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0315 N00014-83-K-0480
9. PERFORMING ORGANIZATION NAME AND ADDRESS School of Psychology Georgia Institute of Technology Atlanta, Georgia 30332		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR170-904 NR475-026
11. CONTROLLING OFFICE NAME AND ADDRESS Organizational Effectiveness Research Programs and Manpower Personnel and Training Technology Office of Naval Research, Arlington, VA 22217		12. REPORT DATE December 9, 1983 13. NUMBER OF PAGES 36
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

Accession For		
NTIS GRA&I	☑	
~~DTIC TAB~~	□	
Unannounced	□	
Justification		
By		
Distribution/		
Availability Codes		
Dist	Avail and/or Special	
A-1		

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Part of the CNR Manpower R+D Program

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Agreement
Interrater Reliability
Response Bias

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This article presents methods for assessing agreement among the judgments made by a single group of judges on a single variable in regard to a single target. For example, the group of judges could be editorial consultants, members of an assessment center, or members of a team. The single target could be a manuscript, a lower-level manager, or a team. The variable on which the target is judged could be overall publishability in the case of the manuscript, managerial potential for the lower-level manager, or team

20. ABSTRACT (continued)

cooperativeness for the team. The methods presented are based on new procedures for estimating interrater reliability. For situations such as the above, these procedures are shown to furnish more accurate and interpretable estimates of agreement than estimates provided by procedures commonly used to estimate agreement, consistency, or interrater reliability. In addition, the proposed methods include processes for controlling for the spurious influences of response biases (e.g., positive leniency, social desirability) on estimates of interrater reliability.

Estimating Within-Group Interrater Reliability With and Without Response Bias

Many occasions arise in research and practice when it is useful to have an estimate of interrater reliability for judgments of a single target by one set of judges. Examples include the needs to estimate interrater reliability among judges' ratings of (a) the level of performance indicated by a potential anchor for a Behaviorally Anchored Ratings Scale (BARS) in the development phases of a BARS, and (b) the overall "publishability" of a manuscript submitted for journal review. In these examples, the "variable" consists of a single item, with a rating scale such as a seven-point performance scale. It is also helpful to have an index of interrater reliability when scores on a variable consist of means taken over items that are indicators of the same construct. Illustrations include estimates of interrater reliability for (a) a single class, where the data are students' mean scores on items measuring an instructor's "consideration of student needs", and (b) for a single team, where agreement among team members' mean scores on items measuring "team cooperativeness" are used in the design of a team development program.

For these illustrations, each of $\underline{K}$ judges has rated a single target (i.e., a potential BARS anchor, a manuscript, an instructor, a team) on a variable having either a single item (level of performance, publishability) or a set of $\underline{J}$ items that measure the same construct (consideration of student needs, team cooperativeness). For both single and multiple item variables, two assumptions are made. First, the item(s) have been shown to have acceptable psychometric properties (e.g., construct validity, internal consistency in the case of multiple items) in prior research; thus we may focus on interrater reliability for an item or items with known measurement qualities. Second, the alternatives on an item's measurement scale are approximately equally spaced (i.e., an approximately interval response scale such as used in Cooper, 1976; Hsu, 1979). The item response alternatives are assumed to be identical for multiple item variables.

The objective of this article is to introduce procedures for estimating interrater reliability for judgments of a single target by one group of judges, given the assumptions above. The term "interrater reliability" is used here to refer to the degree to which judges are "interchangeable", which is to say the extent to which judges "agree" on a set of judgments (Shrout & Fleiss, 1979, p. 425; see also Bartko, 1976). Mathematically, interrater reliability is typically defined as a proportion, which in this case is the proportion of systematic variance in a set of judgments in relation to the total variance in the judgments. Total variance will be decomposed into two components, the first of which is random measurement error variance. This variance is produced by nonsystematic factors such as brief fluctuations in mood and motivation, momentary inattention, uncontrolled administration conditions (e.g., noise, distraction), illness, fatigue, emotional strain, or chance. The second component is systematic variance, which will be further divided into true variance and variance due to systematic error that reflects a common response bias among judges. Response bias creates problems because its contribution to systematic variance serves to inflate the estimate of interrater reliability (cf. Guion, 1965). For example, if a common tendency exists among judges to select socially desirable response alternatives rather than response alternatives reflective of true judgments, then the appearance of high interrater reliability is likely indicative of response bias rather than true agreement among the judges.

This suggests that an estimate of interrater reliability should not only assess the proportion of total variance in a set of judgments that is systematic variance, but also effect controls for that portion of the systematic variance that is due to response bias. Unfortunately, the few attempts to propose methods for estimating interrater reliabilty for a design in which one set of judges has rated a single target fall short of accomplishing these objectives. A procedure presented by Howe (1977) poses problems because errors associated with lack of agreement at a particular time are confounded with errors associated with instability of judgments over time. Procedures recommended by Finn(1970) and James, Wolf, and Demaree (1981) are based on the classic psychometric model, which partitions total variance into true variance and random measurement error variance. The result is that variance due to response bias is regarded as true variance. Thus, estimates of interrater reliability will be spuriously inflated if a response bias is included in the measurements. However, a review of the procedures introduced by Finn (1970) and by James et al. (1981) is instructive because it furnishes a rationale for estimating interrater reliability for the design of interest. These procedures are reviewed below. Included in this overview is a comparison between the Finn and James et al. procedures and other methods that have been commonly used to estimate agreement, consistency, or interrater reliability. Attention is then given to procedures to effect controls for response bias.

METHODS FOR ESTIMATING WITHIN-GROUP INTERRATER RELIABILITY

Overview of Prior Methods

Review of the Finn (1970) and James et al. (1981) procedures begins with the estimator for judgments on a single item and proceeds to that for mean judgments on $\underline{J}$ items. The $\underline{K}$ judges on whom observations are available are viewed as a group in the statistical sense, and thus the estimators are referred to as "within-group interrater reliability coefficients."

Single item estimator. For a single item, designated $\underline{X_j}$, within-group interrater reliability, or simply IRR, was viewed as a function of two variances. These were (a) the observed variance of the scores furnished by the $\underline{K}$ judges on $\underline{X_j}$, or

$s^2_{x_j}$, and (b) the variance on X_j that would be expected in a condition of an IRR of zero. Finn and James et al. assumed that an IRR of zero would occur when all judgments on X_j were due exclusively to random measurement errors. A mathematical definition of "random" (cf. Brunk, 1965) applied to the present design denotes that each alternative on the measurement scale of X_j has an equal likelihood of response and, therefore, that the judgments would be distributed uniformly (i.e., a rectangular distribution). Consequently, James et al. proposed that the variance expected on X_j when judgments are theoretically due exclusively to random measurement errors could be calculated using the equation for the variance of a rectangular or uniform distribution. This equation is: $\sigma_{EU}{}^2 = (A^2-1)/12$ (Mood, Graybill, & Boes, 1974). The subscript "EU" refers to an expected error (E) variance based on a uniform (U) distribution, and A corresponds to the number of alternatives in the response scale for X_j, which is presumed to vary from 1 to A.

