INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/467

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MEASUREMENT ISSUES IN DETERMINING INTERRATER AGREEMENT

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

by

Andrea Harrell Vincent
Norman, Oklahoma
2002

UMI Number: 3045846



UMI Microform 3045846

Copyright 2002 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© Copyright by ANDREA HARRELL VINCENT 2002
All Rights Reserved.

MEASUREMENT ISSUES IN DETERMINING INTERRATER AGREEMENT

A Dissertation APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

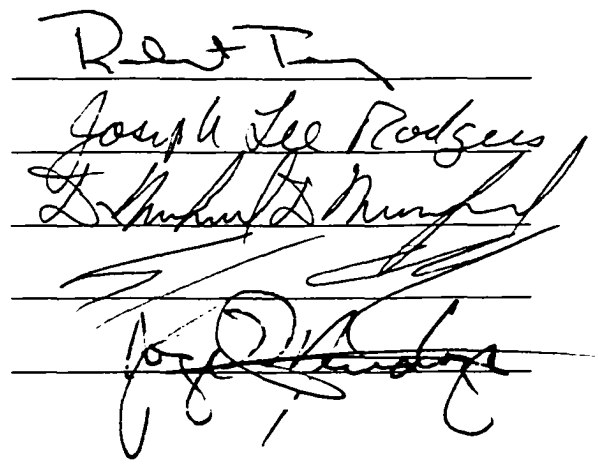


Table of Contents

Abstract	iv
Introduction	1
Kappa	5
Intraclass correlation	7
$r_{WG(J)}$	10
$r^{*}_{WG(J)}$	12
Relational agreement and r_s	13
Latent trait analysis and CSI	15
Summary	18
Study 1	19
Methods	
Data simulation	
Factors manipulated	
True distribution of θ	
Degree of true consensus	20
Slopes	
Number of raters and items	21
Thresholds	22
Estimation and Computation	
Evaluation of Indices	
Results	23
Study 2	39
Methods	
Analyses	40
Results and conclusions	40
Study 3	43
Methods	
Analyses	44
Results and conclusions	
General Discussion	45
References	51

Abstract

Multiple indices have been proposed claiming to measure the amount of agreement between ratings of two or more judges on a multi-item measure. Unfortunately, simulation work based on these indices is lacking; thus we are left with very little understanding of exactly what should be expected of these indices and when they should work. The present investigation seeks to bridge this gap in the literature by comparing several of the more commonly used measures of interrater agreement via an Item Response Theory (IRT) model. The goal is to identify which agreement indices best recover true agreement.

In this manuscript, several agreement indices are compared. Among these are the kappa coefficient κ_m (Fleiss, 1971); the intraclass correlation, *ICC(2, 1)* (Shrout & Fleiss, 1979); several variants of the $r_{WG(J)}$ index (James, Demaree, & Wolf, 1984; Lindell, Brandt, & Whitney, 1999); a measure of agreement for ordinal data (Stine, 1989); and an index derived from a Latent Trait Model (Terry, 2000). Results identify two measures of agreement that consistently recover true agreement. Implications and extensions to the measurement of agreement in multiple contexts are addressed.

Measurement Issues in Determining Interrater Agreement

The field of interpersonal perception research has flourished over the last two decades. Aspects of this research include the elaboration of models describing parameters important to the development of consensus/agreement (e.g., Albright, Kenny, & Malloy, 1988; Kenny, 1991; John & Robins, 1993; Park, Kraus, & Ryan, 1997), accuracy of judgments (e.g., Blackman & Funder, 1998; Funder & Colvin, 1988; Funder, Kolar, & Blackman, 1995; Jussim, 1991), and the measurement of agreement (Janson & Olsson, 2001; Shrout, 1993; Shrout & Fleiss, 1979; Zegers, 1991; Zwick, 1988). Of particular interest to the current study is the measurement of agreement.

The measurement of interrater agreement, while conceptually simple, can be problematic – especially when multiple raters are involved. A variety of approaches have been suggested for use in measuring the degree of agreement between 2 or more raters on a single target variable. Kenny, Albright, Malloy, & Kashy (1994) reviewed several of these approaches which include discrepancy measures, correlational measures, and variance measures (see also Goodwin, 2001; Conway & Schaller, 1998).

Unfortunately, very little simulation work has accompanied the various agreement indices stemming from the aforementioned approaches; thus, researchers are left to struggle to understand the practical differences between them. Each of the agreement indices attend to different aspects of what constitutes agreement which can lead to diverse estimates of agreement. This is problematic for researchers trying to select an agreement index appropriate for their data. The basic question is when should a given agreement index be utilized. To answer this question we must look beyond the derivation and conceptualization of agreement indices. We must move toward applying these

indices to data with known properties to assess the implications of multiple issues related to the measurement of agreement. For example, what is the effect of the number of raters and items on agreement estimates? What effect does the distribution of ratings have on each of these agreement estimates? Does measurement level impact the estimates computed under each approach? Are raters equally perceptive to target stimuli?

The present investigation seeks to bridge this gap in the literature by comparing several approaches to measuring interrater agreement. The primary focus of this effort is to determine the accuracy of each approach in recovering *true* agreement. While in an applied setting *true* agreement is unknown and must be estimated, by simulating data *true* agreement can, in theory, be controlled. This makes it possible to evaluate the accuracy of the agreement estimates in recovering *true* agreement. Extensions to "real" data are also examined.

Agreement vs. Reliability

Much has been written discussing the distinction between agreement and reliability (e.g., Goodwin, 2001; James, Demaree, & Wolf, 1985; Kozlowski & Hattrup, 1992; Shrout, 1993). While these terms are often used interchangeably, their meanings can actually differ substantially. Interrater reliability, sometimes referred to as consistency, measures the degree to which the category of a given response can be predicted from knowledge of another's response category. Thus, reliability measures the linear relationship between the ratings, or the extent to which the rank orders of the ratings match. In contrast, interrater agreement, or consensus, describes the degree to which the ratings of two or more raters exactly match. Thus, agreement measures the

extent to which the ratings of a given rater are interchangeable with ratings of another rater. In this way, agreement can be considered a special case of reliability.

For example, consider two managers who rate the productiveness of three employees on a 6-point scale ranging from 1 (very poor) to 6 (excellent). If one manager gives ratings of 4, 5, and 6 and the other manager gives ratings of 1, 2, and 3 to the same employees, interrater reliability (consistency) is +1. The rank orders of the ratings made by the two managers match exactly. However, even a cursory examination of the ratings reveals that the managers did not fully agree. While they were in agreement as to the relative position of the productiveness of the three employees, they did not agree in the absolute sense on the level of productivity exhibited by the three employees. Measures of interrater agreement (consensus) reflect this lack of strict agreement.

Unfortunately, the fundamental differences between interrater reliability and agreement are often overlooked (Kozlowski et al., 1992). We hope to show the extent to which this confusion can influence 1) the proper selection of an approach to measuring interrater agreement and 2) the interpretation of the resulting agreement estimates.

Why measure agreement?

Quantifying the degree of agreement between ratings made by multiple raters can be useful in many situations. For example, in psychological research it is common to have multiple raters make ratings of individuals on trait adjectives (e.g., ratings of personality or behavior). The average of these ratings is computed to represent the construct or trait being judged under the assumption that there is consensus among raters. Often this assumption goes untested, and as Kenny (1994) points out if there is little

agreement among the ratings, using averages as indicators of a construct could lead to false impressions of a target's characteristics.

Quantifying agreement is also useful in allowing us to examine individual differences in the judgment process and the nature of disagreements between raters (Tanner & Young, 1985). There are many reasons for disagreements among ratings: error, ambiguous rating criteria, acquaintance with target, differential meaning systems, and differential scale usage are but a few. Regardless of the specific sources of disagreements it is important to recognize that ratings contain information about a given target *and* specific information about the rater. For example, in a situation where there is a large discrepancy in ratings, we might be interested in understanding whether there is something inherently different about how this rater makes judgments (e.g., different meaning systems, different use of rating scale, etc). Or perhaps it is something about the target that is more difficult to rate in comparison to others (Kenny, 1993).

Unfortunately, we are limited in our ability to explain these situations from most current interrater agreement indices. While there is growing awareness of the need for statistical modeling approaches in measuring rater agreement (Agresti, 1992; Uebersax, 1992), the most common approach is to compute one of many non-model based, omnibus agreement statistics. This makes it difficult, if not impossible, to isolate the impact of any single rater on overall agreement. To do so would require extensive effort – something to the effect of computing agreement for all k raters then removing raters one at a time, recomputing agreement with each subsequent removal.

Additionally, such omnibus indices provide no mechanism by which *true* rater agreement can be controlled in a simulated setting. We hope to bypass this issue by

simulating data via one of the modeling approaches available for rater data. This provides for control of the desired parameter, *true* agreement, which can be compared to the observed agreement recovered by the various agreement indices.

We chose to simulate data via an Item Response Theory (IRT) model. The goal is to identify which of the agreement indices accurately recover *true* agreement as specified by the IRT model. We recognize that this data simulation approach might bias one of the agreement indices (*CSI*, discussed more later), yet we feel that the potential information to be gleaned from this study outweighs this bias. Nonetheless, we will proceed with due caution when interpreting the results of this analysis.

Nine agreement indices were selected for this comparative study. A brief discussion of each of these indices follows.

Kappa

The kappa coefficient (κ) was introduced by Cohen (1960) as a summary index for measuring chance corrected agreement between 2 raters using a nominal scale of measurement. Kappa compares the agreement in the observed ratings to that expected if the ratings were independent. Since its introduction kappa has been extended to include differential weighting (Cohen, 1968), conditional agreement, associational relationships, and multiple (>2) raters (Fleiss, 1971; Light, 1971).

Let π_{ii} denote the probability that 2 raters place a subject in category i and $\Sigma_i \pi_{ii}$ be the total probability of agreement. If the rater's ratings are independent then $\pi_{ii} = \pi_{i+}\pi_{+i}$, and the probability of agreement equals $\Sigma_i \pi_{i+}\pi_{+i}$. Cohen's κ is expressed by

$$\kappa = \frac{\sum_i \pi_{ii} - \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}. \quad (1)$$

Theoretically, κ can assume values on the interval [-1,1]; however values less than zero indicate a level of agreement that is less than what is expected by chance alone. As such, values less than zero are typically ignored. κ will approach 1 when there is high agreement among raters and will be approximately zero when agreement between raters equals that expected by chance.

Fleiss (1971) extended the logic of κ to the multiple rater situation. Let x_{ij} be the number of ratings on target i (i=1,...,n) in category j (j=1,...,k) then

$$\sum_j x_{ij} = m \quad (2)$$

for all i. The value of κ_m is then given by

$$\kappa_m = 1 - \frac{\sum_i x_{ij}(m - x_{ij})}{nm(m-1)\bar{p}_j\bar{q}_j} \quad (3)$$

where n is the number of targets, m is the number of raters, $\bar{p}_j$ is the overall proportion of ratings in category j, and $\bar{q}_j$ is 1 minus $\bar{p}_j$.

The kappa statistic is generally regarded as being better than simple percentage of agreement measures for assessing nominal level agreement in that it does correct for chance agreement. However, κ has been criticized for the assertion that it is a "chance-

corrected measure of agreement". As Uebersax (2001) discusses, the expected agreement term is computed under the assumption of statistical independence of raters which generally does not hold. κ can reach its theoretical maximum value (1.0) only when the marginal distributions for the raters' sets of data are the same. Additionally, κ suffers from not having a mechanism (e.g., Spearman-Brown prophecy formula) by which its value can be adjusted to account for inclusion or removal of items.

Intraclass Correlation

The intraclass correlation (*ICC*) is in a family of measures derived from generalizability theory (Cronbach, Gleser, Nanda, & Rajartnam, 1972). Cronbach and colleagues blended their variance components estimation methodology with the logic of the analysis of variance (ANOVA) to develop a comprehensive plan for analyzing measurement variation. This plan, generalizability (G) theory, allows for the decomposition of error variation into components attributable to sources other than true-score and error variance (Shavelson & Webb, 1991; Conway & Schaller, 1998; Goodwin, 2001).

The family of *ICC* coefficients compares the discrepancies of ratings of individual targets to the degree to which the targets are distinguished from one another. If the discrepancies between ratings of individual targets are small relative to the degree to which the targets are distinguished from one another, then agreement is considered high. If, however, the discrepancies within targets are as large as the differences between targets, then agreement is judged to be low. The *ICC*, like other G theory techniques, requires that the data be measured on interval- or ratio-level scales.

Shrout and Fleiss (1979) distinguish six forms of the *ICC*. Each of the forms is deemed to be appropriate for specific situations depending on design and conceptual intent. Of specific interest to the measurement of agreement is the form Shrout and Fleiss call the *ICC(2, 1)*.

The *ICC(2, 1)* assumes a random sample of k raters is selected from a larger population. In this sample, each rater rates each of n targets. The rating, x_{ij}, which denotes the i^{th} rating (i=1,...,k) on the j^{th} target (j=1,...,n), can then be specified by the equation

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}. \quad (4)$$

In this equation, μ is the grand mean of the ratings; a_i is the difference between μ and the i^{th} judge's ratings; b_j is the difference between μ and the j^{th} target's true score (the mean across many repeated ratings on the j^{th} target); $(ab)_{ij}$ is the degree to which the i^{th} judge departs from his or her usual rating tendencies when confronted by the j^{th} target; and e_{ij} is the random error in the i^{th} judge's scoring of the j^{th} target. The target component b_j is assumed to vary normally with a mean of zero and variance of σ^2_T, and the error terms e_{ij} are assumed to be independently and normally distributed with a mean of zero and variance σ^2_E.

The variance attributable to these effects is partitioned into separate components by calculating generalizability coefficients. The ratio of target variance to total variance is given by

$$ICC(2,1) = \frac{Var(b)}{Var(a)+Var(b)+Var(ab+e)} \quad (5)$$

where $Var(b)$ is an estimate of target variance σ^2_T, $Var(a)$ is an estimate of rater variance σ^2_J and $Var(ab+e)$ is overall error variance attributable to the combination of rater-specific patterns (ab) and error (e).

The *ICC(2, 1)* estimates the expected correlation that would be obtained between two ratings made by randomly chosen raters; in other words, a_i is a random variable that is assumed to be normally distributed with a mean of zero and variance σ^2_J. As such, the *ICC(2, 1)* can be appropriately interpreted as an agreement index (Shrout & Fleiss, 1979; Shrout, 1993). Among the other forms of the ICC described by Shrout and Fleiss is the *ICC(3, 1)* which treats rater effects as fixed, thus is considered a measure of consistency (or reliability) rather than agreement.

The *ICC(2, 1)* takes values between 0 and 1. Values approaching 1 are obtained when there are few discrepancies in the ratings of a target, thus high agreement. The index will approach 0 when the discrepancies within targets becomes large relative to the variation between targets (Shrout, 1993).

Fleiss and Cohen (1973) have shown that the *ICC* is equivalent to kappa under certain weighting conditions, specifically, when the weights, W_{ij}, are defined by,

$$W_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}, \quad (6)$$

the proportional squared distance between ratings of two raters, i and j, on k items.

The advantage of the *ICC* over some other measures of agreement is its strong conceptual foundation and model-based nature. As previously stated, multiple sources of measurement error can be included in the model expanding the amount of information that the index provides regarding not only agreement but also potential sources of disagreement. Unfortunately, the *ICC* is perceived as a more complex index than some other measures of agreement which could possibly curtail more widespread application (Kenny, et al., 1994).

$r_{WG(J)}$

James, Demaree, and Wolf (1984) presented a method for assessing agreement defined by the formula

$$r_{WG} = 1 - (s_{x_j}^2 / \sigma_{EU}^2), \tag{7}$$

where r_{WG} is the within group interrater agreement for a group of k raters on a single item X_j, $s_{x_j}^2$ is the observed variance on X_j, and σ_{EU}^2 is the variance under a uniform distribution on X_j that would be expected if all ratings were due exclusively to random measurement error. This agreement coefficient was derived for the narrow situation in which agreement must be assessed among ratings of a single target on a single item.

James, et al (1984) extended the logic of assessing agreement for ratings made of a single target on a single item to ratings on multiple items of a single target. This index, $r_{WG(J)}$, is defined by

$$r_{WG(J)} = \frac{J[1-(\overline{s_{xj}}^2 / \sigma_{EU}^2)]}{J[1-(\overline{s_{xj}}^2 / \sigma_{EU}^2)]+\overline{s_{xj}}^2 / \sigma_{EU}^2)}, \quad (8)$$

where $r_{WG(J)}$ is the within-group interrater agreement for raters' mean scores based on J essentially parallel items, $\overline{s_{xj}}^2$ is the mean of the observed variances on the J items, and σ_{EU}^2 is the variance on X_j that would be expected if all judgments were due exclusively to random measurement error. The resulting value estimates the degree to which observed similarity in responses is due to actual agreement between group members. Values of $r_{WG(J)}$ generally range from 0 (no consensus) to 1 (perfect consensus); however, negative values and/or values larger than 1 can be obtained when the observed variance is greater than expected by chance.

The advantage of measures such as $r_{WG(J)}$ is that they do not depend on the observed between-groups variance. However, $r_{WG(J)}$ has suffered criticism for the ambiguity involved in choosing the reference distribution (Kozlowski & Hattrup, 1992). Finn (1970) originally used the uniform distribution as the distribution representing chance responding. James, et al. (1984) also recommended the uniform distribution but introduced other alternatives for use as the reference distribution and suggested that the choice of the reference distribution should be guided by the researcher's knowledge of the phenomenon in question (Lindell, et al., 1999). Additionally, $r_{WG(J)}$ has been criticized because like many variance measures, by definition, it depends on the number of items (J) in the scale. All other factors being equal, values of $r_{WG(J)}$ will be larger as J increases.

$r^{*}_{WG(J)}$

Recognition of $r_{WG(J)}$ as an agreement index rather than a reliability coefficient led Lindell, Brandt, & Whitney (1999) to question the application of the Spearman-Brown correction in deriving Equation 8. James, et al. (1984), following the work of Finn (1970), initially labeled $r_{WG(J)}$ as an index of interrater reliability. However, as derived and conceptually defined, it is clearly a measure of interrater agreement (Kozlowski & Hattrup, 1992; James, Demaree, & Wolf, 1993). As such, Lindell, et al. recommended the $r_{WG(J)}$ index be modified as follows:

$$r^{*}_{WG(J)} = 1 - \frac{\bar{s}_x^2}{\sigma^2_{EU}}, \tag{9}$$

which substitutes the average item variance into the numerator of equation 7. The modified index, $r^{*}_{WG(J)}$, assumes a unifactorial scale with independent items and equal variance. $r^{*}_{WG(J)}$ is a linear function (inverse) of $\overline{s_{xj}}^2$. As the average variance among ratings increases, agreement as assessed by $r^{*}_{WG(J)}$ decreases. The range of $r^{*}_{WG(J)}$ is on the interval [–1, 1] (for a 5-point rating scale). Negative values are obtained when the observed agreement is less than expected by chance. Additionally, $r^{*}_{WG(J)}$ is invariant with respect to the number of items *J*, thus can be expected to consistently yield smaller values for agreement than $r_{WG(J)}$ (Lindell & Brandt, 1999).