The equation for $\sigma_{EU}{}^2$ includes the assumption that the psychological response scale (cf. Guilford, 1954) underlying the measurement scale of X_j is discrete (as well as approximately interval). That is, the underlying psychological response scale for an item is assumed to be composed of a finite set of sequentially ordered, countable categories. Support for this assumption is furnished by psychophysical scaling studies, which have shown that psychological judgments of continuously distributed, physical stimuli tend to involve only a limited number of sequentially ordered categories -- specifically seven categories, plus or minus two (Miller. 1956; Parducci, 1965). Support is also provided by performance evaluation studies where Landy and Farr (1980) concluded that the number of response categories efficiently used in rating scales follows the Miller (1956) "dictum" of seven, plus or minus two categories. Studies of cognitive style and structure furnish additional support for an assumption of discrete response scales. These studies have shown that unidimensional cognitive schemas (attributes, dimensions) are comprised of a modest number of categories, which is a function of limited cognitive information processing capabilities, simplification strategies, and bounded rationality (cf. Dawes, 1976; Goldstein & Blackman, 1978; Scott, Osgood, & Peterson, 1979; Slovic, Fischhoff, & Lichtenstein, 1976; Struefert & Struefert, 1978; Wyer, 1974).

We do not wish to imply that all psychological response scales are discrete or that all judges are equally articulated on a particular discrete, psychological response scale. Nevertheless, the evidence reviewed above suggests that an assumption of discrete scales with seven plus or minus two categories is applicable to many judgment tasks for most judges. Thus, an assumption of discrete response scales was employed to develop estimating procedures. Later in this article it will be shown that changes in the computation of expected variances makes possible the use of the estimating procedures when psychological response scales are assumed to be continuous rather than discrete.

Given $s^2_{x_j}$ and an $\sigma_{EU}{}^2$ based on a discrete response scale, an estimate of IRR may be derived as follows (James et al., 1981). An observed score on item X_j, designated X_{jk} ($k=1, \ldots, K$ judges), may be represented as $X_{jk} = \mu_j + (\bar{X}_j - \mu_j) + e_{jk}$. In this equation, μ_j is the population mean (true score) on the item, $\bar{X}_j$ is the sample mean, and e_{jk} is a random error of measurement. If the X_{jk} are reflective solely of μ_j, then the X_{jk} are entirely devoid of random error variance and $s^2_{x_j} = 0$. This is because variance in the X_{jk} arises only from variation in the e_{jk}, and thus $s^2_{x_j}$ estimates random error variance. On the other hand, if the X_{jk} are a function of random error exclusively and conform to equal likelihood, random responses, then

$s^2_{x_j} = \sigma_{EU}^2$. This suggests that the extent to which the X_{jk} are actually reflective of μ_j, and may be said to reveal nonerror or true variance, is indicated by $(\sigma_{EU}^2 - s^2_{x_j})$.

An IRR estimate is obtained by placing the estimates of the variances in the equation: (true variance)/(true variance + error variance), or $(\sigma_{EU}^2 - s^2_{x_j})/ (\sigma_{EU}^2 - s^2_{x_j}) + s^2_{x_j} = (\sigma_{EU}^2 - s^2_{x_j})/\sigma_{EU}^2$. This equation reduces to the equation proposed by Finn (1970), namely $1 - (s^2_{x_j}/\sigma_{EU}^2)$. The term $(s^2_{x_j}/\sigma_{EU}^2)$ estimates the "proportion of random or error variance present in the observed ratings", and $1 - (s^2_{x_j}/\sigma_{EU}^2)$ "gives the proportion of non-error variance in the ratings, a reliability coefficient" (Finn, 1970, p. 72).

The discussion above is summarized by the following equation:

$$r_{WG(1)} = 1 - (s^2_{x_j}/\sigma_{EU}^2) \qquad (1)$$

where $r_{WG(1)}$ is the within-group interrater reliability for a group of K judges on a single item X_j, $s^2_{x_j}$ is the observed variance on X_j, and σ_{EU}^2 is the variance on X_j that would be expected if all judgments were due exclusively to random measurement error.

Use of Eq. 1 is illustrated in Table 1 for $K = 10$ and $A = 5$, 7, and 9, respectively. The 10 judgments in the $A = 5$ column are distributed approximately uniformly. Thus, we would expect $r_{WG(1)}$ to be approximately zero, which is indeed the case [i.e., $r_{WG(1)} = 1 - (1.73/2.0) = .13$; $\sigma_{EU}^2 = (5^2-1)/12 = 2.0$]. The $A = 7$ column involves judgments clustering about the upper end of the scale, which suggests a high interrater reliability. This is also the case; $r_{WG(1)} = .94$. Finally, judgments in the

A = 9 column cluster about the theoretical midpoint of a 9-point scale, which again suggests a high interrater reliability. The interrater reliability for this data set is .92.

Insert Table 1 about here

Multiple item estimator. The estimate of IRR for judges' mean scores is based on the assumption that the J items (j=1,...,J) are "essentially parallel" indicators of the same construct. This implies that the variances of, and covariances among, the items are approximately equal, respectively, in the underlying domain of items. Given these assumptions, it is possible to estimate IRR among judges' mean scores by applying the Spearman-Brown prophecy formula to Eq. 1 (Finn, 1970; James et al., 1981). James et al. derived several estimating equations, the most direct for computing purposes is as follows:

$$r_{WG(J)} = \frac{J[1 - (\overline{s^2_{x_j}}/\sigma_{EU}^2)]}{J[1 - (\overline{s^2_{x_j}}/\sigma_{EU}^2)] + (\overline{s^2_{x_j}}/\sigma_{EU}^2)} \qquad (2)$$

where $r_{WG(J)}$ is the within-group interrater reliability for judges' mean scores based on J essentially parallel items, $\overline{s^2_{x_j}}$ is the mean of the observed variances on the J items, and σ_{EU}^2 has the same definition as before.

Use of Eq. 2 is illustrated in Table 2. Section A of Table 2 presents simulated data for K = 10 judges who have rated the same target on J = 6 essentially parallel items. All items employ the same discrete, approximately interval five-point scales (i.e., A = 5) and, after reflection if necessary, are scored in the same direction. Cursory examination of the data suggests a high IRR. The estimate furnished by $r_{WG(J)}$, shown in Section B of Table 2, supports this observation. That is,

$\underline{r_{WG(6)}}$ = .97, which denotes a high level of IRR among the judges' means shown at the bottom of the data matrix in Section A. $\underline{r_{WG(J)}}$ will generally be larger than $\underline{r_{WG(1)}}$ because averaging over essentially parallel items reduces the influence of measurement error (cf. Lord & Novick, 1968).

Insert Table 2 about here

Comparison of $\underline{r_{WG(J)}}$ with commonly used methods. Methods commonly used to estimate IRR, agreement, or consistency of judgments (cf. Mitchell, 1979; Shrout & Fleiss, 1979) should not be employed in our present design. To see why this is the case, representatives of these methods and their associated estimates for the data in Section A of Table 2 are presented in Section C of this same table. Whereas the data indicate obvious agreement or consistency, the estimates based on these methods are either low, indeed negative in one case, or uninterpretable (i.e., $\underline{\bar{D}}$ = 1.76). The low values for the intraclass correlations, mean percentage of agreement, and average intercorrelation among judges' profiles occur for multiple reasons. Nevertheless, in one form or another the most salient reason can be traced to the simple fact that the item data in Section A suffer a severe restriction of range, which is quite likely to occur if judges in a single group agree on responses to essentially parallel items.

To expand briefly on this point, the intraclass correlations could not assume high values unless the between-item mean squares were large in relation to the within-item mean squares (cf. Lahey, Downey, & Saal, 1983). Large between-item mean squares were not obtained because the means on the essentially parallel items were almost identical. Mean percentage of agreement was low because this statistic is "insensitive to degrees of agreement, that is, it treats agreement as an all-ornone phenomenon, with no room for partial or incomplete agreement" (Mitchell, 1979, p. 377). In the present illustration, scores of "4" versus "5" in the item data were treated as disagreements when in fact they reflected at least "partial agreement" on a response scale with five alternatives. The mean correlation of -.11 between judges' profiles was due to the small and unreliable variations in shapes among judges' profiles on the items, as well as minimal variance in the judgment matrix (Lahey et al., 1983). Finally, the mean Euclidean distance value of 1.76 furnished no direct basis for inferring level of agreement or estimating interrater reliability. This statistic could be used for other purposes, such as relating comparative levels of agreement to some other variable when data are collected from different groups of judges. However, we believe that most investigators will find it informative to have an estimate of interrater reliability for a group, or, as discussed later, for each one of a set of groups if the level of agreement varies among groups.

Other methods that have been used to estimate agreement or consistency among judgments include r_p (Cattell, 1949), Mahalanobis D (Cronbach & Gleser, 1953; Overall, 1964), and r_c (Cohen, 1969). These methods were not included in Table 2 for the following reasons. Mahalanobis D and r_p require that profile elements (items in this case) be sampled from orthogonal latent variables. This is clearly not applicable because the items in Table 2 are assumed to be essentially parallel measures of a single latent variable or construct. A key problem with r_c is that it considers the direction of scoring of items to be arbitrary. This is often not the case. Furthermore, r_c is a function of the correlations among judges' profiles and thus would be subject to the same problems discussed above for the mean correlation among judges' profiles.

Methods Associated with Response Bias in Expected and Observed Distributions

Several authors (and an anonymous reviewer) have argued that the expected distribution of responses might be nonuniform when no true agreement exists among judges (Hsu, 1979; Selvage, 1976). This suggests the need to consider systematic errors or bias in the expected distribution; that is, systematic factors which render the expected distribution nonuniform when the true IRR is zero. The likely candidates for systematic bias are response sets and response styles, such as central tendency, leniency, and social desirability (cf. Cronbach, 1946, 1950; Guilford, 1954; Guion, 1965; Nunnally, 1978). Response sets and styles are typically regarded as personal characteristics. Our concern here, however, is the systematic effects that response sets and styles have on expected and observed distributions for a group of $\underline{K}$ judges. Consequently, we employ the term "response bias" to indicate systematic biasing of a response distribution due to a common response set or style within a group of judges.

Situations in which different judges employ different response sets or styles are not addressed in this initial effort to consider the effects of systematic bias in judgments. The generalizability of the discussion to follow is therefore limited. Nevertheless, prior research suggests that response bias evolving from a common response set or style among judges is an important issue with the potential for frequent occurrence. Examples include a common tendency for individuals to select socially desirable response alternatives on questionnaires designed to measure affective variables such as anxiety (cf. Nunnally, 1978), for subordinates to exhibit positive leniency when describing supervisors (Schriesheim, 1981), and for judges to select neutral response alternatives when items are ambiguous or when the judges wish to be evasive (cf. Guilford, 1954; Guion, 1965).

To visualize the problem caused by response bias, consider a condition in which there is no true variance in a set of judgments on a single item. Based on Eq. 1, this implies that (a) the distribution of observed judgments should be approximately uniform, (b) $s_{x_j}^2 \cong \sigma_{EU}^2$, and (c) the proportion of true variance in the judgments, given by $r_{WG(1)}$, should be approximately zero. Suppose, however, that a response bias generated by social desirability resulted in a unimodal, negative skew in the observed distribution of judgments. Since $s_{x_j}^2$ (skewed) is by necessity less than $s_{x_j}^2$ (uniform), it follows that $r_{WG(1)}$ based on $s_{x_j}^2$ (skewed) will be positive. Theoretically, what we have done by estimating the proportion of true variance in the judgments by $r_{WG(1)} = 1 - [s_{x_j}^2 \text{ (skewed)}/\sigma_{EU}^2]$ is to regard all systematic sources of variance as true variance. In other words, the clustering of responses in the socially desirable tail of the observed distribution is falsely assumed to be reflective of true agreement when in fact this clustering of judgments is due to systematic bias evolving from social desirability. The result of this false assumption is that $r_{WG(1)}$, and by implication $r_{WG(J)}$, will furnish spuriously high estimates of IRR when judgments are affected by a response bias that causes observed responses to cluster about a subset of response alternatives on a scale.

Spurious inflation of all forms of reliability due to response bias has been addressed at the theoretical level (cf. Guilford, 1954; Guion, 1965; Wherry & Bartlett, 1982), but almost ever considered at the empirical level. We shall attempt to bridge this gap between psychometric theory and data for Eqs. 1 and 2 by proposing three steps that are designed to control for the spurious influences of response bias on estimates of $r_{WG(1)}$ and $r_{WG(J)}$.