Relational Agreement and r_s

Stine (1989) proposes a theory for assessing consensus based on measuring agreement with respect to the empirically meaningful relationships among the ratings. These meaningful relations are defined as those that are invariant with respect to admissible transformations. From this definition, Stine develops a definition of "relational agreement" in which the meaningful relations generated by two raters are compared to one another. Agreement and reliability as previously defined can be subsumed under this larger umbrella of relational agreement in which meaningful relations are determined according to the scale of measurement of the data. The term relational agreement refers to "the degree to which one variable is an admissible transformation of the other variable" (Stine, 1989). If the data are measured on an absolute scale of measurement, then a relational agreement measure is a measure of agreement (as traditionally defined). However, if the data are measured on an interval scale, then the Pearson's correlation coefficient is a measure of relational agreement.

Coefficients appropriate for measuring relational agreement will vary as a function of the degree to which one of the two variables (data generated by the raters) approximates an admissible transformation of the other variable, with transformation admissibility being defined by the scale of measurement for the two variables. If the measures are absolute scale variables, then the coefficient will be a function of the degree to which the numbers generated by one rater exactly match those from the other rater (i.e., the two raters agree). If the measures are ordinal, then the coefficient will be a function of the degree to which the two sets of ratings (one from each rater) have the

same order (i.e., they can be associated with one another using a strictly monotonic function).

Stine proposed a set of coefficients appropriate for measuring relational agreement based largely on the work of Zegers and ten Berge (1985). Zegers and ten Berge proposed a general formula for association coefficients between two variables for the absolute, additive, ratio, and interval scales. Stine extended this work and developed uniforming transformations for the remaining measurement scales. Additionally. Fagot (1993) proposes a generalization of the Zegers-ten Berge theory to the case of multiple judges (for absolute, additive, ratio, and interval scales). This discussion will concentrate on Stine's proposed coefficient for ordinal scales as it is consistent with the manner in which most ratings in the social sciences are made.

The coefficient for ordinal scales is defined by

$$o'_{XY} = 2\sum_{i=1}^{n} R(X_i)R(Y_i) / [\sum_{i=1}^{n} R(X_i)^2 + \sum_{i=1}^{n} R(Y_i)^2], \tag{10}$$

where $R(X_i)$ and $R(Y_i)$ are the respective ranks of ratings X_i and Y_i. Correcting for chance, the equation becomes

$$o'_{XY} = \frac{2[\sum_{i=1}^{n} R(X_i)R(Y_i) - 1/n \sum_{i=1}^{n} R(X_i) \sum_{i=1}^{n} R(Y_i)]}{[\sum_{i=1}^{n} R(X_i)^2 + \sum_{i=1}^{n} R(Y_i)^2 - 2/n \sum_{i=1}^{n} R(X_i) \sum_{i=1}^{n} R(Y_i)]}. \tag{11}$$

Note that for multiple raters, this is equivalent to an average Spearman rank correlation coefficient, r_s (Stine, 1989).

Fagot (1991) demonstrates that the *ICC(2, 1)* can be viewed as a coefficient of relational agreement. In fact, if the appropriate uniforming transformation is applied, a different estimate of *ICC(2, 1)* can be computed for each possible scale of measurement. Thus, the *ICC(2,1)* can be viewed as a family of coefficients of relational agreement unto itself.

Latent Trait Analysis and the CSI

Terry (1995) proposes a model for understanding the process of making judgments which is particularly useful for statistical and measurement purposes in that it is simply a specific case of the general two-parameter (or Birnbaum) IRT model. This model is called the *L*atent *TR*ait model of *I*nter*P*ersonal *P*erception (LaTRIPP). The model is derived from Item Response Theory and proposes that we consider the probability that target i receives a rating from rater j (denoted as $P_{ij}(\theta)$) as a function of three model parameters: θ_i, the level of the trait "possessed" by target i; α_j, the sensitivity of rater j to the criterion θ; and, β_{ij}, the threshold of rater j to the criterion θ. A rating begins with the target's true level of the trait θ. This information, along with random errors and unique variation leads to each rater's impression of the target's trait level. Each rater applies his/her discretionary thresholds to the trait being judged to yield a rating. The parameters captured by this rater model are related by the following equation:

$$P_{ij}(\theta)=\left[1+e^{-1.7\alpha_j(\theta-\beta_{ij})}\right]^{-1}. \qquad (12)$$

This equation describes a simple logistic curve denoting the nonlinear regression of the parameter $P_{ij}(\theta)$ on the trait parameter θ_i, with the α_j and the β_{ij} serving as the regression slope and thresholds, respectively. Figure 1 contains one such curve for a rater (j) with a slope parameter α_j equal to 0.8 and a threshold parameter β_{ij} equal to 1.0.

This model can also be related in terms of the logit model (Hambleton & Swaminathan, 1985). Let $P_{1j}(\theta)$ be the probability of giving a rating of 1 on item j, $P_{2j}(\theta)$ be the probability of giving a rating of 2 on item j, and $P_{ij}(\theta)$ be the probability of giving rating i on item j. Then, $O_{ij} = P_{ij}(\theta)/P_{1j}(\theta)$ is the odds of giving a rating i compared to a rating of 1 as a baseline. The log-odds for giving a higher rating using a rating of 1 as a measure of baseline comparison is given by

$$\ln O_{ij} = 1.7\alpha_j(\theta - \beta_{ij}). \qquad (13)$$

This model suggests that the propensity of receiving a particular rating is a function of the level of the attribute θ residing in the target being judged, the rater's thresholds for responding to varying levels of the attribute, and the sensitivity of the rater being able to discriminate between persons with varying degrees of the attribute. The sensitivity parameter, α_j, relates to the ability of a rater to make distinctions in θ among the pool of targets. A rater with a sensitivity parameter near zero for a given trait would be a poor rater – their ratings would not vary with individual differences in θ. The threshold parameter, β_{ij}, relates to information unrelated to the true trait which influences the rater's propensity to apply a given category rating, sometimes referred to as rater leniency. The values of the β_{ij} parameters represent the trait level necessary to respond

above threshold j with .50 probability (Embretson & Reise, 2000). This parameter differentiates raters who may be more or less likely to apply certain ratings. For example, in a courtroom setting it is necessary to make some determination of the guilt or innocence of the defendant. Jurors will vary in the amount of evidence required to vote the defendant guilty. Those jurors demanding nothing short of DNA evidence in addition to eyewitness testimony would have a high threshold β_j for a guilty vote. Conversely, those jurors who are convinced by little evidence would have a low threshold for giving a guilty vote. In the same way, raters will vary in the thresholds they apply to rate targets on the latent trait. For dichotomous ratings (such as guilty/not guilty), only one threshold is estimated. However, when more than two ordered categories are used (such as a Likert scale) more thresholds are estimated. With five response options, there are m_i=4 thresholds between the response options.

Using this model, a calibration sensitivity index (*CSI*; Terry, 2000) can be computed for each rater. The *CSI* is given by the following equation:

$$CSI = \frac{\alpha_j^{\,2}}{1+\alpha_j^{\,2}}. \tag{14}$$

As shown in (14), the crucial parameter of the *CSI* is the sensitivity parameter α_j from the LaTRIPP model. The calibration sensitivity parameter is roughly analogous to the capability of a particular rater to make distinctions in θ among targets in consensus with other raters. A rater with a high level of calibration sensitivity distinguishes among targets with a high level of consensus with the group of raters. In contrast, a rater with a

low level of calibration sensitivity may distinguish among targets, but these distinctions are inconsistent with the other raters. As shown in Figure 1, this parameter is operationally defined to be proportional to the slope of the rater characteristic curve. Thus, a rater with a calibration sensitivity parameter of zero on a particular characteristic would be depicted as having a flat, horizontal line. This means that the ratings of this particular rater would have no consensus with the other ratings, and, subsequently, would not systematically vary with target changes in the trait being judged.

The *CSI* yields values ranging from 0 (no consensus) to 1 (perfect consensus). The *CSI* provides information about potential sources of disagreement at the disaggregated (i.e., individual) level. As previously mentioned, this information is difficult or impossible to glean from other agreement indices.

Summary

The purpose of this study is to critically review and compare various methods of measuring consensus in groups. As previously mentioned, a large literature has developed devoted to understanding how consensus is achieved. This literature is focused on understanding the impact of various individual and/or group characteristics on subsequent levels of group consensus. While this is certainly a key issue in consensus research, the current study will not attempt to describe nor explain the mechanisms contributing to the development of consensus in groups. For this the reader is referred to Kenny (1991). Rather, this study focuses solely on the measurement of agreement and how current methods compare in recovering *true* agreement among raters.

Individual rater data will be simulated using the LaTRIPP model. This type of simulation has not been done before in the evaluation of consensus indices given there is

no model for mapping individual rater judgment processes onto consensus indices. The resulting consensus estimates will be compared for accuracy in recovering true consensus.

Method

Data Simulation

The general 2-parameter IRT model with five response categories was used to generate the simulated datasets. Item probabilities for five categories per item for a simulated rater were generated using Equation 12. To assign a rating for each simulated rater, item parameters, α and β, were chosen (specific methods discussed below). To generate the ratings, each target was randomly assigned a θ (value on the latent trait being rated). These θ values were combined with the IRT parameters (α and β) to create a "probability" of a given rating from which values were sampled to actually obtain the ratings. For each simulated rater a single random number (Y) was sampled which served as the rating of rater j on item k for target i.

Factors Manipulated

Five factors were manipulated to generate the simulated datasets: shape of the prior (true) distribution of θ, dyadic consensus, degree of true consensus, number of raters, and number of items. In sum, 108 conditions were simulated in this study each replicated ten times. This yielded a total of 1080 simulated data matrices (see Figure 2).

True distribution of θ. Each target was randomly assigned a value on the latent trait θ being rated. For the simulations, ratings were sampled from two prior distributions of θ: normal and uniform. The goal of this manipulation is to identify the effect of the true underlying distribution of θ on the agreement estimates.

Degree of true consensus. The use of IRT models allows data to be generated for any desired level of consensus. For the current study, three levels of consensus were simulated: 1) high consensus – defined as 80% agreement between raters, on average, 2) medium consensus defined as 50% agreement between raters, on average, and 3) low consensus defined as 20% agreement between raters, on average. Since consensus is defined to be proportional to the slope α_j of the item characteristic curve, values of α were determined such that

$$True\,Consensus = \frac{\alpha_j^{\,2}}{1+\alpha_j^{\,2}}. \tag{15}$$

Thus, to achieve data with 20%, 50%, and 80% true consensus, values of α_j were set at 0.5, 1.0, and 2.0, respectively.

Slopes. The slopes α_j can be assumed to be invariant across raters or allowed to vary across pairs of raters. Both of these possibilities were examined. For the first level, constant slopes, α_j was set to be invariant across raters (i.e., $\alpha_j = \alpha$ for all j). As such, the model reduces to the Rasch model (Rasch, 1966). This implies that all judges are equally perceptive (or sensitive) to the trait θ being rated. Setting α to be constant for all j raters produces data with constant consensus for all pairs of raters. For example, for *the high consensus-constant slopes* condition, the ratings for any given pair of raters is constrained to achieve agreement of 80%.

This factor also has unique implications for the scale type of the data. For the logit form of the Rasch model, it has been shown that trait levels have interval scale properties if the distances between raters have invariant meaning for behavior, regardless of the trait

level (Rasch, 1977). That is, if the distance between 2 pairs of raters is equal, then the same difference in log odds for responses to any item is expected. Therefore, data that fit the Rasch model have the additivity property, which justifies interval-level measurement (Embretson & Reise, 2000).

The second level of this factor, varying slopes, allowed α_j to vary across pairs of raters with the restriction that average agreement across all dyads be maintained at the specified consensus level. Thus, for the *high consensus-varying slopes* condition, the consensus between pairs of raters will fluctuate; however, the average consensus across all pairs is constrained to achieve agreement of 80%.

Following the same logic from above, data for this condition can be expected to have ordinal scale type. With varying slopes, an interaction exists between rater response probabilities and item-trait level. Thus, equal distances between pairs of raters no longer imply equal differences in log odds for items; the log odds will vary as a function of item-trait level. Hence the additivity property is lost and interval scale measurement can no longer be justified.

Number of raters and items. Lindell et al. (1999) showed that values for $r_{WG(J)}$ increase as the number of items being rated increase. Other indices may similarly be unduly influenced. These concerns also extend to the number of raters. Therefore, the final two factors that were manipulated in the simulations were the numbers of raters and number of items in the scale. Three rater conditions (j = 2, 5, & 25) and three item conditions (k = 10, 25, & 50) were examined.

Thresholds. The thresholds β_{ij} were set a priori and remained constant throughout the simulations. By specifying constant thresholds, we are simulating data assuming all

raters are using the rating scale equivalently. Four threshold values were specified to simulate data from a Likert scale with 5 response options. The thresholds were chosen at the following values to obtain an approximate symmetric distribution of observed ratings: $\beta_j = -1.5, -.75, .75, 1.5$.

Estimation and Computation

Item parameters for the *CSI* calculation were estimated using Parscale 3 (Muraki & Bock, 1997). The marginal maximum likelihood procedure and EM algorithm were used to estimate the item parameters. Program default values were used for all estimation.

All simulations and computation of the remaining indices were performed using SAS (SAS Institute, Inc., 1990). A total of nine coefficients of interrater agreement were computed on the ratings for each of the simulated conditions: κ_m, *ICC(2,1)*, $r_{WG(J)}$ (uniform reference distribution) , $r_{WG(J)_n}$ (normal reference distribution), $r^{*}_{WG(J)}$ (uniform reference distribution), $r^{*}_{WG(J)_n}$ (normal reference distribution), r_s, CSI_n (normal priors on θ), and CSI_u (uniform priors on θ). Means of the resulting agreement estimates were computed across the ten replications of each of the 108 conditions.

Evaluation of Indices

Differences in estimated mean agreement values computed from the various agreement indices were evaluated via a 9 (index) x 2 (true distribution) x 2 (slopes) x 3 (number of raters) x 3 (number of items) x 3 (degree of true consensus) analysis of variance (ANOVA). Two error indices were computed to evaluate the accuracy of the agreement indices at recovering true agreement: bias and root-mean-square error (RMSE). Bias measures the signed difference between the mean agreement estimate and true agreement. RMSE is the positive square root of the mean square of the residuals.

Values of bias and RMSE close to zero indicate low error in the accuracy of the index in recovering true agreement. RMSE and bias were computed across replications for each of the simulated conditions. These values are given by

$$RMSE(\hat{a}) = \sqrt{\frac{1}{n}\Sigma_i(\hat{a}-a)^2} \tag{16}$$

and

$$Bias(\hat{a}) = \frac{1}{n}\Sigma_i(\hat{a}-a), \tag{17}$$

where a is true agreement (a = 20%, 50%, or 80%), $\hat{a}$ is the agreement estimate for the i^{th} replication and n is the number of replications of each condition. A variance components analysis was then conducted using RMSE and bias values as dependent variables. All variance components analyses were completed using SAS's VARCOMP procedure (SAS Institute, Inc., 1990).

To better understand the nature of the variation in the relative magnitudes of the estimates, an additional variance components analysis was performed. Rank orders of the mean agreement values across index, within a condition, were determined. The variance components analysis was then performed on the resulting rank orders.

Results

Appendix A (Table A1) presents the means and standard deviations of the nine agreement estimates computed on the simulated data. Each mean is computed across ten replications of the 108 simulated conditions. During the estimation phase for the *CSI*

calculations, 22 of the 2160 programs (1.02%) necessary to estimate the slope parameters α failed to converge. This includes estimation for both CSI_n (n=1080) and CSI_u (n=1080). The missing α values were replaced using the mean of the slopes for the remaining replications of that condition.

A 9 (Index) x 2 (True distribution) x 2 (Slope) x 3 (Raters) x 3 (Items) x 3 (Consensus) ANOVA (see Table 1) was conducted to evaluate the effect of each factor on the agreement estimates. As the focus of this research is predominantly on differences in estimates as a function of agreement index, only significant main effects of index and interactions including index are discussed and interpreted.

The results of the ANOVA indicated a significant five-way interaction between index, true distribution of θ, number of raters, number of items, and degree of consensus, $F(64, 8748) = 1.36$, $p<0.05$. The size of this effect is minimal considering the sample size; however, we chose to continue with the interpretation of this finding as though it were a real effect and not simply a result of excessive power. No significant effects of slope were indicated, thus, it was subsequently dropped from the model. As a result of this finding, subsequent analyses were based on 20 replications of each condition rather than the original 10 (the two levels of the slope factor were consolidated).

To better understand the significant five-way interaction, the three-factor model including index, raters, and items conditioning on true distribution of θ (2 levels) and true consensus (3 levels) was evaluated. As a result, six 9 (index) x 3 (raters) x 3 (items) ANOVAs were conducted, one for each of the following: 1) normal distribution, high consensus; 2) normal distribution, medium consensus; 3) normal distribution, low consensus; 4) uniform distribution, high consensus; 5) uniform distribution, medium

consensus; 6) uniform distribution, low consensus. The results of these ANOVAs are illustrated in Figures 3-14. These figures present the average agreement estimates as a function of a) the number of items and b) the number of raters. A cursory examination of these figures further suggests that the significant 5-way interaction that was detected may be a product of sample size. Generally, the estimates stay relatively constant across the levels of the rater and item factors. Some small variation is observed, but not enough to be considered meaningful. However, due to excessive power, the ANOVA is detecting these differences as significant.

Normal Prior Distribution

High true consensus (80%). Results of the three-factor ANOVA indicate a significant main effect of index, $F(8, 1539) = 771.72$, $p<.0001$ and a significant interaction between index and items, $F(16, 1539)=3.32$, $p<.0001$ (see Figures 3 & 4). As illustrated in Figures 3 & 4 variability among the agreement estimates is high. Estimates of agreement across this set of conditions range from .31 to .989.

Figures 3 & 4 suggest $r^{\bullet}_{WG(J)}$ is the most accurate estimate of agreement for the high consensus conditions. $r^{\bullet}_{WG(J)_n}$ also appears to recover true agreement well, especially as the number of items increases. Collapsing across rater and item conditions, the average $r^{\bullet}_{WG(J)}$ consensus estimate is 83.8% (SD = 5%; see Table 2) – a discrepancy from true agreement of 3.8%. The average $r^{\bullet}_{WG(J)_n}$ underestimates consensus by 6.3% (M = 73.7%) which is only slightly worse than $r^{\bullet}_{WG(J)}$; however, the variability across replications is much higher (SD = 21%).

The discrepancy from true consensus is one method of determining the error of the measures in recovering true agreement. Bias and RMSE can also be examined to reveal more precisely the error of estimation. Tables 3 & 4 present the average RMSE and bias of the consensus measures conditioning on the prior θ distribution and true consensus across rater and item conditions. As seen in Table 3, $r^{*}_{WG(J)}$ results in the smallest error of approximation with RMSE = 0.06 and bias = 0.033.

Medium true consensus (50%). The three factor ANOVA indicates a significant main effect for index F(8, 1539) = 234.87, p<.0001. Figures 5 & 6 suggest the CSI_n is most accurate in recovering agreement when true consensus is 50%. The average CSI_n is 52% (SD=6%), a discrepancy from true consensus of only 2%. $r^{*}_{WG(J)}$ also performs well for this set of conditions with an average estimate of 57.6% (SD = 10%). RMSE and bias of CSI_n and $r^{*}_{WG(J)}$ provide additional evidence of the accuracy of these indices. Average RMSE for CSI_n and $r^{*}_{WG(J)}$ are 0.06 and 0.111, respectively. Both indices show a slight positive bias in estimation – bias of CSI_n is 0.019 and bias of $r^{*}_{WG(J)}$ is 0.071.

Low true consensus (20%). Results of the three factor ANOVA indicate a significant main effect for index, F(8, 1539) = 6.39, p<.0001. Overall, the estimates are less variable for this set of conditions. However, as illustrated in Figures 7 & 8, $r_{WG(J)}$ and $r_{WG(J)_n}$ deviate dramatically from the remaining indices with values approaching and/or exceeding 1. For 5 raters and 25 items $r_{WG(J)}$ soars to nearly 3.0.

Across rater and item conditions, r_s and *ICC(2, 1)* outperform the other indices with consensus estimates of 17.9% and 18.4%, respectively. The RMSE results are not completely consistent with these results. On average, the CSI_u estimates showed the

lowest RMSE (RMSE=.111) with the *ICC(2, 1)* and r_s following close behind (RMSE=.128 and RMSE=.132, respectively). Nonetheless, the average bias values support r_s and *ICC(2, 1)* as the best estimates of consensus for true consensus of 20%. Both measures showed a slight negative bias; however, both were less (in an absolute sense) than the bias of CSI_u.

Uniform Prior Distribution

High true consensus (80%). Results of the three factor ANOVA indicate a significant main effect of index, F(8, 1539) = 1506.47, p<.0001; a significant interaction between index and raters, F(16, 1539) = 3.07, p<.0001; and a significant interaction between index and items, F(16, 1539) = 16.05, p<.0001. Figures 9 & 10 suggest CSI_n is responsible for the significant interaction between index and items. CSI_n approaches true consensus as the number of items in the scale increase. While κ_m and CSI_u also show increases as the number of items increase, the CSI_n is certainly the most affected in this case. Similarly, CSI_n shows a decrease in value (away from true consensus) when the number of raters is 25.

Across rater and item conditions, the average estimates suggest CSI_n and $r^*_{WG(J)_n}$ most accurately recover true consensus. CSI_n estimates consensus at 78.2% (SD = 8%), which is only slightly better than $r^*_{WG(J)_n}$ at 82% (SD = 5%; see Table 2). However, the instability of CSI_n suggests $r^*_{WG(J)_n}$ might be the better estimate for this set of conditions, particularly with a small number of items ($k < 25$). The average RMSE of $r^*_{WG(J)_n}$ is 0.051 with bias of 0.02.

Medium true consensus (50%). The consensus measures yield highly variable estimates when true consensus is 50%. Estimates range 31.5% (κ_m) to 96.9% ($r_{WG(J)}$). The three factor ANOVA indicates a significant main effect of index, $F(8, 1539) = 1011.9$, $p<.0001$ and a significant interaction between index and items, $F(16, 1539) = 5.62$, $p<.0001$. As illustrated in Figures 11 & 12, $r^{*}_{WG(J)_n}$ most accurately recovers 50% true consensus, averaging 50.4% across rater and item conditions. CSI_u also performs well when the number of items is 10; however, its performance tapers off as the number of items increases. The RMSE values for CSI_u are lower, on average, than $r^{*}_{WG(J)_n}$, yet estimates of bias support $r^{*}_{WG(J)_n}$.

Low true consensus (20%). Results of the three factor ANOVA indicate a significant main effect of index, $F(8, 1539) = 2.83$, $p<.01$ and a significant interaction between index, items, and raters, $F(32, 1539) = 1.73$, $p<.01$. Averages across item and rater conditions suggest $r_{WG(J)}$ respectably estimates true consensus of 20%. The average $r_{WG(J)}$ is 29.7%. However, the standard deviation of $r_{WG(J)}$ (SD = 5.51 or 551%) suggests further investigation of the estimates is warranted. A cursory glance at Figures 13 & 14 and a RMSE of 2.731, almost three times that of the other indices, uncovers the real story. $r_{WG(J)}$ is drastically effected by the number of raters and items. While the average $r_{WG(J)}$ across raters and items might closely approximate true consensus, $r^{*}_{WG(J)}$ is clearly a better index for this set of conditions. Average RMSE of $r^{*}_{WG(J)}$ is 0.141 with average bias of 0.018. Interestingly, κ_m shows the lowest average RMSE and bias. However, due to clear violations of its assumptions, κ_m is not appropriate for this data.

Variance Components Analyses

Tables A2 and A3 present the error of estimation for each of the simulated conditions as assessed using the RMSE and bias indices. Two variance components analyses were performed: one using RMSE as the dependent variable and one using bias as the dependent variable. Tables 6 and 7 present the estimated variance components and percentage of total variance attributable to each source of variation for RMSE and bias, respectively. In keeping with the analysis plan from the ANOVA, the variance components were computed conditioning on the degree of consensus (high, medium, or low) and true distribution of θ (normal or uniform). Initial estimation included all main effects and interactions. Any near-zero or negative variance components were pooled into error variance and the results presented in Tables 6 and 7 were obtained.

Normal Distribution

High true consensus (80%). As seen in Table 6, the main effect of index accounted for 75% of the total variance in the RMSE estimates – meaning that across items and raters the residuals of the nine consensus indices primarily differed as a function of the index computed. The next two variance estimates are those attributable to the two-way interactions in the model. The index x items interaction accounted for 10% of the total variance, indicating a small proportion of the variance in RMSE is attributable to the number of items being rated. The index x raters interaction accounted for only 1% of the total variance.

Additionally, Table 6 shows the amount of variance associated with the three-way interaction, index x items x raters. In the ANOVA model used here, this interaction is also called the residual because it contains all other sources of error not specified in the

model. In this case, 14% of the variance in RMSE is attributable to the residual component.

Table 7 presents the percentage of total variation in bias values attributable to each measured source of variation. Index explained 96% of the variation in bias values, leaving only 4% to be explained by other sources of variation. Similar to the RMSE results, this indicates that index alone accounts for a large proportion of the variability in the error values.

Medium true consensus (50%). As seen in Table 6, index accounted for 56% of the total variance in RMSE for the medium consensus conditions. As before, this indicates that across items and raters the agreement indices differed among themselves on their average distance from the average consensus value. The index x raters and index x items interactions accounted for a minimal percentage of the total variance: a combined 10%. The three-way interaction accounts for 24% of the variability in RMSE. This is a larger proportion than seen in the high consensus condition. In general, there is less variability among the agreement estimates at 50% consensus, thus less of the variation will be explainable by index alone.

Index accounted for 95% of the total variation in the bias of the consensus measures (see Table 7). The two-way interactions accounted for 2% leaving 3% to be explained by the residual component.

Low true consensus (20%). The low consensus condition shows a different pattern of results than seen in the medium and high consensus conditions. The main effect of index accounted for only 11% of the total variance. For the two-way interactions, the variance components estimates were both zero. However, Table 6 shows the relative

amount of total variance associated with the three-way interaction, index x items x raters, is quite large: 89%. The fact that it is so large means that, to some extent, the agreement indices differed nonsystematically across the items and raters conditions. Often this might be troublesome, but as in the medium consensus condition, there is very little variability (relatively speaking) for the low consensus condition. Those indices which consistently underestimated consensus (i.e., kappa) for the high and medium consensus conditions are much more in line with the other estimates for the low consensus situations. Thus, index alone was less likely to explain a large portion of the variability in the RMSE (and bias) values. Index accounted for 36% of the total variation in bias (see Table 7). The remaining 64% of the variation is attributable to the residual component.

Uniform Distribution

High true consensus (80%). Similar to the results obtained when a normal distribution of θ is assumed, the main effect on index in the high consensus condition for a uniform θ distribution accounted for 85% of the total variance in the RMSE values. The index x items and index x raters interactions accounted for a combined 4% of the total variance. The three-way interaction, index x items x raters, accounted for 11% of the total variance. Overall, these results indicate that the variability among RMSE values is predominantly attributable to differences in the index being computed.

Index accounted for 94% of the total variability in the bias values. The remaining 6% of the variability was dispersed across the two-way interactions and the residual component.

Medium true consensus (50%). As seen in Table 6, index accounted for 89% of the total variability in RMSE. The index x raters and index x items interactions

accounted for 8% of the total variance of the RMSE values. The remaining 3% was attributable to the three-way interaction, index x items x raters.

Index accounted for 97% of the total variation in bias (see Table 7). The two-way interactions, index x items and index x raters, accounted for the remaining 3% of the total variation.

Low true consensus (20%). Similar to the normal θ distribution conditions, the low consensus condition assuming a uniform θ distribution also shows a small percentage of variation in RMSE attributable to index – only 8%. The index x items interaction accounted for 10% of the total variability in RMSE, while the index x raters interaction accounted for 18%. The remaining 64% is accounted for by the residual component. As before, this indicates that the RMSE values for the indices differed nonsystematically across the items and raters conditions.

Index accounted for only 8% of the total variation in bias. The remaining 92% of the variation is attributable to the residual component. Again, this is most likely due to the lower overall variability in estimates between the agreement indices. Thus, the remaining variability was due to factors other than index or the index x items and index x raters interactions.

Rank orders

An initial inspection of the consensus indices revealed that the consensus measures tended to maintain a consistent ordering in magnitude regardless of specific condition characteristics. Therefore, an additional variance components analysis was performed to determine what percentage of the variability in the rank orders could be explained by index. The rank orders of the agreement indices served as the dependent

variable. As previously stated, rank orders were computed for the means of the 20 replications of each of the conditions across agreement index. The average rank order for each index is presented in Table 8. The estimated variance components and percentage of total variation for the rank order data are presented in Table 9. Five factors were included in the variance components analysis: index, true distribution, raters, items, and degree of consensus. Negative and near-zero variance components were pooled into error and the results presented in Table 9 were obtained (Cronbach, et. al., 1972). As shown, the only positive, non-zero variance component was that of index, accounting for 72% of the total variability in the rank orders. These findings confirm our suspicions, suggesting that the consensus measures tend to order themselves similarly for any given set of data despite key differences that might exist within the data.

Conclusions

This simulation study extends the literature on the measurement of agreement in a number of ways. First, consensus measures were evaluated against simulated individual rater data with known properties. To date, this type of simulation has not been performed in conjunction with the study of the measurement of agreement given there has been no model available for mapping responses of individual raters onto agreement indices. Terry's LaTRIPP model (2000) allows consensus research to move beyond the simple simulation of variance components or other aggregate properties of data (e.g., variances, etc.) by modeling underlying rater processes. The LaTRIPP model recognizes and models the different thresholds and sensitivities that raters may have and mathematically relates them to the common latent trait which underlies the entire rating process.

A second contribution of this study is the identification of factors that differentially influence estimates of rater agreement. Results of this study suggest that agreement estimates are significantly related to the unique combination of five factors; this combination is represented by the significant 5-way interaction between the number of raters, number of items, true consensus, choice of prior distribution of θ, and agreement index. The size of this interaction effect and subsequent graphical examination of the simulated data suggest that the interaction is primarily an artifact of excessive power due to the large sample size (N=9720; 9 indices computed on 1080 simulated data sets).

The only factor manipulated in this study which did not have a significant influence on the agreement estimates was the slope factor. Results suggest there is no difference in agreement estimates computed on data generated from the 2-parameter model (α_j varying) versus the simpler Rasch model ($\alpha_j = \alpha$ for all j). This is somewhat surprising in light of previous research (Murphy et al., 2002) and theoretical considerations. It seems natural to assume raters differ in their sensitivities to subject stimuli. However, at least for the current data, this conclusion is not supported.

This finding also has interesting implications for the effect of scale type on agreement estimates. As discussed previously, the Rasch model implies interval scale data, while the 2-parameter model implies, at best, ordinal scale data. If the Rasch model fits as well as the 2-parameter model, it follows that scale type is not an important factor for IRT models in the estimation of consensus. While many argue for the importance of scale type in estimating IRT models, Velleman & Wilkinson (1993) argue that scale type of the data does not really matter for many important circumstances, such as the use of

parametric methods for analyzing data. The current data tend to support Velleman and Wilkinson's position.

This does not imply, however, that there are not advantages to fitting a model where individual raters' slopes are estimated. The slope is an individual differences measure of the sensitivity of a rater for a given value of the latent trait. As such, estimates of the individual slopes are useful for identifying raters who are better at distinguishing among different θ levels in a pool of targets ("good" raters) versus those who cannot make these fine distinctions ("poor" raters).

Taken together, the results stemming from the simulated data suggest the most accurate measures of consensus are $r^{*}_{WG(J)}$, CSI_{n}, and CSI_{u}. When the distribution of the latent trait was assumed to be approximately normal in shape, the most accurate consensus measures were $r^{*}_{WG(J)}$ and CSI_{n}. κ_{m}, r_{s}, $ICC(2,1)$, $r^{*}_{WG(J)_{n}}$, and CSI_{u} consistently *under*estimated true agreement in the simulated data; whereas, $r_{WG(J)}$ and $r_{WG(J)_{n}}$ consistently *over*estimated true agreement. When the distribution of the latent trait was assumed to be approximately uniform, the most accurate measures of consensus were $r^{*}_{WG(J)}$ and CSI_{u}. κ_{m} and CSI_{u} consistently *under*estimated true agreement here, while r_{s}, $ICC(2,1)$, $r_{WG(J)}$, and $r_{WG(J)_{n}}$ *over*estimated consensus, on the average. These findings imply that the choice of the agreement index and the distribution of the latent trait will impact an index's ordering relative to true consensus.

These findings are particularly interesting because they imply that the choice of the reference distribution does influence the accuracy of the estimates of the *CSI*. If the latent trait is presumed normal (or uniform) in shape, better estimates will be obtained

when the prior distribution is properly specified. This is also true for $r^{*}_{WG(J)}$. For $r^{*}_{WG(J)}$ using the uniform expected variance yields the most accurate estimates when the true underlying distribution of θ is normal, and $r^{*}_{WG(J)_n}$ is most accurate when the true underlying distribution of θ is uniform. Further investigation is necessary to uncover the cause of this mismatch.

From the outset of this study, I expected the accuracy of the CSI to far exceed the accuracy of the other measures of agreement. Therefore, I was surprised at the relatively poor performance of the *CSI* for the high and low consensus conditions. There have long been concerns about the ability of IRT estimation routines to recover parameter values. Results of the current study suggest there may still be unresolved issues regarding IRT parameter estimation. Specifically, the current data suggest the prior distributions typically placed on the parameters during estimation may be overly influential on resulting parameter estimates.

By default, PARSCALE places priors on model parameters to aid parameter estimation. Relevant to the current study is the specification of a prior distribution for α_j, the item discriminations. A log normal prior distribution with mean = 0 and SD = 0.5 is the default for PARSCALE (Muraki, et al., 1997). Generally, the effect of the priors on the resulting estimates should be minimal given sufficiently large sample sizes for both the number of items and raters. For the current data, the priors appear to be somewhat too influential in the estimation of consensus values when the consensus is either high (~80%) or low (~20%).

A mean of 0 for $\log(\alpha)$ corresponds to $\alpha = 1$. As previously shown, when $\alpha = 1$, rater consensus is 50%. The current study simulated data for true consensus of 80%,

50%, and 20%. Consensus is accurately recovered by the *CSI* when true consensus equals 50%. However, as the value of true consensus moves away from 50%, the *CSI* becomes less accurate. For example, for the 80% true consensus conditions, to accurately recover consensus, slopes estimates should approach $\alpha = 2$. However, CSI_n consistently *under*estimated true consensus possibly indicating that the slope estimates are being constrained by the prior – pulled toward the prior of $\alpha = 1$. For the 20% true consensus conditions, the slopes estimates should approach $\alpha = 0.5$. However, for these conditions, estimates of CSI_n consistently *over*estimated true consensus - again pulled toward the value of the prior of $\alpha = 1$.

IRT parameter estimates can be obtained without placing any priors on the distribution of the slopes. Subsequently, we attempted this variant of estimating *CSI*. Unfortunately, PARSCALE encountered more difficulties converging to a solution than when priors were specified. When a normal latent trait distribution was assumed, 38 of the 1080 programs (3.5%) necessary for parameter estimation failed to converge. When a uniform latent trait distribution was assumed 66 of 1080 programs (6%) failed to converge. These non-convergence issues are most likely due to a few raters whose ratings look dissimilar to those of the other raters. Priors tend to help reign in these values. However, without priors, the estimation encounters difficulty in estimating parameters for these few outlying raters - especially with the relatively small number of raters and items in this investigation. While this is not an especially large percentage of the data, for situations where multiple replications are not available (for mean replacement, etc.) the lack of convergence presents a problem. A potential solution to this estimation problem will be presented at the conclusion of this manuscript.

Lindell's modified index, $r^{*}_{WG(J)}$, results in values which accurately recover true agreement, but these values are generally too high. While it could be argued that $r^{*}_{WG(J)}$ proved to be the best index for assessing agreement (across all conditions), we propose that the *ICC(2,1)* or the *CSI* might be better estimates. The *ICC(2, 1)* and *CSI* are conservative, but they will not overestimate consensus. Lindell, et al. (1999) point out small sample sizes can bias values of $r^{*}_{WG(J)}$. They suggest when the number of raters is 2, any level of disagreement greater than random response yields "improper" values for $r^{*}_{WG(J)}$. However, when the number of raters is at least 10, extreme levels of disagreement are required in the sample to yield improper values of $r^{*}_{WG(J)}$ (when using a 5-point scale). The current study did not replicate these findings, but the lowest level of consensus examined in this study (20%) is still greater than chance agreement. Consequently, we may not have examined consensus at values low enough to replicate the previous results.

The *ICC(2, 1)* proved to be relatively accurate for the low consensus conditions. The *ICC(2, 1)* estimates were generally too low for medium to high levels of true consensus. Yet, for low consensus levels *ICC(2, 1)* recovered true agreement as well as $r^{*}_{WG(J)}$ and CSI_{n}. One possible explanation for the observed result follows. As previously discussed, the CSI_{n} is a nonlinear model of rater judgment processes. In this model, consensus is accounted for with the slope of the rater characteristic curve. As consensus increases, the slope of the curve also increases, subsequently increasing the degree of nonlinearity in the response function. A linear model fits less well as the degree of nonlinearity increases; thus, the linear model will always underestimate consensus. As consensus within a group decreases, the slope decreases, making the rater characteristic

curve flatter, i.e., more linear. With this flattened slope, a linear model (such as *ICC(2, 1)*) will fit as well or better than nonlinear models.

$r_{WG(J)}$ and $r_{WG(J)_n}$ were the worst estimates (i.e., least accurate) of consensus across all conditions. Globally, the estimates from these indices are too high, drastically overestimating true consensus. In addition, the estimates are unstable and can show extreme effects of the number of raters and items. This is consistent with previous studies which suggest values of $r_{WG(J)}$ and $r_{WG(J)_n}$ depend on the number of items in the scale and the number of raters (e.g., Cohen, Doveh, & Eick, 2001; Lindell, 1999; James, et al, 1984). All of the remaining indices proved to be relatively stable across the rater and items conditions.

Results of the variance components analyses of RMSE and bias further elucidate the impact of the choice of agreement index and its effects on the resulting agreement estimates. For medium to high consensus, the primary source of variability in the average error of estimation was index. For low consensus, although results suggest index does not explain a substantial amount of the variability, these results must be evaluated in the context of all of the agreement indices. When true consensus was low (20%) the variability of estimates was reduced. Indices that had been consistently underestimating true consensus in the high and medium consensus conditions (e.g., κ_m and r_s) were much more similar when raters agreed less. Since the variability of the agreement estimates is less initially, naturally less of the variability in average deviation from true consensus is attributable to index.