Step 1

Ask the question: If there is no true variance in the judgments and the true IRR is zero, then what form of distribution would be *expected* to result from response bias, and, of course, some random measurement error? This distribution reflects one's hypothesis about response bias and is referred to as the "null distribution". (If no systematic bias is expected, then the null distribution is uniform.) The hypothesis may reflect application of knowledge from prior research to the present context. For example, it may be reasonable to expect a skewed null distribution in interest inventories as a result of a conscious or unconscious bias to present oneself in a favorable manner (cf. Guion, 1965; Nunnally, 1978). Performance evaluations by supervisors might be expected to be skewed as a result of such things as a bias that leaders should be supportive of subordinates, or a bias motivated by the need to avoid attributions by others that poor subordinate performance is due to inadequate leadership (cf. Guilford, 1954; James & White, in press; Landy & Farr, 1980). The hypothesis may also reflect knowledge that conditions conducive to a common response set or style were inherent in the research design [see extensive discussions by Cronbach (1946, 1950) and Guilford (1954)]. To illustrate, a triangular distribution that reflects a central tendency response bias (on a discrete scale) might be expected if inexperienced or indifferent judges with low articulation and little or no training are asked to respond to complex and/or ambiguous items (Guilford, 1954; Scott et al., 1979). Central tendency and a triangular distribution might also result from other types of common biases, such as when judges are purposefully cautious or evasive because responses to items are not collected on a confidential basis and political reasons exist for not departing from the neutral alternatives on the scales (cf. Guion, 1965).

We will focus on the null distributions presumably generated by central tendency (triangular distribution) and social desirability and positive leniency (skewed distributions) because these are the sources of bias that appear to have been empirically supported in research (cf. Anastasi, 1982; Landy & Farr, 1980; Nunnally, 1978; Wherry & Bartlett, 1982). Discussions concerning skewed distributions apply to positively or negatively skewed distributions, although negatively skewed distributions will be used in illustrations. Response sets or styles such as acquiescence, criticalness, deviant-response tendencies, and extreme response tendencies are not considered here because there is little evidence of their generalizability over instruments or ability to account for more than small fractions of variance (cf. Nunnally, 1978; Rorer, 1965). Other response sets or styles such as impulsion can be neutralized with good measurement techniques (Guilford, 1954). Correction for guessing is not addressed because our interest focuses on items for which there is no correct response (Rorer, 1965). Finally, halo and implicit theories regarding covariation among variables that presumably measure different constructs are beyond the confines of our single variable design.

Empirical evidence should also be used to propose a null distribution. This issue is more meaningfully treated in a later context. For the present, this first step is considered satisfied when the investigator hypothesizes a null distribution that corresponds to the response bias expected as a result of central tendency, social desirability, or positive leniency.

Step 2

Given a null distribution, the next step is to derive an expected variance, or "*EV*". This is the variance expected when all *systematic* variance is due to response bias. Calculation of *EVs* for triangular and skewed distributions is illustrated for five-point scales, where we again presume discrete, approximately equally spaced, psychological response scales.

Triangular null distribution. The expected proportions of judgments for a five-point, discrete scale (i.e., $A = 5$) in a triangular distribution are: 1 = .11, 2 = .22, 3 = .33, 4 = .22, 5 = .11 (Messick, 1982). Equations for *EVs* for the null distribution are (Messick, 1982, Eq. 4):

$$\sigma_{ET}^2 = \begin{cases} \dfrac{(A-1)(A+3)}{24} & \text{for } A \text{ odd, and} \\[2ex] \dfrac{A^2 + 2A - 2}{24} & \text{for } A \text{ even,} \end{cases}$$

where σ_{ET}^2 denotes the *EV* for a triangular null distribution. For $A = 5$, the value of σ_{ET}^2 is 1.33

Negatively skewed null distribution. Three distributions were arbitrarily selected to simulate different as well as representative degrees of skew. The first distribution possesses a small skew, the second a moderate skew, and the third a large skew. The *EV* for the distribution possessing a small skew is designated by σ_{ESS}^2. The *EV* for the distribution with a moderate skew is denoted by σ_{EMS}^2, while that from the large skew is σ_{ELS}^2. For the small skew condition, the proportions of judgments on a five-point, discrete scale were set as follows: 1 = .05, 2 = .15, 3 = .20, 4 = .35, and 5 = .25. The expected value of this null distribution is 3.6, and σ_{ESS}^2 is 1.34. The expected value and the expected variance were determined by employing definitional equations for discrete random variables (cf. Mood et al., 1974). For example, the expected value or $E(X)$ is equal to $\Sigma_{i} a_i p_i$, where a_i is a scale value and p_i is the probability of occurrence of that scale value (i.e., the proportions assigned above). Thus, $E(X) = 1(.05) + \ldots + 5\,(.25) = 3.6$. The expected variance is given by $E([X-E(X)]^2)$, which is equal to: $(1-3.6)^2(.05) + \ldots + (5-3.6)^2(.25) = 1.34$. The proportions of judgments for the moderate skew condition were set at: 1 = 0, 2 = .10, 3 = .15, 4 = .40, and 5 = .35. The expected value is 4.0, and σ_{EMS}^2 is .90. Finally, the proportions for the large skew condition were set at: 1 = 0, 2 = 0, 3 = .10, 4 = .40, and 5 = .50. The expected value is 4.4 and σ_{ELS}^2 is .44.

There are, of course, an infinite number of negatively skewed distributions. Similarly, a central tendency response bias does not connote a perfectly triangular distribution. However, these null distributions are representative and furnish a basis for demonstrating the principles involved in estimating $\underline{r_{WG(1)}}$ and $\underline{r_{WG(J)}}$ in the presence of response bias.

Step 3

This step consists of (a) replacing $\underline{\sigma_{EU}}^2$ in Eqs. 1 or 2 with the EV for the proposed null distribution, and (b) using the values furnished by Eqs. 1 or 2 as estimates of within-group interrater reliability ($s_{x_j}^2$ or $\overline{s_{x_j}^2}$ remain the same). To illustrate, consider the data in Section A of Table 3, which simulate the observed judgments of 10 judges on 6 essentially parallel items. Inspection of these data indicates a clustering of judgments about the scale midpoint of "3". If it is believed that these data reflect a central tendency response bias, and that the null distribution is approximately triangular, then $\underline{\sigma_{ET}}^2$ should be used in place of $\underline{\sigma_{EU}}^2$ in Eqs. 1 and 2. As shown in Section B of Table 3, the resulting estimate of $\underline{r_{WG(1)}}$ is .50 for each item, and $\underline{r_{WG(6)}}$ is .86.

Insert Table 3 about here

The logic of this approach is that if $\underline{\sigma_{ET}}^2$ is an accurate reflection of the null distribution, then $s_{x_j}^2$ (or $\overline{s_{x_j}^2}$) $\cong \underline{\sigma_{ET}}^2$ implies that all systematic variance in the judgments is a function of a central tendency response bias. Consequently, $\underline{r_{WG(1)}}$ (or $\underline{r_{WG(J)}}$) $\cong$ 0 because there is no true variance in the judgments. However, if $s_{x_j}^2$ (or $\overline{s_{x_j}^2}$) $< \underline{\sigma_{ET}}^2$, then we infer that not all of the systematic variance in the judgments is due to response bias. Rather, at the item level, $(s_{x_j}^2/\underline{\sigma_{ET}}^2)$ reflects the proportion of variance that is attributable to systematic response bias and random error. This is (.67/1.33) or .50 for each item in Table 3.

Accordingly, $r_{WG(1)}$ and $r_{WG(J)}$ estimate the proportion of variance that is systematic and nonbiased, which is to say the proportion of true variance in the single item judgments and mean judgments. Thus, although $r_{WG(1)} = .50$ is modest, $r_{WG(6)} = .86$ suggests that a substantial proportion of the variance in the mean judgments in Table 3 is true variance.

Applications and logic for skewed null distributions follow a similar course. The values for σ_{ESS}^2, σ_{EMS}^2, and σ_{ELS}^2 from Step 2 were used in Eqs. 1 and 2 to estimate IRRs for the data shown in Section A of Table 2. The resulting estimates are reported in Table 4. Of initial importance is the point that the observed data in Table 2 reflect a severe negative skew. If it is believed that this negative skew is at least partially a function of a social desirability or positive leniency response bias, then one might argue for a null distribution involving a large skew. Accordingly, if σ_{ELS}^2 were set equal to .44, then the estimate of the proportion of systematic variance that is true variance in item 1 is given by $r_{WG(1)} = 1 - (s_{x_j}^2/\sigma_{ELS}^2) = 1 - (.28/.44) = .36$. However, $r_{WG(J)} = .78$, a result of the fact that the modest amount of true variance at the item level is emphasized for judges' means over six items.

Insert Table 4 about here

One might also believe that while a social desirability or positive leniency response bias is a possibility, the bias is not large. This suggests that the null distribution should reflect a moderate or even a small skew. In these cases, the proportions of observed variance attributable to response bias and random error are substantially reduced at the item level in comparison to σ_{ELS}^2 (i.e., $s_{x_j}^2/\sigma_{EMS}^2 =$

.28/.90 = .31; and $s_{x_j}^2/\sigma_{ESS}^2$ = .28/1.34 = .21). It follows that estimates of $r_{WG(1)}$ and $r_{WG(J)}$ will be comparatively higher. As shown in Table 4, this is the case.

In summary, estimates of IRRs in conditions of response bias are obtained by calculating an *EV* for a null distribution that reflects the influence of the presumed response bias, and employing that *EV* in Eqs. 1 and 2. We proceed now to the issue that the accuracy of the IRR estimate is dependent on the accuracy of the null distribution and the *EV*, and therein is likely to be a problem.

Multiple Null Distributions

A key issue in attempting to hypothesize a null distribution is that true scores tend to be confounded with systematic errors (Guilford, 1954; Nunnally, 1978). As a case in point, Guilford (1954, p. 451) reported that a clustering of observed judgments about the middle alternatives on a response scale may reflect a "genuinely moderate amount of the trait [target in our terms] indicated by the items" rather than a central tendency. For example, consider the data in Table 3 that were used to illustrate a central tendency response bias. An equally plausible explanation for these data is that μ_j in the equation $X_{jk} = \mu_j + (\bar{X}_j - \mu_j) + e_{jk}$ is equal to "3"; that is, the true score is 3. This connotes that all systematic variance is true variance, and that all variation in the observed judgments is due to random influences (i.e., the e_{jk}). These assumptions imply that the null distribution is the uniform distribution. Thus, $\sigma_{EU}^2 = 2.0$ should be used in Eqs. 1 and 2, and, as shown in Section C of Table 3, the estimates of $r_{WG(1)}$ and $r_{WG(6)}$ should be .66 and .92, respectively.

The same possibility applies for skewed distributions. The data in Table 2 might indicate that the high ratings by multiple judges of a department head who is being considered for a vice-presidency reflect the selection, training, and experience of the department head--that is, true ability--rather than positive leniency on the part of the judges (cf. Bernardin & Pence, 1980; Borman, 1979). If this is the case, then the uniform null distribution should be used for estimation purposes. This null would replace the nulls that were based on skews and employed to obtain the IRR estimates reported in Table 4. Estimates of $r_{WG(1)}$ and $r_{WG(J)}$ based on σ_{EU}^2 are reported at the bottom of Table 4 for comparison purposes.

In short, a particular observed distribution may be consistent with any number of null distributions, including the uniform distribution and nulls involving response bias. It follows that the shape of the observed distribution should *not* in general be employed to propose a null distribution. Rather, evidence other than the observed distribution should be used to propose a null, after which the proposed null can be compared to the observed distribution and a subjective goodness of fit test conducted. The question here is whether the proposed null could have generated the observed distribution. For example, if a null distribution with a large skew and $\sigma_{ELS}^2 = .44$ were hypothesized, but the observed distribution has a moderate skew with $s_{x_j}^2 = .90$, then this particular null distribution has been disconfirmed. Allowance should be made for sampling error, where, for example, an obtained $s_{x_j}^2$ of .46 could well be consistent with an σ_{ELS}^2 of .44.

This procedure is designed to reduce the number of viable null distributions. However, as discussed in greater detail shortly, it is still likely that multiple null distributions will be consistent with a specific observed distribution. This underscores the need to obtain evidence other than the observed distribution to propose nulls. This other evidence consists of the aforementioned use of knowledge from prior research. A stronger base for hypothesizing a null is furnished by obtaining empirical data on the judges, target, item(s), and/or judgment context at hand. The many possibilities that exist here are reviewed in numerous articles and texts. They include (a) comparisons between the judgments of concern and objective indicators of a target, (b) comparisons among judgments from the same judges on diverse scales, (c) use of "lie scales" or "faking keys" to identify judges who presumably are not answering frankly or honestly, (d) having judges respond to the same items in honest versus realistic conditions, (e) comparisons between the responses of the judges and those of independent observers or "expert judges", and (f) having judges respond to amiguous items, or simply to response scales for which there are no items (cf. Anastasi, 1982; Berg & Rapaport, 1954; Borman, 1978; Cronbach, 1946, 1950; Damarin & Messick, 1965; Guilford, 1954; Guion, 1965; Nunnally, 1978; Rorer, 1965; Schriesheim, 1981; van Heerden & Hoogstraten, 1979). Rater training is another possibility, although there is the potential problem of introducing bias into judgments, as in the case of introducing a central tendency bias by training designed to avoid positive leniency (Bernardin & Pence, 1980). Finally, the advent of latent trait theory (cf. Guion & Ironson, 1983; Hulin, Drasgow, & Parsons, 1983; Lord, 1980) and latent variable structural analysis (cf. James, Mulaik, & Brett, 1982; Jöreskog & Sörbom, 1979) may furnish novel and productive approaches for dealing with response bias.