Studies 2 and 3

As the goal of this study was to better understand the appropriate application of the various measures of consensus, the study was extended to include an investigation of two additional datasets (non-simulated data). The goals of the subsequent studies are three-fold: 1) to validate the results obtained from the simulated data on two different types of perception data, 2) to replicate the ordered patterns observed for the agreement indices in the simulated data, and 3) advance the understanding of the ramifications associated with choice of consensus measure.

Study 2

Method

Participants. The data used in this study were collected as part of a job analysis project (Reiter-Palmon, Clifton, Connelly, Uhlman, & Mumford, 1991). Surveys were sent to all salespeople at a national company. Data from one of two sales divisions were chosen for this analysis. A total of 108 surveys were returned by the salespeople in this division, a return rate of 52%. The information collected was anonymous, so no background information or identifying information was collected with the surveys.

Procedure. Participants (N = 108) took part in a task rating survey. In this survey, the participants were presented with a list of 161 tasks that were generated in meetings with subject matter experts. The participants rated each task on four dimensions: importance, relative time spent, criticality, and performance. The rating scales are shown in Appendix B.

The rating scales were comprised of Likert scale ratings with choices ranging from 0 (e.g., does not apply) to 5 (e.g., extremely important). For the dimensions of

importance and criticality a rating of zero was treated as missing data. A zero rating on these dimensions indicated a rater did not have knowledge of the particular task, and thus was not qualified to rate the item. In contrast, a zero rating on the dimensions of relative time spent and performance was deemed a meaningful alternative indicating lack of time spent or lack of importance of the task.

Analyses. Nine coefficients of interrater agreement were computed on the ratings for each of the four dimensions: κ_m, *ICC(2,1)*, $r_{WG(J)}$, $r_{WG(J)_n}$, $r^{\bullet}_{WG(J)}$, $r^{\bullet}_{WG(J)_n}$, r_s, CSI_n, and CSI_u. While it is impossible to evaluate the accuracy of the estimates in this data (because true agreement is unknown), we can look for similar patterns of results among the indices as compared to the simulation study and perhaps draw inferences about the true nature of agreement in this application.

Results and Conclusions

Table 10 presents the agreement estimates computed for the job analysis data. Using the simulation study results as a guide, I first examined the values for $r^{\bullet}_{WG(J)}$ and CSI_n. In general, all four of the rated dimensions appeared to have moderate levels of consensus in the task ratings. Similar relationships between $r^{\bullet}_{WG(J)}$ and CSI_n were observed as were seen in the simulation results. For example, $r^{\bullet}_{WG(J)}$ is higher than CSI_n across the four dimensions. Additionally, CSI_u estimates closely follow in value the CSI_n estimates.

Estimated values for consensus using $r^{\bullet}_{WG(J)}$ and CSI_n are identical for the importance dimension – suggesting true consensus among raters was, indeed, around 50%. However, across the remaining three dimensions the discrepancies between $r^{\bullet}_{WG(J)}$

and CSI_n increase reaching a maximum for the performance dimension. This creates some difficulty when trying to determine accuracy of estimation. Nonetheless, overall, consensus appears to be in the 40-60% range for each of the four dimensions.

The remaining indices were generally consistent with the simulation results assuming the normal latent trait distribution. First, $r_{WG(J)}$ and $r_{WG(J)_n}$ were high – inflated by the influence of the 161 tasks rated in the survey. Estimates of $r_{WG(J)}$ and $r_{WG(J)_n}$ ranged between .944 and .997 across the four dimensions.

Second, as observed in the simulations, κ_m, *ICC(2,1)*, and r_s underestimate consensus. These low estimates are most likely due to violations of relevant statistical assumptions. For example, poor estimates of κ_m are most likely due to its application to data that are not nominal. In the same way, the *ICC(2,1)* and r_s will tend to underestimate consensus as the data become increasingly nonlinear. If consensus is estimated at roughly 50% for this data, then a fair degree of nonlinearity is expected to be present in the data; thus, low estimates of *ICC(2,1)* and r_s are obtained.

From a more applied standpoint, the results of the consensus analysis are contrary to the findings of Lindell et al. (1998). This prior study found generally low values of $r^{\bullet}_{WG(J)}$ for task time spent (.19 - .38) and low to moderate values for task importance (.22 - .57); whereas, the current data suggested a moderate degree of consensus for both relative time spent ($r^{\bullet}_{WG(J)} = .652$) and importance ($r^{\bullet}_{WG(J)} = .513$). These discrepancies are most likely attributable to differences in the types of organizations being analyzed and variable job requirements and not to any important differences in actual data characteristics.

One concern that arises after examining the agreement estimates is the behavior of the consensus estimates for the performance dimension. It could be argued that agreement for this dimension should roughly approximate chance agreement. Ratings of performance are directed to be ratings of the raters' "own" performance. Thus, most of the agreement in the ratings should be attributable to chance. Following this line of reasoning, several of the estimates appear to be capitalizing on chance agreement in the task ratings. Specifically, $r_{WG(J)}$, $r_{WG(J)_n}$, $r^{*}_{WG(J)}$, and $r^{*}_{WG(J)_n}$ values for this dimension are the highest of the four dimensions.

It could also be argued, though, that some minimum threshold of performance is essential to maintaining employment. Thus, consensus in ratings might actually be expected to be high due to the restriction in the range of the responders and thus, their responses. Unfortunately, from the current analyses, the cause of the variable consensus estimates for the performance dimension remains unclear.

Study 3

Study 3 further extends the simulation study by examining agreement on nomination data. The data used come from a study that examined individual differences in person perception (Murphy, DeBacker, & Terry, 2001).

Method

Participants. A total of 106 (64 females; 42 males) fourth-, fifth-, and sixth-grade children participated in the study, ranging in age from 9 to 13 years ($\underline{M}$ = 11.05, $\underline{SD}$ = .86). The sample was relatively ethnically diverse with 37% Caucasian, 22% African American, 7% Hispanic, 6% Native American, and 28% Mixed Origin. Participants were recruited from 7 classrooms in two elementary schools in a mid-sized mid-western town.

Data from one representative classroom were selected for the current investigation. Fifteen (N=15) children from this classroom participated in the study.

Procedure. Data collection occurred in the spring of the school year. Participants were asked to complete several questionnaires through the course of the study. At the end of the study participants were asked to make judgments of peers. It is this judgment data that were analyzed for the current study.

Judgments of peers. To assess children's judgments of peers, children participated in a sociometric nomination procedure. Children nominated an unlimited number of classmates (see Terry, 2000) who fit the description of 12 characteristics (adapted from Coie & Dodge, 1988). Nominations from three of these statements were analyzed for the current study.

Specifically, children were asked to nominate peers who "...are leaders who are good to have in charge," "...are well-liked by many kids, " and "...stay by themselves and away from other kids." Children were read the relevant statement (e.g., "These kids are leaders who are good to have in charge") and marked selections on their class roster. Class rosters contained lists of participants' names with the various behavioral statements across the top. Children were told to make a mark in the box for each peer who fit the description and were allowed to nominate an unlimited number of classmates for each of the statements.

Analyses. Consensus estimates among the binary nominations on the three target characteristics were computed for each of the three trait nominations separately. Nine indices of consensus were computed: κ_m, ICC(2,1), $r_{WG(J)}$, $r_{WG(J)_n}$, $r^{*}_{WG(J)}$, $r^{*}_{WG(J)_n}$, r_s, CSI_n, and CSI_u. All parameter estimates necessary for the CSI_n and CSI_u were obtained

using BILOG 3 (Mislevy & Bock, 1990) which allows researchers to estimate the parameters of IRT models on questionnaire data with dichotomous item response formats.

Results and Conclusions

Table 11 presents the results of the agreement analyses. Higher consensus was found among judges for the shy/loner characteristic than for the more positive characteristics of leadership and likeability. Research suggests that negative characteristics may be particularly salient to children because they can threaten children's sense of well-being or indicate social deviance (Murphy et al., 2001). Additionally, withdrawal from social interaction, which is the behavior addressed by the shy/loner statement, is viewed as odd behavior during middle and late childhood. Children may attend more to peers who more frequently engage in such "undesirable" behaviors such that there is relatively high consensus among judges about who possesses these negative characteristics (Rubin, Bukowski, & Parker, 1998).

While generally the data do follow the same patterns as seen in both the simulation data and the job analysis data, there are some important departures. First, the discrepancies between the estimates of CSI_n and $r^*_{WG(J)}$ are large for the three trait characteristics. Further study is needed to determine whether this is an artifact of the binary data. In other words, future investigations should consider the effect of the number of scale points (i.e., response options) on agreement estimates.

Second, note that for binary data (A=2) the expected variance for the uniform distribution is $(A^2 - 1)/12 = 0.25$, and the expected variance for the normal distribution (when estimated using a triangular distribution) is $(A^2 + 2A - 2)/24 = 0.25$. Given that the

expected variances for the two different reference distributions are equivalent, $r_{WG(J)}$ and $r_{WG(J)_n}$ should yield equivalent values for the current data. Similarly, $r^{*}_{WG(J)}$ and $r^{*}_{WG(J)_n}$ should be equivalent. This is, in fact, the case, as seen in Table 11.

While this is a convenient property, it does leads to some concerns for the accuracy of $r^{*}_{WG(J)}$ in this context. Results obtained from Study 1 suggested $r^{*}_{WG(J)_n}$ was not a particularly good estimate of consensus. Yet, for binary data, values of $r^{*}_{WG(J)}$ are equivalent to $r^{*}_{WG(J)_n}$. Is it the case that $r^{*}_{WG(J)}$ is not as accurate when the number of response options is less than 5? Questions such as this remain to be resolved in future research.

General Discussion

The present examination of the statistical properties of nine measures of consensus contributes to an understanding of the estimates of agreement researchers can expect to obtain from some of the more commonly used measures of agreement. Specifically, the effects of factors such as the number of raters, the number of items, true distribution of θ, constant/varying slopes, and degree of true consensus on agreement estimates were examined. The initial study was extended to investigate two different types of perception data.

It is common for different literatures to cling to preferred, or traditional, measures for assessing consensus among raters. For example, the medical literature often uses variants of Cohen's kappa. The relevant data are often diagnosis classifications or data that are similarly nominal in scale type. $r_{WG(J)}$ and its variants are almost exclusively cited in the industrial/organizational psychology literature. Social psychologists and

sociologists frequently report *ICC*s. Regardless of discipline, it is important that estimates of consensus 1) measure what they say they are measuring and 2) accurately estimate true agreement among raters. While research on the measurement of consensus has proceeded within each of these literatures, it is rare to find research that attempts to bridge this expanse and incorporate many of these different measures of consensus into one comprehensive evaluation.

The results of the current investigation suggest the choice of consensus measure is important and will, no doubt, influence conclusions drawn about the degree of consensus among raters. Studies 1 and 2 revealed relatively consistent orderings for the magnitude of the consensus estimates across a number of factors. However, study 3 showed some important discrepancies that need to be investigated further. Of most importance for the current investigation is the unresolved issue of the impact of the number of scale points, or response options on consensus measures. Previous research has shown that values of $r_{WG(J)}$ computed on dichotomous data provide estimates of agreement that range from slightly higher to substantially lower than those for $r_{WG(J)}$ computed from polytomous ratings (Lindell & Brandt, 1999). The disparities were deemed to be attributable to coarse categorizations imposed by the dichotomous response. The same pattern was shown to hold for estimates obtained from $r^{*}_{WG(J)}$. While we did not explicitly look at the impact of varying the number of scale points, the current data combined with minimal previous research suggest that further investigation of this factor on each of the agreement estimates is warranted.

In subsequent research it will also be important to examine the impact of the choice of the reference distribution on consensus estimates. The present studies suggested

that some of the measures actually perform better when there is a mismatch between the assumed reference distribution and the observed distribution of ratings. Further investigation is needed to identify why indices do not perform better when the observed and expected distributions of ratings match.

Overall, the results lead me to conclude that $r^{*}_{WG(J)}$, CSI_{n} (for normal θ), and CSI_{u} (for uniform θ) are the most accurate estimates of true consensus. Much additional work remains to be done; however, in light of these considerations, it is clear that both $r^{*}_{WG(J)}$ and *CSI* consistently estimate true consensus with less error than the remaining indices. Regardless of the accuracy of the indices, for estimation it is important to consider the nature of the error in the estimates. The potential for overestimating consensus when using $r^{*}_{WG(J)}$ must be weighed against the desire to obtain estimates with the least possible error. For those who value erring on the conservative side, the *ICC(2,1)* or the *CSI* might be the better choice.

One potential criticism for the current findings is the argument that the estimates of the *CSI* might be biased due to the data simulation procedure. As previously described, the data were simulated using IRT models. As the *CSI* indices are also based in IRT, one might wonder whether they have a distinct advantage over the other agreement indices. While this was a concern initially, the results of Study 2 dampen this argument. The pattern of the relationships between the indices of agreement first established using the simulated data continue to exist even in a much different type of perception data – data that were collected from "real" raters, not generated using IRT models.

It is puzzling to ponder why $r^{*}_{WG(J)}$ is not as bad as the other consensus measures. The $r^{*}_{WG(J)}$ is an omnibus index without any parameters relating the agreement estimates to the rating process. $r^{*}_{WG(J)}$ is simply a function of the average observed variance in ratings in reference to the chosen null distribution. It is difficult to derive any plausible explanations for the observed results. Future research should investigate the mathematical and theoretical ties between $r^{*}_{WG(J)}$ and the model based indices (*CSI* and *ICC*) for potential explanations. Despite its accuracy, $r_{WG(J)}$ is not model based, thus estimates provide little information about individual raters and the ratings we would expect from them for a given item.

Overall, results suggest $r_{WG(J)}$ and $r_{WG(J)_n}$ are not good measures of consensus. These indices consistently overestimate true consensus due to the disproportionate influence of the number of items and raters on agreement estimates. As expected, the influence of sample size was most profound in the job analysis data which consisted of the most items. While less prominent, this effect is seen across all simulated conditions (with the number of items ranging from 10 to 50) and the sociometric data (with only 15 items).

Except for conditions of low consensus, the *ICC(2, 1)* is a less accurate index for obtaining agreement estimates. The *ICC(2, 1)* assumes the variance of the ratings is essentially the same across raters and targets. This model is not strictly applicable when some raters differentiate targets/items more than others (Shrout, 1993). As mentioned, however, for low consensus, *ICC(2, 1)* performs admirably.

Future research should also investigate the effect of varying the threshold parameters. As thresholds become less uniform throughout θ, the more the observed ratings for a given rater will deviate from symmetry – resulting in a "leniency" or "harshness" bias in the observed rating distributions. In other words, as the thresholds deviate from a uniform distribution, the variance of the observed ratings decreases. The *CSI* should be invariant to varying thresholds. However, the impact of the expected reduction in variance on measures intimately tied to the variance of the observed distribution of ratings remains to be seen.

A potential solution to the convergence difficulties with IRT parameter estimation did arise from this research. If multiple replications of data is not available, one approach to resolving the estimation problems is to use our "best guess" for what the prior slope estimates should be. Results of the current investigation suggest $r^{*}_{WG(J)}$ is an accurate estimate of consensus. As consensus is a function of the slope parameter, we could conceivably compute $r^{*}_{WG(J)}$ to get an estimate of the slope. This "prior" could then be incorporated into the parameter estimation phase allowing us to bypass the previous convergence problems encountered without placing priors on the slopes. Additionally, this new "prior" should not have the same adverse effect on the slope estimates that we encountered in the current study. Putting the two approaches together may give us even better estimates of consensus with much smaller deviation, at least for data with 5-point scales. At this point, it is still unclear whether $r^{*}_{WG(J)}$ accurately estimates consensus as the number of response categories becomes small.

Although there are multiple unresolved issues related to the measurement of consensus, the present findings illustrate the importance of studying the measurement of

consensus in rater responses. Unfortunately, the scope of the current investigation was already immense, preventing further factors from being included and more intently examined.

References

Agresti, A. (1992). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

Albright, L., Kenny, D.A., & Malloy, T.E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55(3),* 387-395.

Blackman, M.C. & Funder, D.C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology, 34,* 164-181.

Cohen, A, Doveh, E. & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods, 6(3),* 297-310.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Coie, J.J. & Dodge, K.A. (1988). Multiple sources of data on social behavior and social status in the school: A cross-age comparison. *Child Development, 59,* 815-829.

Conger, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88(2),* 322-328.

Conway III, L.G. & Schaller, M. (1998). Methods for the measurement of consensual beliefs within groups. *Group Dynamics: Theory, Research, and Practice, 2(4),* 241-252.

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

Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: LEA Publishers.

Fagot, R.F. (1993). A generalized family of coefficients of relational agreement for numerical scales. *Psychometrika, 58(2)*, 357-370.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions.* New York: John Wiley & Sons, 2nd ed.

Funder, D.C. & Colvin, C.R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55(1)*, 149-158.

Fundar, D.C., Kolar, D.C., & Blackman, M.C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69(4)*, 656-672.

Goodwin, L.D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5(1)*, 13-34.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Norwell, MA: Kluwer-Nijhoff Publishing.

James, L.R., Demaree, R.G. & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69(1)*, 85-98.

James, L.R., Demaree, R.G. & Wolf, G. (1993). r_{WG}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78(2),* 306-309.

Janson, H. & Olsson, U. (2001). A measure of agreement for interval of nominal multivariate observations. *Educational and Psychological Measurement, 61(2),* 277-289.

John, O.P. & Robins, R.W. (1993). Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61(4),* 521-551.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review, 98(1),* 54-73.

Kenny, D.A., Albright, L., Malloy, T.E., & Kashy, D.A. (1994). Consensus in interpersonal perception: Acquaintance and the big five. *Psychological Bulletin, 116(2),* 245-258.

Kenny, D.A. (1993). A coming-of-age for research on interpersonal perception. *Journal of Personality, 61(4),* 789-807.

Kenny, D.A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98(2),* 155-163.

Kozlowski, S.W.J. & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77(2),* 161-167.

Light, R.J. (1971), Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76(5)* 365-377.

Lindell, M.K. (2001). Assessing and testing interrater agreement on a single target using multi-item rating scales. *Applied Psychological Measurement, 25(1),* 89-99.

Lindell, M.K. & Brandt, C.J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the CVI, T, $r_{WG(J)}$, and $r^{*}_{WG(J)}$ indexes. *Journal of Applied Psychology, 84(4)*, 640-647.

Lindell, M.K., Brandt, C.J. & Whitney, D.J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23(2),* 127-135.

Lindell, M.K., Clause, C.S., Brandt, C.J. & Landis, R.J. (1998). Relationship between organizational context and job analysis task ratings. *Journal of Applied Psychology, 83(5)*, 769-776.

Mislevy, R.J. & Bock, R.D. (1990). *BILOG-3; Item analysis and test scoring with binary logistic models* [Computer software]. Mooresville, IN: Scientific Software.

Mitchell, S.K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86(2),* 376-390.

Muraki, E. & Bock, R.R. (1997). *PARSCALE 3.0: IRT based item analysis and test scoring for rating-scale data.* Chicago, IL: Scientific Software, Inc.

Murphy, B.C., DeBacker, T. K., Terry, R. (Submitted). "I Don't See Them That Way": Individual Differences in Children's Judgments of Peers. *Developmental Psychology.*

Park, B., Kraus, S., & Ryan, C.S. (1997). Longitudinal change in consensus as a function of acquaintance and agreement in liking. *Journal of Personality and Social Psychology, 72(3),* 604-616.