It is reasonable to expect that the stronger the empirical base, the more informed the decision-making regarding nulls. Nevertheless, it is also the case that, at the present time, no method or set of methods furnishes an infallible basis for completely unraveling the confounding between true scores and response bias (Nunnally, 1978). This implies that while empirical efforts may substantially reduce the list of possible nulls, the end-product of these efforts will likely be the identification of several possible nulls, including, perhaps often, the uniform distribution. Moreover, one must consider the likelihood that even though a strong case has been built for response bias, a number of viable alternatives remain for the shape of the null distribution (e.g., various degrees of skew).

To deal with these issues, we suggest the following additions to the three steps described earlier for estimating IRR in conditions of response bias. These suggestions are designed to obtain a range of IRR estimates within which the true IRR is likely to fall.

Suggested Estimation Procedures When the Null Distribution May Assume Multiple Forms

First, in Step 1 gather as much pertinent information regarding null distributions as possible, including empirical data designed to identify response bias for the judges in the sample at hand. The recommendation is based on the logic that if a fallible decision is unavoidable, then let us attempt to reduce the degree of fallibility as much as possible by considering multiple sources of information.

Second, use this information to propose a small but inclusive *set* of null distributions that represent the major forms of anticipated response bias. For example, one might anticipate a small to moderate positive leniency response bias, in which case a small and a moderate negative skew might be proposed. Or, a moderate to large anticipated central tendency might involve the triangular distribution as well as a more peaked distribution with higher proportions of judges in the middle response alternatives. If the uniform distribution is a possibility, then it too should be included in the set. The rationale here is that even though we cannot pinpoint a particular null with a high degree of confidence, we can place bounds on the most likely types of nulls and thereby increase the likelihood that the true null likes somewhere in this range of distributions.

Third, proceed to Step 2 and identify the smallest and largest *EV*s for the set of null distributions. Fourth, go to Step 3 and compute estimates of $\underline{r_{WG(1)}}$ and/or $\underline{r_{WG(J)}}$ for the smallest and largest *EV*s. These estimates represent the range in which the true interrater reliability is believed to occur. More sophisticated statistical systems involving families of null distributions, weighting of _EVs_ or reliability estimates to reflect differential likelihoods of the value of the true interrater reliability, and computation of weighted and unweighted averages of _EVs_ and reliability estimates are possible and await future embellishments on the procedures suggested here.

The range of estimates for $r_{WG(1)}$ may or may not vary considerably as a function of the difference in the magnitudes of the smallest and largest _EVs_. A different and somewhat fortunate condition exists for $r_{WG(J)}$. In general, the range between the estimates of $r_{WG(J)}$ will decrease as _J_ increases. A dramatic example of this condition is shown in Table 4, where the "50 point" range for estimates of $r_{WG(1)}$ (.36 to .86) is reduced to 19 points for $r_{WG(6)}$ (.78 to .97). This generalization is contingent on there being at least a modest amount of true variance in the judgments (i.e., $s^2_{x_j}$ is less than the smallest _EV_). Inasmuch as we would often expect this to be the case, it is possible to state that ambiguity resulting from discrepencies in the highest and lowest estimates of $r_{WG(J)}$ will decrease as _J_ increases. Indeed, if the difference between the smallest and largest _EVs_ is itself modest, then the estimates will converge rapidly as _J_ increases to only four or five items. Thus, investigators are encouraged to employ mult_ple items and $r_{WG(J)}$ rathe_ than $r_{WG(1)}$. There are, of course, numerous psychometric reasons for encouraging the use of multiple indicators of a construct, not the least of which is reduction of the influence of random measurement errors.

DISCUSSION

The estimating procedures for within-group interrater reliability have the potential for frequent usage inasmuch as many real-world problems involve judgments of a single target on a single variable by one group of judges. Examples presented earlier included the need to estimate IRR for anchors in BARS development, publishability of a particular manuscript, faculty evaluations from a particular class, team cooperativeness for a specific team, and judgments of the ability of a department head to assume a vice-presidency. The list of additional examples is lengthy and involves such things as estimating IRRs for job incumbents' perceptions of job characteristics prior to implementing a job enrichment program, for ratings of an individual's potential in a management assessment center, and for city council members' judgments of the feasibility of implementing a new snow removal program.

It is important to note that we are not suggesting that the methods presented here compete with the traditional intraclass correlation (ICC) designs (cf. Shrout & Fleiss, 1979). These designs require that multiple targets be rated, either in a factorial design in which each target is rated by a different set of judges, or in a repeated measures design where each judge rates all (or most) targets. In contrast, the procedures for estimating interrater reliability presented here apply to situations in which only a single target is available. However, occasions may occur where $r_{WG(1)}$ and $r_{WG(J)}$ are useful in multiple target designs. An example is when violation of the homogeneity of within-target variance assumption precludes meaningful interpretation of an ICC. In this situation, it would be useful to obtain separate estimates of $r_{WG(1)}$ or $r_{WG(J)}$ for each group or target. This stimulates the additional point that the targets in organizational research are often "groups", such as workgroups, departments, or organizations (e.g., climate studies in which judges are nested in organizations--cf. James, 1982). Given heterogeneity of within-group variances, estimates of $r_{WG(1)}$ or $r_{WG(J)}$ could be calculated for each group and then used as measures of a group attribute in group level studies. To illustrate, it may be of interest to identify the conditions that contribute to high versus moderate to low within-group IRRs on a variable such as group cooperativeness.

Such uses, indeed all uses of $\underline{r_{WG(1)}}$ and $\underline{r_{WG(J)}}$, presuppose prior demonstrations of the construct validity of a variable, as stipulated at the first of this article. Alternatively, consideration of response bias and null distributions using the present methods might address construct validity issues not addressed in prior research. That is, if use of the proposed methods indicates that responses on a presumably "valid" variable may have been generated primarily by response bias rather than a true score on a target, then the construct validity of the variable requires further review. Also in need of review, due to benign neglect, is the effect of response bias on other forms of reliability estimation (e.g., ICC, alpha, kappa, test-retest). Corrections for spurious inflation of reliability estimates resulting from systematic response bias (Guion, 1965) should not be limited to the methods proposed here.

This article is concluded with an overview of several statistical concerns. First, evidence reviewed earlier suggested that an assumption of discrete psychological response scales with seven plus or minus two categories is applicable to many judgment tasks for most judges. Nevertheless, an investigator may choose to employ the assumption that the underlying psychological response scale is continuous. This denotes that the $\underline{A}$ alternatives on a response scale are assumed to be "only representative of the possible values along the continuum" from 1 to $\underline{A}$ (Selvage, 1976, p. 606). Equations 1 and 2 are still employed to estimate IRR for variables with continuous response scales. However, the choice of assumption regarding continuous versus discrete response scales affects both the computation and magnitude of IRR estimates. Specifically, computing procedures for EVs (expected variances) under the "continuous assumption" differ from those described for the "discrete assumption". Furthermore, the values of the EVs based on a continuous assumption are less than the values of EVs based on a discrete assumption. The result is that IRR estimates based on a continuous assumption are generally lower than estimates based on a discrete assumption. For example, with *A* = 5, the *EV* for a discrete uniform distribution was presented earlier as $(A^2-1)/12 = 2.0$. The *EV* for a continuous uniform distribution is given by $(A-1)^2/12$, which is 1.33 (Mood et al., 1974; Selvage, 1976). If the $s^2_{x_j}$ for a single item variable is .50, then $r_{WG(1)}$ is .75 using the *EV* for the discrete uniform distribution and .62 using the *EV* for the continuous uniform distribution [i.e., 1 - (.50/1.33)]. Other comparisons of IRRs may be more or less dramatic (cf. Selvage, 1976). This issue is not pursued here given our belief that the discrete assumption is generally appropriate, although estimating procedures for other types of continuous null distributions have been developed and are available from the authors.

Second, like *ICCs*, Eqs. 1 and 2 furnish statistically biased estimates of IRR (Selvage, 1976), but the bias is expected to be minimal for a small number of judges and essentially negligible for a large number of judges (e.g., 10 or more). Third, the estimates of $r_{WG(1)}$ and $r_{WG(J)}$ will assume negative values whenever $s^2_{x_j}$ or $\overline{s^2_{x_j}}$ exceeds the value of an *EV*. As indicated earlier, such a condition disconfirms the null distribution on which the *EV* was based. In addition, if $s^2_{x_j}$ or $\overline{s^2_{x_j}}$ exceeds the *EV* for the uniform distribution (i.e., σ_{EU}^2), then the negative estimate of IRR should be set equal to .00 because variances greater than σ_{EU}^2 are associated with distributions that reflect serious degrees of disagreement. A similar recommendation was made for *ICCs* by Bartko (1976).

Fourth, and finally, special attention should be given to the fact that it is possible to manipulate estimates of $r_{WG(1)}$ and $r_{WG(J)}$ by constructing artificial and unrealistic measurement scales. Adding a spurious number of alternatives to a scale merely to inflate the size of an *EV*, and therefore the estimate of IRR, is poor research practice at best. Conversely, use of only a few alternatives (e.g., *A* = 3) in a scale when it is likely that the category-width or articulation on the psychological measurement scale contains a larger number of categories may result in artificially low estimates of IRR. In the final analysis, use of the methods described in this article rests on the assumption that measurement scales are sensitive to, and limited to, psychometrically reliable differentiation on psychological measurement scales (cf. Guilford, 1954). Valid scaling procedures in conjunction with professional and ethical judgment should satisfy this criterion.

References

Anastasi, A. *Psychological testing*. New York: MacMillan, 1982.

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

Berg, J.A., & Rapaport, G.M. Response bias in an unstructured questionnaire. *The Journal of Psychology*, 1954, *38*, 475-481.

Bernardin, H.J., & Pence, E.C. Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, 1980, *65*, 60-66.

Borman, W.C. Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology*, 1978, *63*, 135-144.

Borman, W. C. Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology,* 1979, *64*, 410-421.

Brunk, H.D. *An introduction to mathematical statistics.* Waltham, MA: Blaisdell Publishing Co., 1965.

Cattell, R. B. r_p and other coefficients of pattern similarity. *Psychometri' ,* 1949, *14*, 279-298.

Cohen, J. r_c: A profile similarity coefficient invariant over variable reflection. *Psychological Bulletin*, 1969, *71*, 281-284.

Cooper, M. An exact probability test for use with Likert-type scales. *Educational and Psychological Measurement*, 1976, *36*, 647-655.

Cronbach, L.J. Response sets and test validity. *Educational & Psychological Measurement*, 1946, *6*, 475-494.

Cronbach, L.J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 3-31.

Cronbach, L.J., & Gleser, G.C. Assessing similarity between profiles. Psychological Bulletin, 1953, 50, 456-473.

Damarin, F.L., & Messick, S.J. Response styles as personality variables: A theoretical integration of multivariate research. Research Bulletin, No. 65-10, Princeton, NJ: Educational Testing Service, 1965.

Dawes, R.M. Shallow psychology. In J.S. Carroll & J. W. Payne (Eds.), Cognition and social behavior. Lawrence Erlbaum: Hillsdale, NJ, 1976.

Finn, R. H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Goldstein, K.M., & Blackman, S. Cognitive style. New York: Wiley, 1978.

Guilford, J.P. Psychometric methods. New York: McGraw-Hill, 1954.

Guion, R.M. Personnel testing. New York: McGraw-Hill, 1965.

Guion, R.M., & Ironson, G.H. Latent trait theory for organizational research. Organizational Behavior and Human Performance, 1983, 31, 54-87.

Howe, J.G. Group climate: An exploratory analysis of construct validity. Organizational Behavior and Human Performance, 1977, 19, 106-125.

Hsu, L. Agreement or disagreement of a set of Likert-type ratings. Educational and Psychological Measurement, 1979, 39, 291-295.

Hulin, C.L., Drasgow, F., & Parsons, C.K. Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin, 1983.

James, L.R. Aggregation bias in estimates of perceptual agreement. Journal of Applied Psychology, 1982, 67, 219-229.

James, L.R., Mulaik, S.A., & Brett, J.M. Causal analysis: Assumptions, models, and data. Beverly Hills: Sage, 1982.

James, L.R. & White, J.F. Cross-situational specificity in managers' perceptions of subordinate performance, attributions, and leader behaviors. Personnel Psychology, in press.

James, L.R., Wolf, G., & Demaree, R.G. Estimating interrater reliability in incomplete designs. Fort Worth, TX: Institute of Behavioral Research, Texas Christian University, 1981.