Rasch, G. (1966). An item analysis that takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49-57.

Rasch, G. (1977). On specific objectivity: An attempt at formulating the request for generality and validity of scientific statement. In M. Gliegvad (Ed.), *The Danish Yearbook of Philosophy* (pp. 58-94). Copenhagen: Munksgaard.

Reiter-Palmon, R., Clifton, T., Connelly, M.S., Uhlman, C.E., & Mumford, M.D. (1991). *Describing Sales Position Requirements: The Job Analysis* (CBCS Report 91-1 Vol.1). Fairfax, VA: General Electric Lighting Division.

Rubin, K.H., Bukowski, W., & Parker, J.G. (1998). Peer interactions, relationships, and groups. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (pp. 619-700). New York: Wiley.

SAS Institute, Inc. (1990). SAS/STAT user's guide: Version 6 (4th ed.). Cary, NC: Author.

Schmidt, F. L. & Hunter, J.E. (1989). Interobserver reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74(2),* 368-370.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage Publications, Inc.

Shrout, P.E. & Fleiss, J.L (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2),* 420-428.

Shrout, P.E. (1993). Analyzing consensus in personality judgments: A variance components approach. *Journal of Personality, 61(4),* 769-788.

Stine, W.W. (1989). Interobserver relational agreement. *Psychological Bulletin, 106(2),* 341-347.

Stine, W.W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin, 105(1),* 147-155.

Tanner, M.A. & Young, M.A. (1985). Modeling ordinal scale disagreement. *Psychological Bulletin, 98(2),* 408-415.

Terry, R. (2000). Recent advances in measurement theory and the use of sociometric techniques. In A.H.N. Cillessen & W.M. Bukowski (Eds.), Recent advances in the measurement of acceptance and rejection in the peer system. New Directions for Child and Adolescent Development, No. 88. San Francisco: Josey Bass.

Terry, R. (1995). *A latent trait model of interpersonal perception: Implications for sociometric assessment.* Unpublished manuscript.

Uebersax, J.S. (1988). Validity inferences from interobserver agreement, *Psychological Bulletin, 104(3),* 405-416.

Uebersax, J.S. (2002). *Statistical methods for rater agreement.* [online]. Available: http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm

Velleman, P.F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47(1),* 65-72.

Zegers, F.E. & ten Berge, J.M.F. (1985). A family of association coefficients for metric scales, *Psychometrika, 50(1),* 17-24.

Zegers, F.E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement, 15(4),* 321-333.

Zwick, R. (1988). Another look at interrater agreement, *Psychological Bulletin, 103(3),* 374-378.

Appendix A

Table A1

Means and Standard Deviations of Agreement Estimates

Condition[a]	κ_m	*SD*	r_s	*SD*	$r_{WG(J)}$	*SD*	*ICC*	*SD*	CSI_a	*SD*	CSI_u	*SD*	$r^*_{WG(J)}$	*SD*	$r_{wg(J)_a}$	*SD*	$r^*_{wg(J)_a}$	*SD*
Normal θ Distribution																		
1 1 1 1	0.22	0.23	0.57	0.24	0.98	0.01	0.59	0.24	0.58	0.07	0.48	0.08	0.83	0.06	0.96	0.02	0.74	0.10
1 1 1 2	0.02	0.22	0.33	0.36	0.92	0.06	0.31	0.30	0.51	0.07	0.40	0.10	0.58	0.17	1.61	2.42	0.36	0.26
1 1 1 3	0.02	0.12	0.26	0.30	1.77	2.93	0.26	0.32	0.48	0.09	0.33	0.12	0.20	0.29	2.15	2.37	-0.20	0.44
1 1 2 1	0.35	0.18	0.69	0.10	0.99	0.00	0.64	0.15	0.68	0.08	0.47	0.10	0.84	0.05	0.99	0.00	0.76	0.07
1 1 2 2	0.06	0.13	0.41	0.18	0.96	0.02	0.37	0.13	0.49	0.09	0.27	0.07	0.53	0.12	1.32	1.29	0.29	0.18
1 1 2 3	0.09	0.10	0.22	0.19	0.94	0.29	0.22	0.12	0.42	0.07	0.21	0.04	0.14	0.20	1.13	0.27	-0.29	0.31
1 1 3 1	0.32	0.11	0.69	0.10	1.00	0.00	0.70	0.09	0.73	0.08	0.48	0.13	0.84	0.04	0.99	0.00	0.76	0.07
1 1 3 2	0.14	0.09	0.40	0.10	0.98	0.01	0.43	0.13	0.49	0.07	0.23	0.05	0.55	0.08	0.93	0.11	0.33	0.12
1 1 3 3	0.05	0.08	0.19	0.16	0.94	0.18	0.20	0.05	0.37	0.08	0.16	0.05	0.10	0.18	1.07	0.09	-0.36	0.27
1 2 1 1	0.30	0.13	0.65	0.14	0.98	0.00	0.70	0.11	0.63	0.05	0.54	0.06	0.85	0.04	0.97	0.01	0.78	0.05
1 2 1 2	0.09	0.08	0.36	0.13	0.91	0.03	0.40	0.17	0.51	0.04	0.37	0.04	0.51	0.10	0.71	0.21	0.26	0.15
1 2 1 3	0.09	0.06	0.21	0.13	0.17	0.87	0.23	0.17	0.47	0.03	0.34	0.02	0.11	0.13	1.74	1.05	-0.34	0.19
1 2 2 1	0.31	0.06	0.65	0.07	0.99	0.00	0.66	0.08	0.69	0.04	0.50	0.06	0.84	0.04	0.99	0.00	0.77	0.05
1 2 2 2	0.17	0.05	0.49	0.08	0.97	0.01	0.50	0.07	0.55	0.04	0.31	0.04	0.60	0.06	0.94	0.03	0.40	0.09
1 2 2 3	0.04	0.05	0.15	0.07	0.77	2.02	0.15	0.07	0.38	0.04	0.18	0.03	0.05	0.09	1.17	0.05	-0.43	0.14
1 2 3 1	0.36	0.04	0.69	0.04	1.00	0.00	0.71	0.06	0.76	0.03	0.51	0.05	0.85	0.01	0.99	0.00	0.77	0.02
1 2 3 2	0.15	0.03	0.45	0.05	0.99	0.00	0.45	0.09	0.54	0.03	0.24	0.03	0.59	0.05	0.97	0.01	0.39	0.07
1 2 3 3	0.05	0.03	0.18	0.06	0.76	0.17	0.17	0.05	0.33	0.04	0.12	0.02	0.09	0.06	1.09	0.03	-0.37	0.09
1 3 1 1	0.33	0.11	0.69	0.10	0.98	0.00	0.69	0.10	0.62	0.03	0.55	0.04	0.84	0.04	0.97	0.01	0.76	0.05
1 3 1 2	0.14	0.06	0.43	0.13	0.93	0.01	0.40	0.10	0.56	0.09	0.49	0.04	0.58	0.02	0.85	0.02	0.37	0.04
1 3 1 3	0.06	0.02	0.19	0.05	0.39	0.22	0.19	0.16	0.57	0.18	0.46	0.00	0.08	0.07	1.76	0.70	-0.39	0.10
1 3 2 1	0.35	0.04	0.70	0.06	0.99	0.00	0.70	0.03	0.66	0.03	0.54	0.05	0.85	0.01	0.99	0.00	0.77	0.02
1 3 2 2	0.15	0.05	0.45	0.08	0.97	0.00	0.45	0.05	0.53	0.03	0.41	0.02	0.59	0.03	0.94	0.01	0.38	0.04
1 3 2 3	0.05	0.02	0.17	0.05	0.69	0.11	0.18	0.06	0.43	0.02	0.38	0.00	0.09	0.04	1.18	0.03	-0.37	0.06
1 3 3 1	0.35	0.05	0.68	0.06	1.00	0.00	0.69	0.06	0.70	0.03	0.53	0.05	0.85	0.02	0.99	0.00	0.77	0.02
1 3 3 2	0.14	0.03	0.43	0.06	0.99	0.00	0.43	0.06	0.52	0.03	0.33	0.02	0.58	0.03	0.97	0.01	0.37	0.04
1 3 3 3	0.05	0.02	0.17	0.04	0.82	0.08	0.17	0.04	0.37	0.02	0.26	0.02	0.10	0.03	1.08	0.01	-0.36	0.05

Condition[a]	κ_m	SD	r_s	SD	$r_{WG(J)}$	SD	ICC	SD	CSI_n	SD	CSI_u	SD	$r^*_{WG(J)}$	SD	$r_{wg(J)_n}$	SD	$r^*_{wg(J)_n}$	SD
2 1 1 1	0.20	0.25	0.63	0.21	0.97	0.01	0.56	0.21	0.58	0.06	0.48	0.08	0.80	0.09	0.98	0.09	0.50	0.62
2 1 1 2	0.16	0.27	0.44	0.31	0.93	0.06	0.41	0.33	0.54	0.10	0.42	0.13	0.61	0.22	0.72	0.36	0.42	0.33
2 1 1 3	-0.08	0.14	0.01	0.42	2.68	3.88	0.00	0.37	0.42	0.08	0.27	0.09	0.04	0.34	1.18	0.48	-0.45	0.51
2 1 2 1	0.31	0.15	0.67	0.13	0.99	0.00	0.57	0.19	0.66	0.09	0.46	0.11	0.84	0.06	0.99	0.01	0.76	0.10
2 1 2 2	0.10	0.12	0.45	0.18	0.97	0.02	0.36	0.16	0.53	0.09	0.30	0.08	0.62	0.13	0.84	0.37	0.43	0.19
2 1 2 3	0.08	0.09	0.27	0.22	0.83	0.31	0.10	0.15	0.44	0.08	0.22	0.06	0.22	0.28	0.84	0.33	-0.17	0.42
2 1 3 1	0.37	0.12	0.72	0.07	1.00	0.00	0.74	0.08	0.76	0.06	0.52	0.09	0.87	0.03	1.00	0.00	0.80	0.05
2 1 3 2	0.11	0.06	0.39	0.12	0.99	0.01	0.39	0.12	0.48	0.08	0.22	0.05	0.59	0.10	0.94	0.10	0.38	0.15
2 1 3 3	0.06	0.10	0.20	0.14	0.80	0.41	0.19	0.07	0.37	0.07	0.15	0.04	0.13	0.15	1.15	0.10	-0.31	0.23
2 2 1 1	0.26	0.15	0.58	0.18	0.98	0.01	0.66	0.13	0.60	0.06	0.48	0.06	0.82	0.04	1.00	0.12	0.60	0.43
2 2 1 2	0.12	0.07	0.37	0.16	0.92	0.02	0.43	0.16	0.51	0.05	0.37	0.06	0.53	0.07	0.78	0.08	0.29	0.10
2 2 1 3	-0.00	0.06	0.10	0.18	1.22	1.59	0.27	0.23	0.43	0.05	0.31	0.04	0.01	0.19	0.56	3.06	-0.50	0.30
2 2 2 1	0.32	0.05	0.69	0.07	0.99	0.00	0.69	0.06	0.70	0.04	0.50	0.06	0.85	0.01	0.99	0.00	0.77	0.02
2 2 2 2	0.12	0.06	0.42	0.12	0.97	0.01	0.42	0.11	0.51	0.06	0.27	0.04	0.55	0.06	0.92	0.04	0.33	0.09
2 2 2 3	0.07	0.06	0.20	0.10	13.90	41.23	0.20	0.10	0.39	0.05	0.19	0.03	0.11	0.12	1.27	0.21	-0.34	0.18
2 2 3 1	0.32	0.04	0.67	0.07	1.00	0.00	0.66	0.09	0.73	0.04	0.47	0.06	0.84	0.02	0.99	0.00	0.76	0.02
2 2 3 2	0.14	0.03	0.46	0.06	0.99	0.00	0.39	0.07	0.54	0.05	0.24	0.03	0.62	0.08	0.97	0.01	0.39	0.07
2 2 3 3	0.07	0.03	0.20	0.05	0.71	0.15	0.15	0.05	0.34	0.04	0.12	0.02	0.07	0.05	1.08	0.02	-0.40	0.08
2 3 1 1	0.36	0.06	0.73	0.08	0.98	0.00	0.74	0.07	0.64	0.02	0.59	0.03	0.79	0.12	1.02	0.17	0.65	0.38
2 3 1 2	0.12	0.06	0.39	0.12	0.93	0.01	0.40	0.11	0.52	0.02	0.47	0.01	0.58	0.05	0.85	0.04	0.38	0.07
2 3 1 3	0.05	0.02	0.19	0.07	0.18	0.71	0.26	0.28	0.49	0.01	0.47	0.01	0.08	0.09	1.68	0.32	-0.39	0.13
2 3 2 1	0.34	0.07	0.65	0.08	0.99	0.00	0.68	0.06	0.64	0.04	0.51	0.07	0.85	0.01	0.99	0.00	0.77	0.02
2 3 2 2	0.12	0.02	0.38	0.07	0.97	0.00	0.46	0.06	0.50	0.03	0.40	0.01	0.57	0.03	0.93	0.01	0.35	0.04
2 3 2 3	0.05	0.02	0.15	0.05	0.61	0.17	0.20	0.05	0.43	0.01	0.38	0.00	0.07	0.04	1.17	0.02	-0.40	0.06
2 3 3 1	0.36	0.05	0.70	0.05	1.00	0.00	0.70	0.06	0.71	0.03	0.54	0.04	0.84	0.01	0.99	0.00	0.77	0.01
2 3 3 2	0.14	0.03	0.42	0.05	0.99	0.00	0.43	0.05	0.52	0.03	0.33	0.02	0.57	0.02	0.97	0.00	0.36	0.03
2 3 3 3	0.06	0.02	0.17	0.05	0.77	0.10	0.17	0.05	0.37	0.02	0.26	0.02	0.08	0.05	1.08	0.02	-0.38	0.07
Uniform θ Distribution																		
1 1 1 1	0.47	0.17	0.88	0.07	0.99	0.01	0.90	0.06	0.68	0.03	0.59	0.05	0.88	0.07	0.98	0.02	0.82	0.10
1 1 1 2	0.37	0.20	0.83	0.07	0.97	0.02	0.80	0.11	0.64	0.04	0.53	0.07	0.76	0.13	0.94	0.05	0.64	0.20

Condition[a]	κ_m	SD	r_t	SD	$r_{WG(J)}$	SD	ICC	SD	CSI_n	SD	CSI_w	SD	$r^*_{WG(J)}$	SD	$r_{wg(J)_n}$	SD	$r^*_{wg(J)_n}$	SD
1 1 1 3	0.14	0.20	0.39	0.22	0.28	4.81	0.39	0.21	0.51	0.07	0.37	0.09	0.19	0.29	1.46	0.95	-0.21	0.44
1 1 2 1	0.56	0.08	0.89	0.04	0.99	0.00	0.88	0.04	0.82	0.04	0.63	0.06	0.88	0.04	0.99	0.00	0.82	0.07
1 1 2 2	0.31	0.14	0.73	0.10	0.98	0.01	0.75	0.09	0.68	0.08	0.43	0.08	0.70	0.10	0.96	0.02	0.55	0.15
1 1 2 3	0.15	0.12	0.41	0.14	0.82	0.16	0.39	0.17	0.50	0.07	0.26	0.04	0.25	0.19	1.16	0.25	-0.13	0.28
1 1 3 1	0.58	0.09	0.91	0.03	1.00	0.00	0.90	0.05	0.89	0.03	0.68	0.06	0.89	0.03	1.00	0.00	0.84	0.03
1 1 3 2	0.29	0.07	0.68	0.06	0.99	0.00	0.71	0.09	0.69	0.05	0.36	0.05	0.65	0.06	0.98	0.01	0.48	0.09
1 1 3 3	0.12	0.06	0.34	0.08	-4.25	16.08	0.40	0.07	0.44	0.06	0.19	0.04	0.13	0.11	0.71	1.26	-0.31	0.16
1 2 1 1	0.52	0.11	0.86	0.07	0.99	0.01	0.89	0.04	0.72	0.03	0.61	0.04	0.86	0.05	0.98	0.01	0.81	0.07
1 2 1 2	0.33	0.10	0.75	0.10	0.96	0.02	0.73	0.09	0.64	0.05	0.51	0.07	0.70	0.10	0.91	0.07	0.54	0.16
1 2 1 3	0.13	0.10	0.40	0.17	0.51	0.75	0.42	0.13	0.53	0.06	0.39	0.06	0.25	0.19	0.03	2.94	-0.12	0.28
1 2 2 1	0.55	0.08	0.88	0.05	0.99	0.00	0.90	0.05	0.83	0.04	0.66	0.06	0.89	0.03	0.99	0.00	0.83	0.05
1 2 2 2	0.32	0.08	0.73	0.04	0.98	0.00	0.72	0.04	0.69	0.04	0.43	0.06	0.68	0.05	0.96	0.01	0.52	0.07
1 2 2 3	0.13	0.05	0.35	0.11	0.84	0.12	0.37	0.11	0.48	0.05	0.24	0.03	0.23	0.11	0.94	0.94	-0.15	0.17
1 2 3 1	0.54	0.03	0.90	0.01	1.00	0.00	0.89	0.04	0.88	0.01	0.68	0.03	0.88	0.01	1.00	0.00	0.83	0.02
1 2 3 2	0.30	0.04	0.71	0.04	0.99	0.00	0.73	0.05	0.72	0.03	0.39	0.04	0.67	0.04	0.98	0.00	0.51	0.06
1 2 3 3	0.14	0.04	0.37	0.05	0.92	0.03	0.44	0.06	0.46	0.05	0.18	0.03	0.21	0.06	1.19	0.13	-0.19	0.09
1 3 1 1	0.56	0.09	0.89	0.05	0.99	0.00	0.91	0.03	0.68	0.02	0.63	0.01	0.89	0.03	0.98	0.01	0.83	0.05
1 3 1 2	0.31	0.11	0.70	0.10	0.95	0.01	0.68	0.06	0.59	0.03	0.50	0.03	0.67	0.06	0.90	0.03	0.50	0.10
1 3 1 3	0.14	0.03	0.40	0.08	0.75	0.07	0.34	0.08	0.52	0.02	0.47	0.01	0.24	0.07	-1.81	10.38	-0.14	0.11
1 3 2 1	0.55	0.07	0.89	0.03	0.99	0.00	0.90	0.01	0.77	0.02	0.65	0.02	0.88	0.03	0.99	0.00	0.82	0.04
1 3 2 2	0.33	0.05	0.73	0.05	0.98	0.00	0.71	0.04	0.64	0.03	0.48	0.03	0.70	0.03	0.97	0.01	0.54	0.04
1 3 2 3	0.18	0.04	0.44	0.06	0.90	0.03	0.37	0.05	0.51	0.02	0.40	0.01	0.27	0.06	1.00	1.22	-0.10	0.09
1 3 3 1	0.54	0.05	0.88	0.03	1.00	0.00	0.88	0.02	0.80	0.02	0.66	0.03	0.88	0.01	1.00	0.00	0.81	0.02
1 3 3 2	0.32	0.02	0.71	0.03	0.99	0.00	0.72	0.03	0.66	0.01	0.42	0.02	0.68	0.02	0.98	0.00	0.52	0.02
1 3 3 3	0.16	0.02	0.39	0.04	0.94	0.01	0.39	0.04	0.49	0.02	0.31	0.01	0.23	0.03	1.20	0.07	-0.15	0.05
2 1 1 1	0.38	0.25	0.81	0.18	0.98	0.01	0.82	0.15	0.67	0.07	0.57	0.09	0.87	0.05	0.97	0.01	0.80	0.08
2 1 1 2	0.19	0.10	0.55	0.19	0.86	0.19	0.54	0.19	0.56	0.06	0.41	0.07	0.51	0.23	0.82	0.30	0.26	0.35
2 1 1 3	0.06	0.13	0.34	0.37	2.08	2.92	0.30	0.35	0.50	0.11	0.36	0.13	0.22	0.36	1.37	0.79	-0.17	0.54
2 1 2 1	0.62	0.13	0.90	0.04	1.00	0.00	0.89	0.05	0.83	0.04	0.66	0.07	0.90	0.04	0.99	0.00	0.84	0.07
2 1 2 2	0.28	0.11	0.70	0.07	0.98	0.00	0.69	0.13	0.66	0.05	0.50	0.03	0.66	0.06	0.96	0.01	0.49	0.08
2 1 2 3	0.12	0.08	0.35	0.15	0.70	0.61	0.41	0.15	0.48	0.13	0.20	0.08	0.17	0.21	0.99	0.55	-0.25	0.32
2 1 3 1	0.55	0.04	0.88	0.04	1.00	0.00	0.87	0.07	0.87	0.03	0.64	0.05	0.88	0.03	1.00	0.00	0.82	0.04