Jöreskog, K.G., & Sörbom, D. Advances in factor analysis and structural equation models. Cambridge, MA: Abt Books, 1979.

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

Landy, F.J., & Farr, J.L. Performance ratings. Psychological Bulletin, 1980, 87, 72-107.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum, 1980.

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

Messick, D.M. Some cheap tricks for making inferences about distribution shapes from variances. Educational and Psychological Measurement, 1982, 42, 749-758.

Miller, G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. The Psychological Review, 1956, 63, 81-97.

Mitchell, S.K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin, 1979, 86, 376-390.

Mood, A.M., Graybill, F.A., & Boes, D.C. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1974.

Nunnally, J.C. *Psychometric theory*. New York: McGraw-Hill, 1978.

Overall, J.E. Note on multivariate methods for profile analysis. *Psychological Bulletin*, 1964, *61*, 195-198.

Parducci, A. Category judgment: A range-frequency model. *Psychological Review*, 1965, *72*, 407-418.

Rorer, L.G. The great response-style myth. *Psychological Bulletin*, 1965, *63*, 129-154.

Schriesheim, C.A. The effect of grouping or randomizing items on leniency response bias. *Educational and Psychological Measurement*, 1981, *41*, 401-411.

Scott, W.A., Osgood, D.W., & Peterson, C. *Cognitive structure: Theory and measurement of individual differences*. New York: Wiley, 1979.

Selvage, R. Comments on the analysis of variance strategy for computation of intraclass reliability. *Educational and Psychological Measurement*, 1976, *36*, 605-609.

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

Slovic, P., Fischhoff, B., & Lichtenstein, S. Cognitive processes and societal risk taking. In J.S. Carrol & J.W. Payne (Eds.), *Cognition and social behavior*. Hillsdale, NJ: Lawrence Erlbaum, 1976.

Streufert, S., & Streufert, S.C. *Behavior in the complex environment*. New York: Wiley, 1978.

van Heerden, J., & Hoogstraten, J. Response tendency in a questionnaire without questions. *Applied Psychological Measurement*, 1979, *3*, 117-121.

Wherry, R.J., & Bartlett, C.J. The control of bias in ratings: A theory or ratings. *Personnel Psychology*, 1982, *35*, 521-551.

Wyer, R. S. Cognitive organization and change: An information processing approach. Wiley: New York, 1974.

Footnote

Support for this project was provided under Office of Naval Research Contract N00014-80-C-0315, Office of Naval Research Project NR170-903. Opinions expressed are those of the authors and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy. The authors wish to thank David W. Bracken, Allan P. Jones, Benjamin Schneider, and S. B. Sells for their helpful suggestions and advice. Requests for reprints should be sent to Lawrence R. James, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332.

Table 1

Within-Group Interrater Reliabilities for Items

with Varying Numbers of Alternatives

	Judgments Furnished by 10 Judges		
	A = 5	*A* = 7	*A* = 9
	5	6	4
	2	6	4
	3	7	4
	5	7	5
	2	7	5
	3	7	5
	1	7	5
	4	6	5
	3	7	6
	4	7	6
Mean:	3.2	6.7	4.9
$s^2_{x_j}$:	1.73	.23	.54
σ^2_{EU}:	2.0	4.0	6.67
$r_{WG(1)}$:	.13	.94	.92

Table 2

Illustrations of a New Method and Commonly Used Methods for Computing Within-Group Interrater Reliability

A. Data

	Judge											
Item	1	2	3	4	5	6	7	8	9	10	Mean	$s_{x_j}^2$
1	5	4	5	4	5	4	5	4	5	4	4.5	.28
2	4	5	4	5	4	5	4	5	4	5	4.5	.28
3	5	5	5	5	5	4	4	4	4	4	4.5	.28
4	4	4	5	5	4	4	5	5	4	4	4.4	.27
5	5	4	4	5	5	5	4	5	4	5	4.6	.27
6	5	4	4	5	4	4	5	5	4	4	4.4	.27
Mean	4.7	4.3	4.5	4.8	4.5	4.3	4.5	4.7	4.2	4.3		.275

B. $r_{WG(6)}$

$$r_{WG(6)} = \frac{6\,[1 - (.275/2)]}{6[1 - (.275/2)] + (.275/2)}$$

$$= .97$$

C. Commonly Used Methods

1. Intraclass Correlations (Shrout & Fleiss, 1979):

   a. For a single rating [ICC(2,1)] = .09

   b. For mean item ratings [ICC(2,10)] = .58

2. Mean Percentage Agreement = 45%

3. Mean Correlation Between Judges' Profiles = -.11

4. Mean Euclidean Distance Between Judges' Profiles ($\bar{D}$) = 1.76

Table 3

Computation of $r_{WG(1)}$ and $r_{WG(J)}$ Based on Triangular and Uniform Null Distributions

A. Data

	Judge											
Item	1	2	3	4	5	6	7	8	9	10	Mean	$s_{x_j}^2$
1	3	2	4	3	3	4	2	3	4	2	3.0	.67
2	4	3	3	3	2	2	3	4	2	4	3.0	.67
3	3	3	4	2	2	4	2	4	3	3	3.0	.67
4	2	2	3	2	4	3	3	3	4	4	3.0	.67
5	4	4	4	2	2	2	3	3	3	3	3.0	.67
6	3	4	3	4	3	3	4	2	2	2	3.0	.67
Mean	3.17	3.0	3.5	2.67	2.67	3.0	2.83	3.17	3.00	3.00		

B. $r_{WG(1)}$ and $r_{WG(J)}$ Based on σ^2_{ET}

$r_{WG(1)}$ = 1 - (.67/1.33) = .50; for each item

$$r_{WG(6)} = \frac{6[1 - (.67/1.33)]}{6[1 - (.67/1.33)] + (.67/1.33)} = .86$$

C. $r_{WG(1)}$ and $r_{WG(J)}$ Based on σ^2_{EU}

$r_{WG(1)}$ = 1 - (.67/2) = .66; for each item

$$r_{WG(6)} = \frac{6[1 - (.67/2)]}{6[1 - .67/2)] + (.67/2)} = .92$$

Table 4

Estimates of $\underline{r_{WG(1)}}$ and $\underline{r_{WG(J)}}$ Based on Skewed and Uniform Null Distributions

Data From Table 2

	Item 1	Judges' Means
Null Distribution	$s^2_{x_j} = .28$	$s^2_{\bar{x}_j} = .275$
Large Skew	.36	.78
Moderate Skew	.69	.93
Small Skew	.79	.96
Uniform	.86	.97