Condition[a]	κ_m	SD	r_s	SD	$r_{WG(J)}$	SD	ICC	SD	CSI_n	SD	CSI_u	SD	$r^*_{WG(J)}$	SD	$r_{wg(J)_n}$	SD	$r^*_{wg(J)_n}$	SD
2 1 3 2	0.34	0.09	0.68	0.09	0.99	0.00	0.73	0.05	0.70	0.07	0.40	0.10	0.64	0.09	0.97	0.02	0.45	0.14
2 1 3 3	0.17	0.13	0.38	0.19	-4.19	16.10	0.42	0.12	0.48	0.13	0.20	0.08	0.20	0.22	0.97	0.20	-0.20	0.34
2 2 1 1	0.50	0.12	0.90	0.05	0.99	0.01	0.90	0.04	0.73	0.03	0.62	0.05	0.88	0.05	0.98	0.01	0.81	0.08
2 2 1 2	0.27	0.11	0.68	0.15	0.94	0.03	0.71	0.06	0.63	0.05	0.49	0.06	0.63	0.12	0.85	0.14	0.45	0.18
2 2 1 3	0.13	0.06	0.41	0.16	0.72	0.21	0.36	0.15	0.53	0.06	0.39	0.06	0.27	0.15	-1.17	6.69	-0.10	0.22
2 2 2 1	0.54	0.10	0.87	0.05	0.99	0.00	0.88	0.05	0.82	0.04	0.65	0.06	0.88	0.03	0.99	0.00	0.82	0.04
2 2 2 2	0.35	0.08	0.72	0.07	0.98	0.00	0.74	0.06	0.71	0.04	0.47	0.06	0.70	0.06	0.97	0.01	0.55	0.09
2 2 2 3	0.18	0.07	0.43	0.12	0.88	0.11	0.43	0.12	0.53	0.07	0.29	0.06	0.29	0.13	1.29	1.08	-0.07	0.19
2 2 3 1	0.55	0.04	0.89	0.01	1.00	0.00	0.89	0.03	0.87	0.02	0.66	0.04	0.88	0.01	1.00	0.00	0.81	0.02
2 2 3 2	0.33	0.06	0.72	0.05	0.99	0.00	0.73	0.05	0.73	0.05	0.40	0.06	0.69	0.06	0.98	0.01	0.53	0.09
2 2 3 3	0.16	0.03	0.39	0.06	0.93	0.03	0.41	0.07	0.48	0.04	0.19	0.03	0.22	0.06	1.26	0.25	-0.17	0.09
2 3 1 1	0.51	0.08	0.84	0.07	0.99	0.00	0.88	0.04	0.66	0.03	0.61	0.03	0.88	0.02	0.98	0.00	0.81	0.03
2 3 1 2	0.33	0.05	0.72	0.07	0.95	0.01	0.68	0.09	0.60	0.03	0.51	0.04	0.68	0.05	0.91	0.03	0.52	0.08
2 3 1 3	0.16	0.05	0.42	0.13	0.71	0.13	0.38	0.18	0.52	0.02	0.47	0.01	0.24	0.14	1.69	5.26	-0.15	0.21
2 3 2 1	0.54	0.07	0.89	0.04	0.99	0.00	0.89	0.01	0.76	0.02	0.64	0.02	0.88	0.02	0.99	0.00	0.82	0.03
2 3 2 2	0.33	0.04	0.73	0.05	0.98	0.00	0.73	0.05	0.63	0.02	0.45	0.02	0.68	0.03	0.96	0.01	0.51	0.04
2 3 2 3	0.17	0.04	0.41	0.07	0.88	0.05	0.39	0.06	0.51	0.03	0.40	0.01	0.25	0.08	1.18	0.76	-0.13	0.11
2 3 3 1	0.54	0.05	0.88	0.03	1.00	0.00	0.89	0.03	0.81	0.02	0.66	0.03	0.87	0.02	1.00	0.00	0.81	0.03
2 3 3 2	0.35	0.12	0.70	0.03	0.99	0.00	0.70	0.03	0.65	0.02	0.42	0.02	0.62	0.10	0.98	0.00	0.49	0.04
2 3 3 3	0.17	0.02	0.40	0.04	0.93	0.01	0.40	0.04	0.49	0.02	0.31	0.00	0.23	0.04	1.21	0.10	-0.16	0.06

Note. The values represent means (and standard deviations) across the ten replications of each simulated condition.

[a]*Condition* represents the combinations of the various factors manipulated in the study. The following coding scheme is used - Column 1: Slopes. 1 =Constant, 2=Varying; Column 2: No. of raters. 1= 2 raters, 2 = 5 raters, 3 = 25 raters; Column 3: No. of items. 1 = 10 items, 2 = 25 items, 3 = 50 items; Column 4: True Consensus. 1 = 80%, 2 = 50%, 3 = 20%.

Table A2

RMSE for estimates of agreement versus true consensus

Condition[a]	κ_m	r_s	$r_{WG(J)}$	*ICC*	CSI_n	CSI_u	$\dot{r}_{WG(J)}$	$r_{WG(J)_n}$	$\dot{r}_{WG(J)_n}$
Normal θ Distribution									
1 1 1	0.63	0.29	0.18	0.31	0.23	0.33	0.08	0.18	0.38
1 1 2	0.47	0.34	0.43	0.34	0.09	0.14	0.21	1.48	0.30
1 1 3	0.26	0.37	3.81	0.22	0.26	0.14	0.32	2.03	0.69
1 2 1	0.49	0.17	0.19	0.19	0.15	0.35	0.07	0.19	0.09
1 2 2	0.44	0.19	0.47	0.26	0.09	0.22	0.14	0.98	0.23
1 2 3	0.15	0.20	0.74	0.25	0.24	0.05	0.24	0.84	0.56
1 3 1	0.47	0.12	0.20	0.15	0.09	0.32	0.07	0.19	0.06
1 3 2	0.38	0.15	0.48	0.17	0.07	0.28	0.11	0.45	0.20
1 3 3	0.17	0.14	0.74	0.25	0.19	0.06	0.18	0.91	0.58
2 1 1	0.54	0.24	0.18	0.37	0.19	0.30	0.05	0.20	0.26
2 1 2	0.40	0.19	0.41	0.26	0.05	0.14	0.08	0.29	0.26
2 1 3	0.17	0.16	1.32	0.16	0.25	0.13	0.21	2.38	0.66
2 2 1	0.48	0.15	0.19	0.19	0.11	0.31	0.05	0.19	0.05
2 2 2	0.36	0.11	0.47	0.09	0.06	0.21	0.10	0.43	0.16
2 2 3	0.16	0.09	21.72	0.09	0.19	0.03	0.16	1.03	0.61
2 3 1	0.47	0.13	0.20	0.12	0.07	0.32	0.05	0.19	0.04
2 3 2	0.35	0.07	0.49	0.06	0.05	0.26	0.10	0.47	0.13
2 3 3	0.14	0.06	0.56	0.04	0.14	0.08	0.13	0.88	0.59
3 1 1	0.46	0.12	0.18	0.11	0.18	0.23	0.08	0.22	0.23
3 1 2	0.38	0.15	0.43	0.15	0.07	0.04	0.09	0.35	0.14
3 1 3	0.15	0.06	0.48	0.06	0.35	0.27	0.14	1.60	0.60
3 2 1	0.46	0.14	0.19	0.14	0.15	0.28	0.05	0.19	0.04
3 2 2	0.37	0.11	0.47	0.11	0.03	0.09	0.08	0.43	0.14
3 2 3	0.15	0.06	0.47	0.05	0.23	0.18	0.12	0.97	0.58
3 3 1	0.45	0.12	0.20	0.13	0.10	0.27	0.05	0.19	0.04
3 3 2	0.36	0.09	0.49	0.09	0.04	0.17	0.08	0.47	0.14
3 3 3	0.15	0.05	0.61	0.04	0.17	0.06	0.12	0.88	0.57
Uniform θ Distribution									
1 1 1	0.43	0.14	0.19	0.13	0.14	0.23	0.09	0.18	0.08
1 1 2	0.28	0.26	0.43	0.25	0.12	0.09	0.26	0.43	0.32
1 1 3	0.19	0.33	3.96	0.31	0.32	0.19	0.31	1.47	0.61
1 2 1	0.24	0.10	0.19	0.10	0.04	0.16	0.10	0.19	0.07
1 2 2	0.24	0.23	0.48	0.23	0.18	0.07	0.20	0.46	0.11
1 2 3	0.12	0.23	0.70	0.23	0.30	0.08	0.19	0.97	0.48
1 3 1	0.24	0.10	0.20	0.10	0.08	0.15	0.09	0.20	0.05
1 3 2	0.20	0.19	0.49	0.19	0.20	0.14	0.16	0.48	0.12
1 3 3	0.11	0.21	15.89	0.21	0.28	0.06	0.17	1.04	0.52
2 1 1	0.31	0.10	0.19	0.10	0.08	0.19	0.09	0.18	0.07
2 1 2	0.22	0.25	0.45	0.24	0.14	0.06	0.20	0.40	0.17

Condition[a]	κ_m	r_s	$r_{WG(J)}$	ICC	CSI_n	CSI_u	$\dot{r}_{WG(J)}$	$r_{WG(J)_n}$	$\dot{r}_{WG(J)_n}$
2 1 3	0.11	0.26	0.67	0.25	0.33	0.20	0.17	4.65	0.39
2 2 1	0.27	0.09	0.19	0.10	0.04	0.16	0.09	0.19	0.05
2 2 2	0.18	0.23	0.48	0.24	0.20	0.08	0.20	0.46	0.08
2 2 3	0.08	0.22	0.67	0.23	0.31	0.08	0.13	1.33	0.36
2 3 1	0.25	0.09	0.20	0.10	0.08	0.14	0.08	0.20	0.03
2 3 2	0.19	0.22	0.49	0.23	0.23	0.12	0.19	0.48	0.07
2 3 3	0.06	0.19	0.73	0.19	0.27	0.03	0.06	1.04	0.39
3 1 1	0.27	0.09	0.19	0.10	0.13	0.18	0.09	0.18	0.04
3 1 2	0.20	0.23	0.45	0.23	0.10	0.03	0.18	0.41	0.08
3 1 3	0.06	0.23	0.54	0.23	0.32	0.27	0.11	7.63	0.38
3 2 1	0.26	0.09	0.19	0.10	0.04	0.15	0.08	0.19	0.04
3 2 2	0.17	0.23	0.48	0.24	0.14	0.05	0.19	0.47	0.05
3 2 3	0.05	0.23	0.69	0.24	0.31	0.20	0.09	1.31	0.33
3 3 1	0.26	0.09	0.20	0.09	0.02	0.14	0.08	0.20	0.03
3 3 2	0.18	0.21	0.49	0.21	0.16	0.08	0.17	0.48	0.03
3 3 3	0.04	0.20	0.74	0.20	0.29	0.11	0.05	1.01	0.36

[a]*Condition* represents the combinations of the various factors manipulated in the study. The following coding scheme is used: Column 1: No. of raters. 1 = 2 raters, 2 = 5 raters, 3 = 25 raters; Column 2: No. of items. 1 = 10 items, 2 = 25 items, 3 = 50 items; Column 3: True Consensus. 1 = 80%, 2 = 50%, 3 = 20%.

Table A3

Bias for estimates of agreement versus true consensus

Condition[a]	κ_m	r_s	$r_{WG(J)}$	*ICC*	CSI_n	CSI_u	$r^{*}_{WG(J)}$	$r_{WG(J)_a}$	$r^{*}_{WG(J)_a}$
Normal θ Distribution									
1 1 1	-0.59	-0.20	0.18	-0.22	-0.22	-0.32	0.01	0.17	-0.18
1 1 2	-0.41	-0.12	0.42	-0.14	0.03	-0.09	0.09	0.67	-0.11
1 1 3	-0.23	-0.06	2.02	-0.07	0.25	0.10	-0.05	1.47	-0.52
1 2 1	-0.47	-0.12	0.19	-0.12	-0.13	-0.33	0.04	0.19	-0.04
1 2 2	-0.42	-0.07	0.47	-0.09	0.01	-0.21	0.07	0.58	-0.14
1 2 3	-0.11	0.05	0.69	0.05	0.23	0.02	-0.02	0.79	-0.43
1 3 1	-0.45	-0.09	0.20	-0.09	-0.06	-0.30	0.05	0.19	-0.02
1 3 2	-0.37	-0.11	0.48	-0.10	-0.01	-0.28	0.09	0.43	-0.14
1 3 3	-0.14	-0.01	0.67	0.03	0.17	-0.04	-0.06	0.91	-0.53
2 1 1	-0.52	-0.19	0.18	-0.19	-0.19	-0.29	0.03	0.19	-0.11
2 1 2	-0.39	-0.13	0.41	-0.13	0.01	-0.13	0.02	0.25	-0.23
2 1 3	-0.16	-0.04	0.49	-0.04	0.25	0.12	-0.11	0.95	-0.62
2 2 1	-0.48	-0.13	0.19	-0.13	-0.10	-0.30	0.04	0.19	-0.03
2 2 2	-0.36	-0.05	0.47	-0.04	0.03	-0.21	0.07	0.43	-0.14
2 2 3	-0.15	-0.02	7.13	-0.02	0.18	-0.02	-0.11	1.02	-0.59
2 3 1	-0.46	-0.12	0.20	-0.11	-0.06	-0.31	0.04	0.19	-0.03
2 3 2	-0.35	-0.05	0.49	-0.04	0.04	-0.26	0.09	0.47	-0.11
2 3 3	-0.14	-0.01	0.53	-0.01	0.14	-0.08	-0.11	0.88	-0.59
3 1 1	-0.46	-0.09	0.18	-0.08	-0.17	-0.23	0.01	0.20	-0.10
3 1 2	-0.37	-0.09	0.43	-0.09	0.04	-0.02	0.08	0.35	-0.13
3 1 3	-0.14	-0.01	0.08	-0.01	0.33	0.27	-0.11	1.52	-0.59
3 2 1	-0.46	-0.12	0.19	-0.12	-0.15	-0.28	0.04	0.19	-0.03
3 2 2	-0.37	-0.08	0.47	-0.08	0.01	-0.09	0.07	0.43	-0.13
3 2 3	-0.15	-0.04	0.45	-0.04	0.23	0.18	-0.11	0.97	-0.58
3 3 1	-0.44	-0.11	0.20	-0.11	-0.10	-0.27	0.04	0.19	-0.03
3 3 2	-0.36	-0.07	0.49	-0.07	0.02	-0.17	0.07	0.47	-0.14
3 3 3	-0.15	-0.03	0.60	-0.03	0.17	0.06	-0.51	0.88	-0.57
Uniform θ Distribution									
1 1 1	-0.38	0.05	0.19	0.06	-0.13	-0.22	0.06	0.18	0.01
1 1 2	-0.22	0.19	0.41	0.17	0.10	-0.03	0.14	0.38	-0.05
1 1 3	-0.10	0.16	0.98	0.15	0.30	0.16	-0.01	1.21	-0.39
1 2 1	-0.21	0.10	0.19	0.10	0.02	-0.15	0.08	0.19	0.03
1 2 2	-0.21	0.22	0.48	0.22	0.17	-0.03	0.17	0.46	0.02
1 2 3	-0.06	0.18	0.56	0.19	0.29	0.03	0.00	0.87	-0.39
1 3 1	-0.23	0.09	0.20	0.10	0.08	-0.14	0.08	0.20	0.03
1 3 2	-0.18	0.18	0.49	0.18	0.19	-0.12	0.14	0.47	-0.03
1 3 3	-0.06	0.16	-4.42	0.16	0.26	-0.00	-0.03	0.64	-0.45
2 1 1	-0.29	0.08	0.19	0.09	-0.08	-0.18	0.06	0.18	0.01
2 1 2	-0.20	0.21	0.45	0.22	0.14	-0.00	0.16	0.38	-0.00

Condition[a]	κ_m	r_s	$r_{WG(J)}$	*ICC*	CSI_n	CSI_u	$r^*_{WG(J)}$	$r_{WG(J)_n}$	$r^*_{WG(J)_n}$
2 1 3	-0.07	0.20	0.41	0.20	0.33	0.19	0.06	-0.77	-0.31
2 2 1	-0.25	0.08	0.19	0.09	0.03	-0.14	0.07	0.19	0.03
2 2 2	-0.16	0.22	0.48	0.23	0.20	-0.05	0.17	0.46	0.03
2 2 3	-0.05	0.19	0.66	0.20	0.31	0.07	0.04	0.92	-0.31
2 3 1	-0.25	0.09	0.20	0.09	0.08	-0.13	0.07	0.20	0.02
2 3 2	-0.18	0.22	0.49	0.22	0.22	-0.11	0.16	0.48	0.02
2 3 3	-0.05	0.18	0.73	0.18	0.27	-0.01	0.01	1.02	-0.38
3 1 1	-0.26	0.07	0.19	0.08	-0.13	-0.18	0.07	0.18	0.02
3 1 2	-0.18	0.21	0.45	0.22	0.09	0.01	0.15	0.41	0.01
3 1 3	-0.05	0.21	0.53	0.21	0.32	0.27	0.02	-0.26	-0.34
3 2 1	-0.25	0.09	0.19	0.09	-0.04	-0.15	0.07	0.19	0.02
3 2 2	-0.17	0.23	0.48	0.23	0.13	-0.04	0.17	0.47	0.03
3 2 3	-0.03	0.22	0.69	0.23	0.31	0.20	0.05	0.89	-0.31
3 3 1	-0.26	0.08	0.20	0.09	0.01	-0.14	0.07	0.20	0.01
3 3 2	-0.18	0.21	0.49	0.21	0.16	-0.08	0.14	0.48	0.01
3 3 3	-0.04	0.20	0.74	0.20	0.29	0.11	0.03	1.00	-0.36

[a]*Condition* represents the combinations of the various factors manipulated in the study. The following coding scheme is used: Column 1: No. of raters. ! = 2 raters, 2 = 5 raters, 3 = 25 raters; Column 1: No. of items. 1 = 10 items, 2 = 25 items, 3 = 50 items; Column 3: True Consensus. 1 = 80%, 2 = 50%, 3 = 20%

Appendix B

Importance – How important is this task to job performance?

0=Does not apply
1=Unimportant
2=Somewhat important
3=Important
4=Very important
5=Extremely important

Relative Time Spent – How much time do you spend performing this task relative to other tasks?

0=Never do this
1=Very small amount
2=Small amount
3=Average (same) amount
4=Large amount
5=Very large amount

Criticality – What will happen if the task is inadequately performed?

0=Does not apply
1=No serious consequences
2=Least serious consequences
3=Moderately serious consequences
4=Serious consequences
5=Most serious consequences

Performance – How well do you perform the task?

0=Does not apply
1=Poorly
2=Below average
3=Average
4=Above average
5=Excellent

Table 1

Analysis of variance results

Source	df	MS	F	p
Index (I)	8	71.36	27.58	<.0001
Prior (P)	1	7.61	2.94	0.086
Slope (S)	1	1.39	0.54	0.463
Raters (R)	2	2.23	0.86	0.422
Items (It)	2	5.33	2.06	0.127
Consensus (C)	2	129.86	50.19	<.0001
I x P	8	12.57	4.86	<.0001
I x S	8	2.93	1.13	0.336
I x R	16	3.37	1.30	0.185
I x It	16	4.03	1.56	0.071
I x C	16	10.51	4.06	<.0001
I x P x S	8	2.26	0.87	0.538
I x P x R	16	2.26	0.88	0.598
I x S x R	16	2.35	0.91	0.559
I x P x It	16	1.92	0.74	0.754
I x S x It	16	1.40	0.54	0.923
I x R x It	32	3.30	1.27	0.138
I x P x C	16	5.84	2.26	0.003
I x S x C	16	2.74	1.06	0.389
I x R x C	32	3.22	1.25	0.161
I x It x C	32	4.01	1.55	0.025
I x P x S x R	16	2.27	0.88	0.597
I x P x S x It	16	1.98	0.76	0.728
I x P x R x It	32	3.44	1.33	0.102
I x S x R x It	32	2.13	0.82	0.746
I x P x S x C	16	2.06	0.80	0.691
I x P x R x C	32	2.23	0.86	0.691
I x P x It x C	32	1.78	0.69	0.91
I x S x It x C	32	1.38	0.53	0.986
I x R x It x C	64	3.19	1.23	0.101
I x P x S x R x It	32	1.76	0.68	0.91
I x P x S x R x C	32	2.25	0.87	0.677
I x P x S x It x C	32	1.84	0.71	0.885
I x P x R x It x C	64	3.52	1.36	0.029
I x S x R x It x C	64	2.09	0.81	0.086
I x P x S x R x It x C	64	1.80	0.70	0.969

Table 2

Means and standard deviations of agreement estimates conditioning on degree of agreement and true distribution of θ

		Consensus		
Distribution	Index	High	Medium	Low
Normal	κ_m	.318 (.13)	.122 (.10)	.048 (.08)
	r_s	.670 (.12)	.415 (.15)	.179 (.17)
	ICC (2,1)	.671 (.12)	.412 (.15)	.184 (.17)
	CSI_n	.669 (.08)	.520 (.06)	.416 (.09)
	CSI_u	.507 (.08)	.337 (.10)	.267 (.11)
	$r_{WG(J)}$	.989 (.01)	.959 (.04)	1.61 (9.81)
	$r_{WG(J)_a}$	.989 (.05)	.952 (.66)	1.24 (.99)
	$r^{\bullet}_{WG(J)}$	.838 (.05)	.576 (.10)	.097 (.16)
	$r^{\bullet}_{WG(J)_a}$	.737 (.21)	.360 (.14)	-.358 (.25)
Uniform	κ_m	.534 (.11)	.315 (.10)	.144 (.08)
	r_s	.880 (.06)	.709 (.10)	.389 (.14)
	ICC (2,1)	.887 (.05)	.710 (.09)	.391 (.14)
	CSI_n	.782 (.08)	.657 (.06)	.497 (.07)
	CSI_u	.639 (.05)	.450 (.07)	.313 (.11)
	$r_{WG(J)}$	.993 (.01)	.969 (.05)	.297 (5.51)
	$r_{WG(J)_a}$	.988 (.01)	.944 (.09)	.815 (3.26)
	$r^{\bullet}_{WG(J)}$	.879 (.03)	.668 (.10)	.227 (.16)
	$r^{\bullet}_{WG(J)_a}$	.820 (.05)	.504 (.15)	-.162 (.24)

Table 3

Average RMSE values (and standard deviations) of agreement estimates versus true consensus conditioning on degree of agreement and true distribution of θ

Distribution	Index	Consensus		
		High	Medium	Low
Normal	κ_m	.495 (.06)	.390 (.05)	.165 (.04)
	r_s	.165 (.06)	.156 (.08)	.132 (.10)
	ICC (2,1)	.192 (.10)	.169 (.10)	.128 (.11)
	CSI_n	.141 (.05)	.006 (.02)	.224 (.06)
	CSI_u	.300 (.04)	.174 (.08)	.111 (.07)
	$r_{WG(J)}$	.190 (.01)	.460 (.03)	3.383 (9.56)
	$r_{WG(J)_a}$	.195 (.03)	.594 (.55)	1.282 (.68)
	$r^{*}_{WG(J)}$	.060 (.02)	.111 (.04)	.181 (.07)
	$r^{*}_{WG(J)_a}$	.131 (.18)	.188 (.06)	.606 (.07)
Uniform	κ_m	.282 (.06)	.207 (.04)	.091 (.05)
	r_s	.099 (.02)	.227 (.03)	.232 (.05)
	ICC (2,1)	.102 (.01)	.228 (.03)	.232 (.04)
	CSI_n	.073 (.04)	.163 (.04)	.304 (.02)
	CSI_u	.168 (.03)	.080 (.04)	.136 (.08)
	$r_{WG(J)}$	.193 (.01)	.472 (.02)	2.731 (4.91)
	$r_{WG(J)_a}$	.188 (.01)	.452 (.03)	2.271 (2.49)
	$r^{*}_{WG(J)}$	.087 (.01)	.193 (.03)	.141 (.08)
	$r^{*}_{WG(J)_a}$	.051 (.02)	.115 (.09)	.424 (.09)

Table 4

Average bias values (and standard deviations) of agreement estimates versus true consensus conditioning on degree of agreement and true distribution of θ

		Consensus		
Distribution	Index	High (.80)	Medium (.50)	Low (.20)
Normal	κ_m	-.482 (.05)	-.378 (.04)	-.150 (.04)
	r_s	-.129 (.04)	-.085 (.04)	-.021 (.06)
	ICC (2,1)	-.129 (.05)	-.088 (.04)	-.016 (.06)
	CSI_n	-.131 (.06)	.019 (.02)	.216 (.06)
	CSI_u	-.293 (.03)	-.163 (.08)	.068 (.11)
	$r_{WG(J)}$	.189 (.01)	.459 (.03)	1.409 (3.12)
	$r_{WG(J)_a}$	.189 (.01)	.452 (.21)	1.043 (.34)
	$r^*_{WG(J)}$	.033 (.02)	.071 (.03)	-.132 (.20)
	$r^*_{WG(J)_a}$	-.063 (.08)	-.139 (.05)	-.558 (.08)
Uniform	κ_m	-.266 (.05)	-.187 (.04)	-.056 (.03)
	r_s	.080 (.02)	.209 (.05)	.189 (.03)
	ICC (2,1)	.087 (.01)	.210 (.05)	.191 (.03)
	CSI_n	-.018 (.08)	.157 (.05)	.297 (.03)
	CSI_u	-.161 (.03)	-.050 (.05)	.113 (.10)
	$r_{WG(J)}$	.193 (.01)	.470 (.03)	.097 (1.68)
	$r_{WG(J)_a}$	.188 (.01)	.444 (.05)	.615 (.91)
	$r^*_{WG(J)}$	.071 (.01)	.156 (.05)	.018 (.04)
	$r^*_{WG(J)_a}$	.020 (.01)	.003 (.07)	-.362 (.06)

Table 5

Analysis of variance results (simple effects)

Distribution	Consensus	Source	df	MS	F	p
Normal	High	Index	8	8.38	771.72	<.0001
		Index*Items	16	.036	3.32	<.0001
		Index*Raters	16	.015	1.39	.1368
		Index*Items*Raters	32	.007	.61	.9602
	Medium	Index	8	14.02	234.87	<.0001
		Index*Items	16	.067	1.13	.3194
		Index*Raters	16	.010	1.69	.0427
		Index*Items*Raters	32	.033	.55	.9817
	Low	Index	8	69.17	6.39	<.0001
		Index*Items	16	9.64	.89	.5804
		Index*Raters	16	10.09	.93	.5316
		Index*Items*Raters	32	12.42	1.15	.2617
Uniform	High	Index	8	4.17	1506.5	<.0001
		Index*Items	16	.044	16.05	<.0001
		Index*Raters	16	.009	3.07	<.0001
		Index*Items*Raters	32	.004	1.54	.0281
	Medium	Index	8	8.32	1011.9	<.0001
		Index*Items	16	.046	5.62	<.0001
		Index*Raters	16	.009	1.21	.2541
		Index*Items*Raters	32	.002	.22	1.000
	Low	Index	8	12.56	2.83	.0040
		Index*Items	16	7.69	1.73	.0352
		Index*Raters	16	6.31	1.42	.1226
		Index*Items*Raters	32	7.69	1.73	.0070

Table 6

Estimated variance components for RMSE of the simulated data

	Source of Variation	df	Mean Square	Estimated Variance Component	Percentage of Total Variance
Normal θ distribution					
Consensus = High					
	Index	8	0.1412	0.0149	75
	Index*Items	16	0.0067	0.0019	10
	Index*Raters	16	0.0018	0.0003	1
Consensus = Med					
	Index	8	0.2948	0.0298	56
	Index*Items	16	0.0097	0	0
	Index*Raters	16	0.0298	0.0057	10
Consensus = Low					
	Index	8	10.491	0.7051	11
	Index*Items	16	4.8562	0	0
	Index*Raters	16	5.1506	0	0
Uniform θ distribution					
Consensus = High					
	Index	8	0.0498	0.0054	85
	Index*Items	16	0.0012	0.0003	4
	Index*Raters	16	0.0001	0	0
Consensus = Medium					
	Index	8	0.1684	0.0182	89
	Index*Items	16	0.0594	0.0011	6
	Index*Raters	16	0.0016	0.0004	2
Consensus = Low					
	Index	8	9.2894	0.3461	8
	Index*Items	16	3.9146	0.4334	10
	Index*Raters	16	4.8743	0.7533	18

Table 7

Estimated variance components for bias of the simulated data

	Source of Variation	df	Mean Square	Estimated Variance Component	Percentage of Total Variance
Normal θ distribution					
Consensus = High					
	Index	8	0.4178	0.0462	96
	Index*Items	16	0.0017	0.0004	1
	Index*Raters	16	0.0007	0.0001	0
Consensus = Med					
	Index	8	0.7003	0.0771	95
	Index*Items	16	0.0035	0.0007	1
	Index*Raters	16	0.0049	0.0011	1
Consensus = Low					
	Index	8	3.4799	0.3471	36
	Index*Items	16	0.4779	0	0
	Index*Raters	16	0.5030	0	0
Uniform θ distribution					
Consensus = High					
	Index	8	0.2076	0.0228	94
	Index*Items	16	0.0021	0.0006	3
	Index*Raters	16	0.0004	0.00006	<1
Consensus = Med					
	Index	8	0.4169	0.0460	97
	Index*Items	16	0.0023	0.0008	2
	Index*Raters	16	0.0005	0.0001	<1
Consensus = Low					
	Index	8	0.6302	0.0353	8
	Index*Items	16	0.3830	0	0
	Index*Raters	16	0.3141	0	0

Table 8

Means and standard deviations of rank orders by agreement index

Index	Mean	SD
κ_m	8.61	0.60
r_s	4.93	1.20
ICC (2,1)	4.72	1.28
CSI_n	4.61	1.66
CSI_u	6.72	1.66
$r_{wg(J)}$	1.50	1.19
$r_{wg(J)_u}$	1.91	1.39
$r^*_{wg(J)}$	4.89	1.62
$r^*_{wg(J)_u}$	7.11	1.69

Note: Ranks computed in descending order (thus, a rank of 1 indicates the highest agreement estimate).

Means and standard deviations are across 54 consolidated conditions.

Table 9

Estimated variance components for rank orders

Source of Variation	*df*	*Mean Square*	*Estimated Variance Component*	*Percentage of Total Variance*
Index	8	5633.78	5.2146	72
Error	9711	2.0314	2.0314	28

Table 10

Comparison of agreement indices for Study 2 data

	Dimension			
Index	Importance	Time Spent	Criticality	Performance
κ_m	.062	.065	.070	.055
r_s	.270	.252	.275	.187
ICC (2,1)	.206	.241	.227	.199
CSI_n	.511	.416	.417	.375
CSI_u	.495	.391	.373	.342
$r_{WG(J)}$	.944	.997	.995	.997
$r_{WG(J)_u}$	.983	.993	.987	.995
$r^*_{WG(J)}$	.513	.652	.543	.707
$r^*_{WG(J)_u}$	.268	.470	.313	.555

Table 11

Comparison of agreement indices for Study 3 data

	Dimension		
Index	Leadership	Well-liked	Shy/Loner
κ_m	.117	.175	.447
r_s	.195	.257	.487
ICC (2,1)	.154	.212	.485
CSI_n	.524	.358	.732
CSI_u	.329	.206	.945
$r_{WG(J)}$	.780	.878	.984
$r_{WG(J)_n}$	.780	.878	.984
$r^*_{WG(J)}$	.191	.325	.799
$r^*_{WG(J)_n}$	.191	.325	.799

Figure 1. Person Perception Function with Alpha = 0.8 and Beta = 1.0

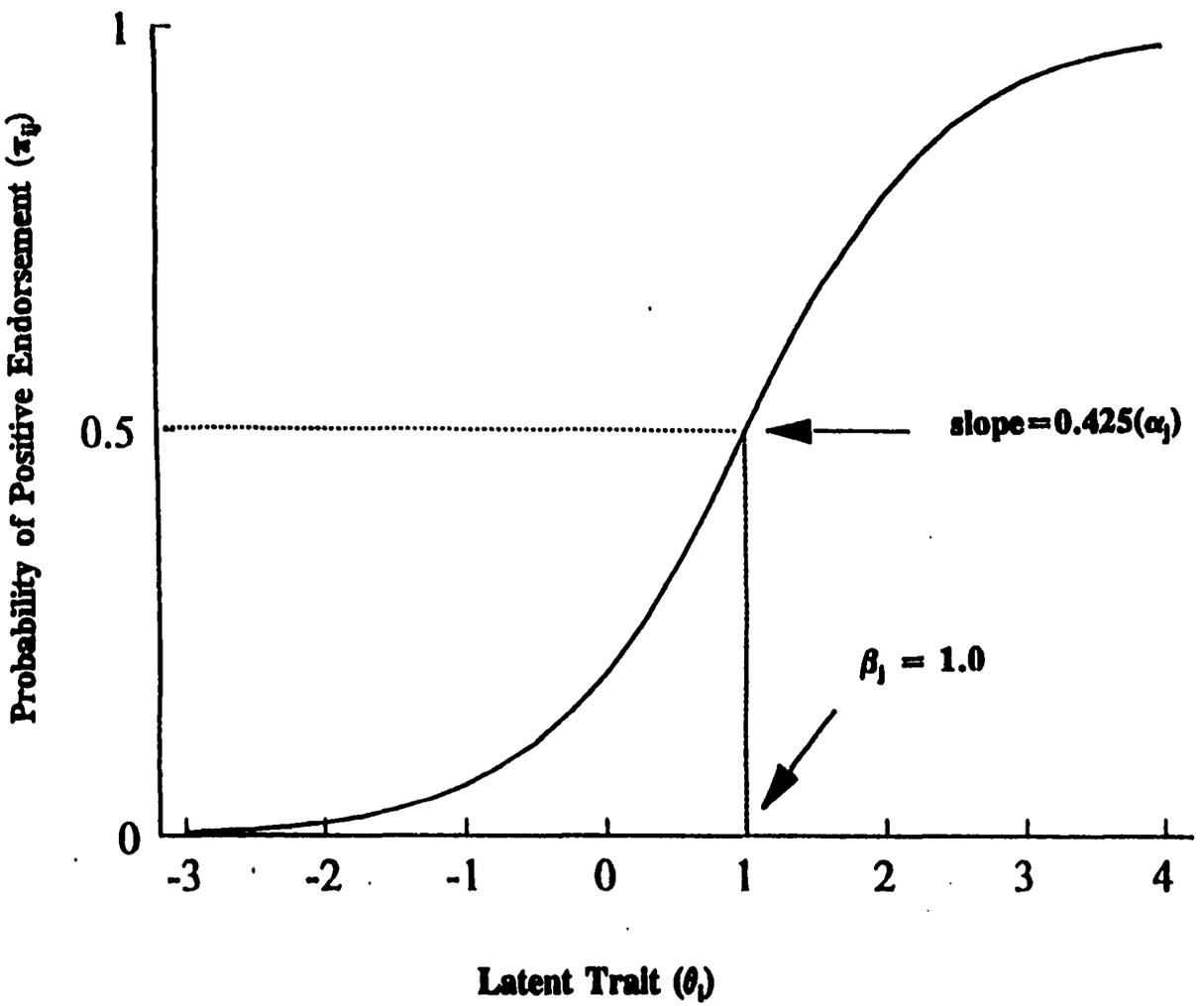


Figure 2. Simulation Design

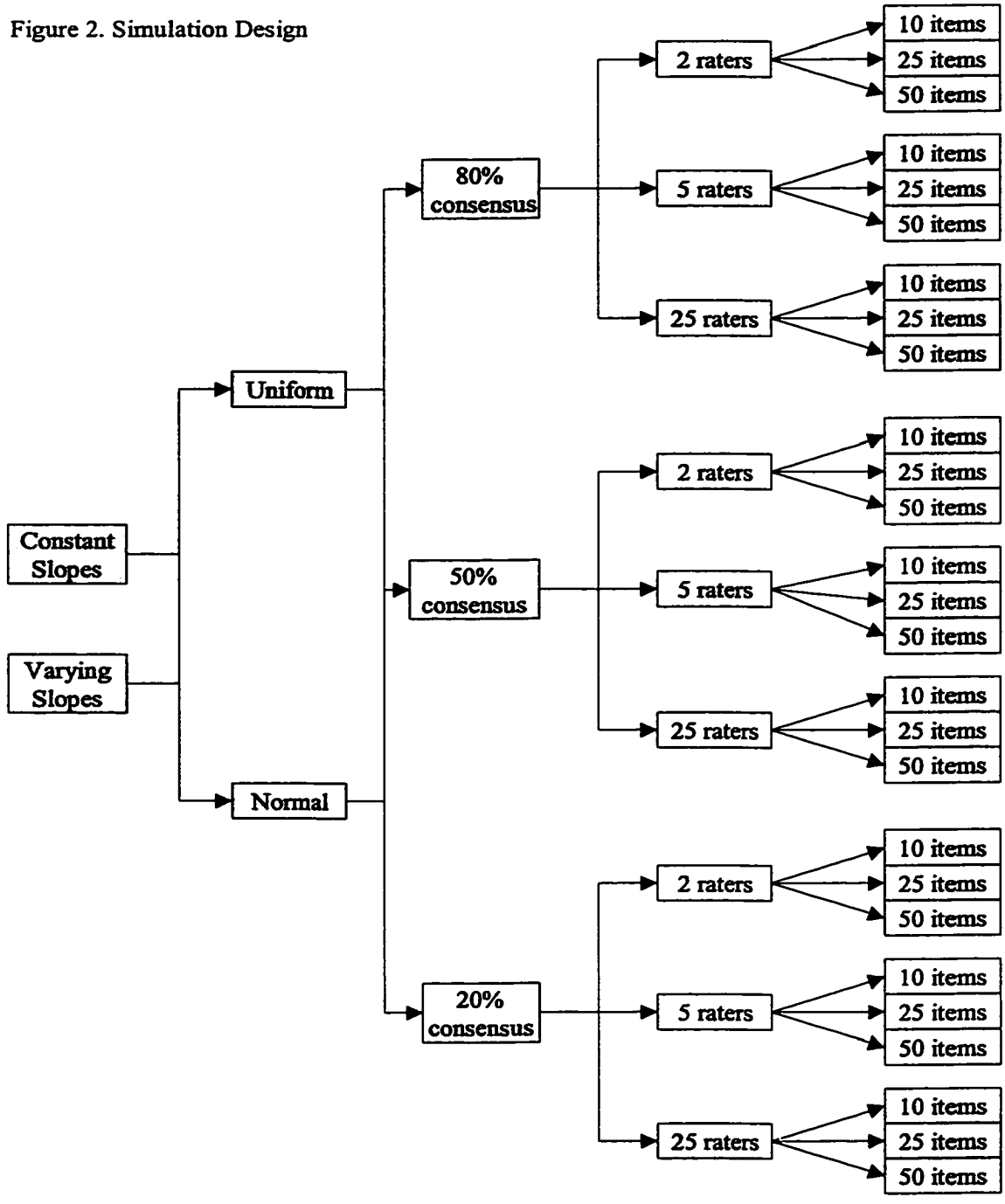


Figure 3. Normal True Distribution, 80% Consensus

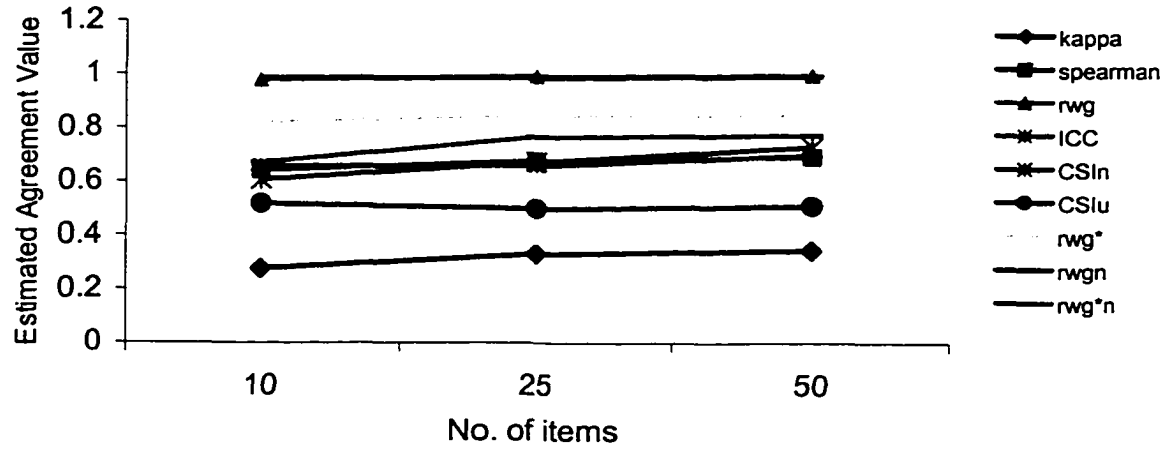


Figure 4. Normal True Distribution, 80% Consensus

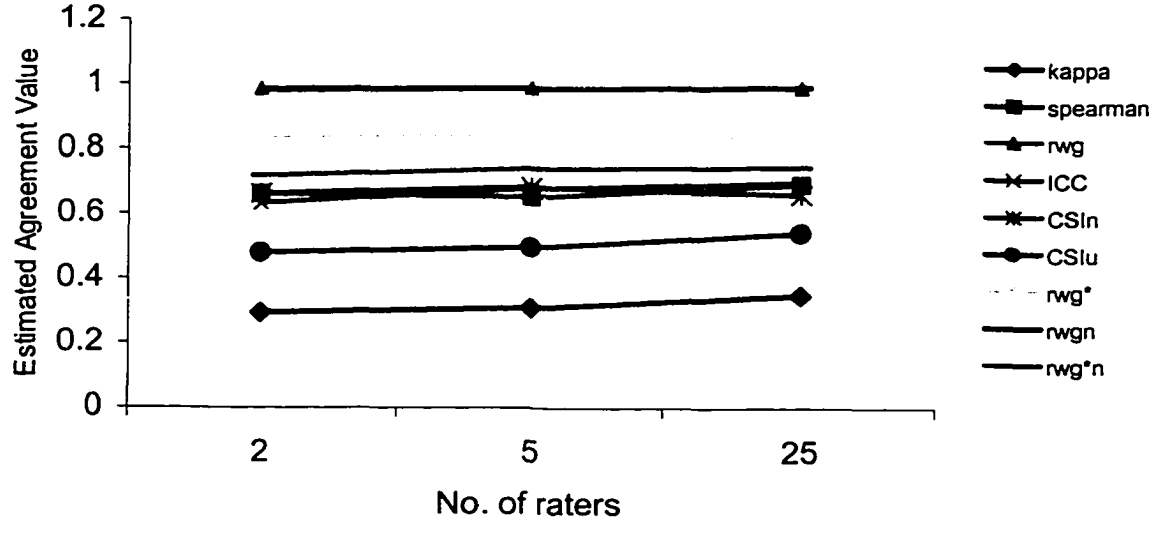


Figure 5. Normal True Distribution, 50% Consensus

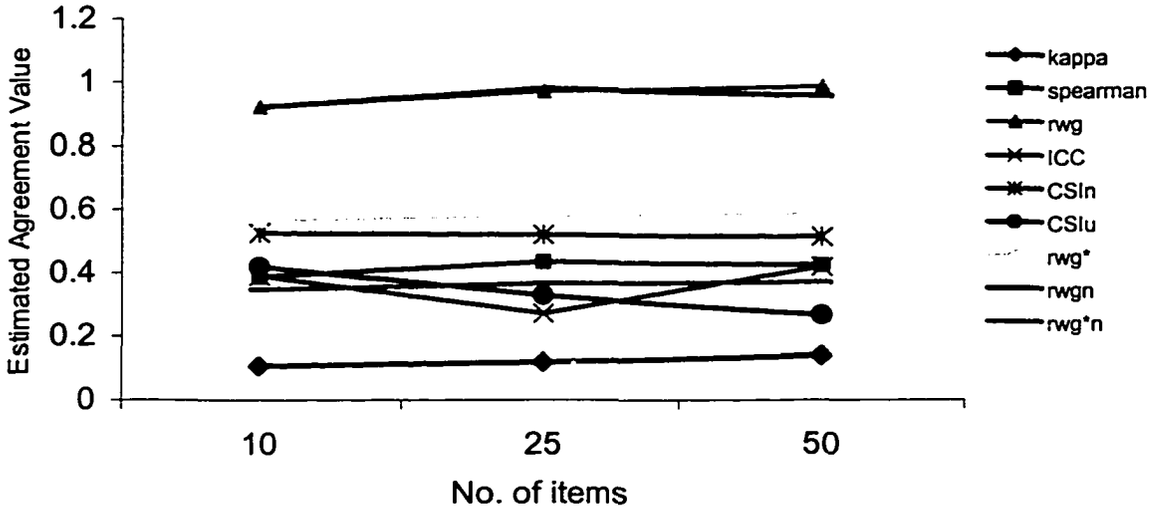


Figure 6. Normal True Distribution, 50% Consensus

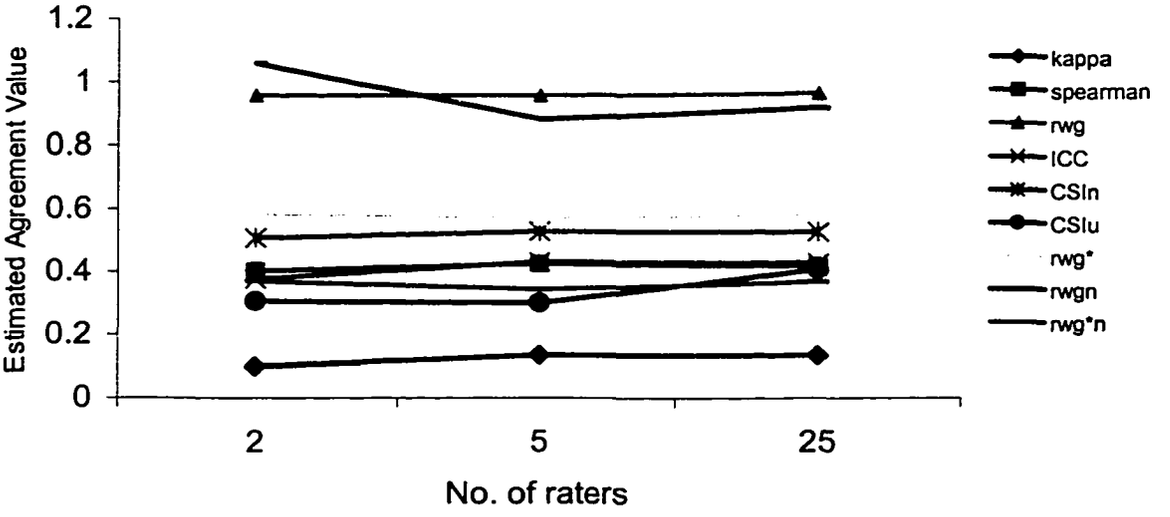


Figure 7. Normal True Distribution, 20% Consensus

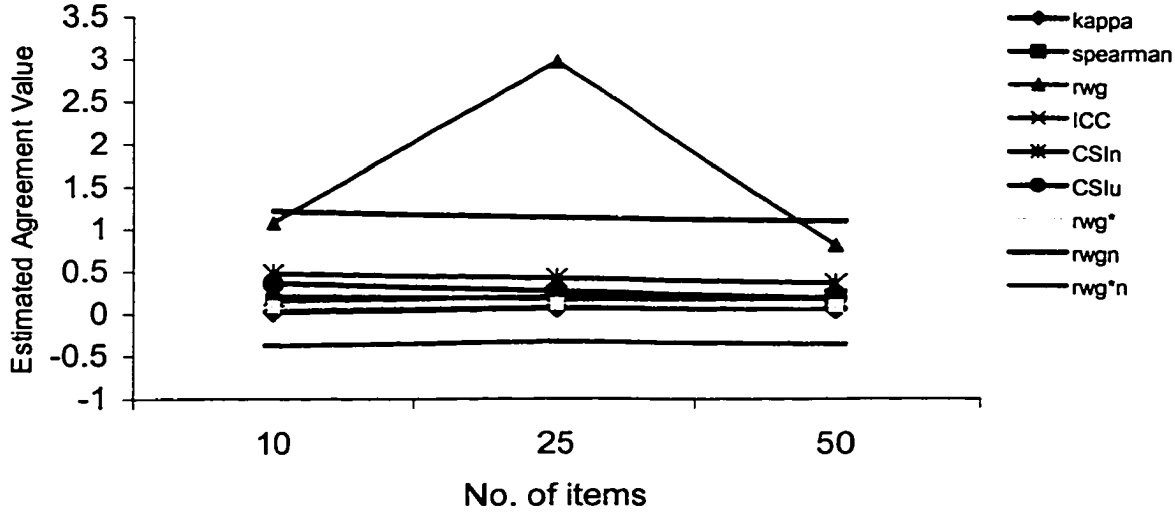


Figure 8. Normal True Distribution, 20% Consensus

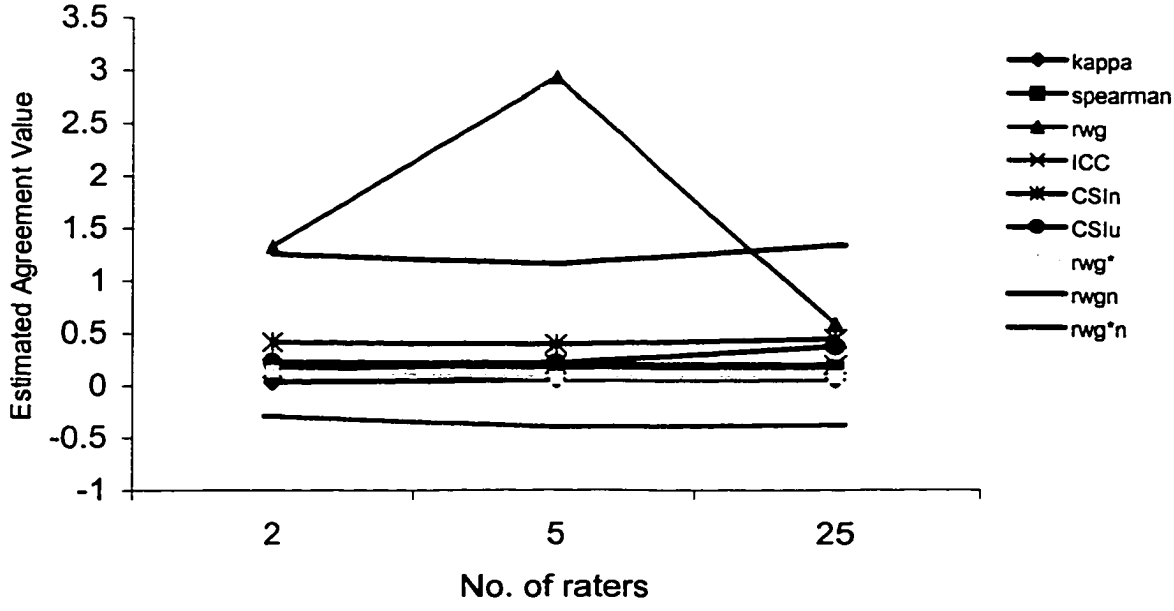


Figure 9. Uniform True Distribution, 80% Consensus

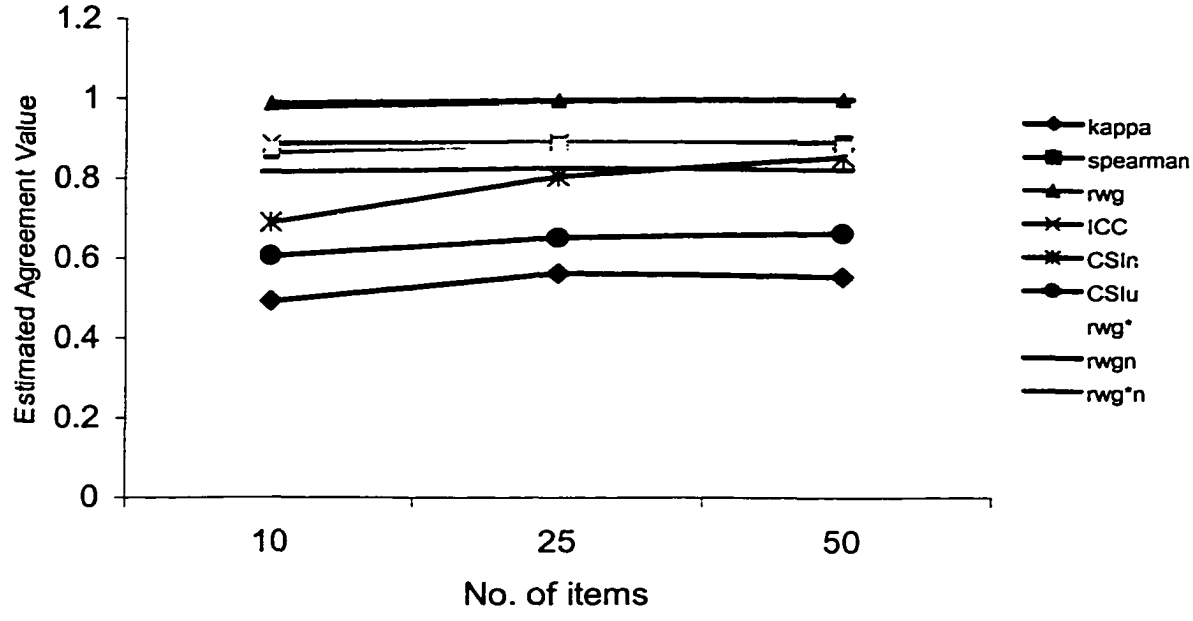


Figure 10. Uniform True Distribution, 80% Consensus

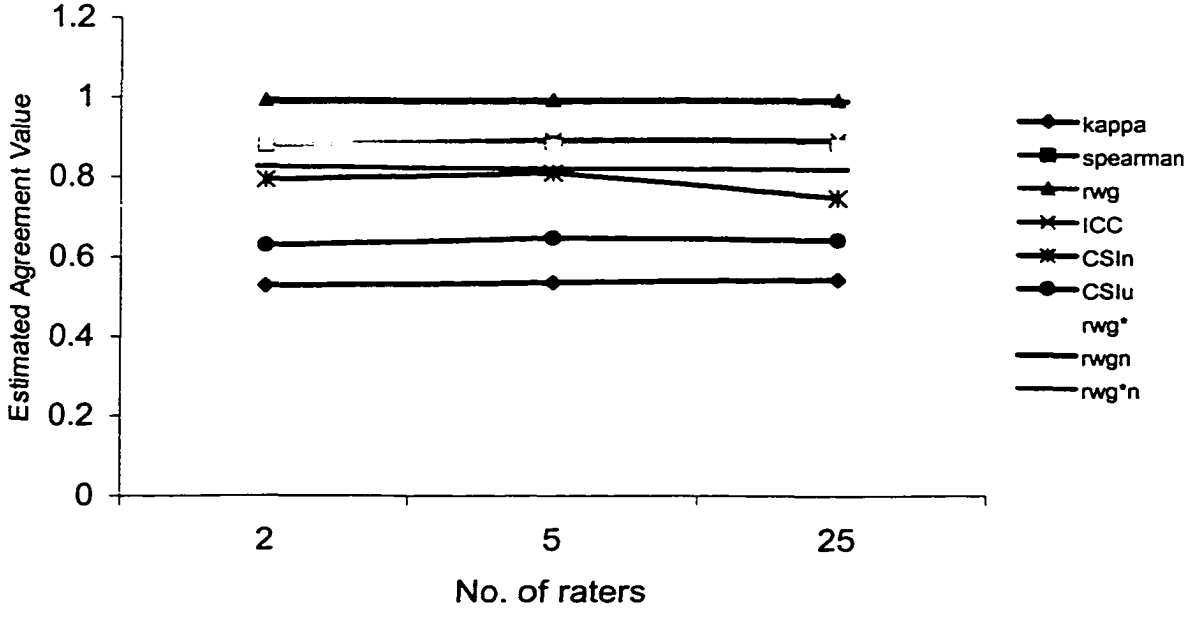


Figure 11. Uniform True Distribution, 50% Consensus

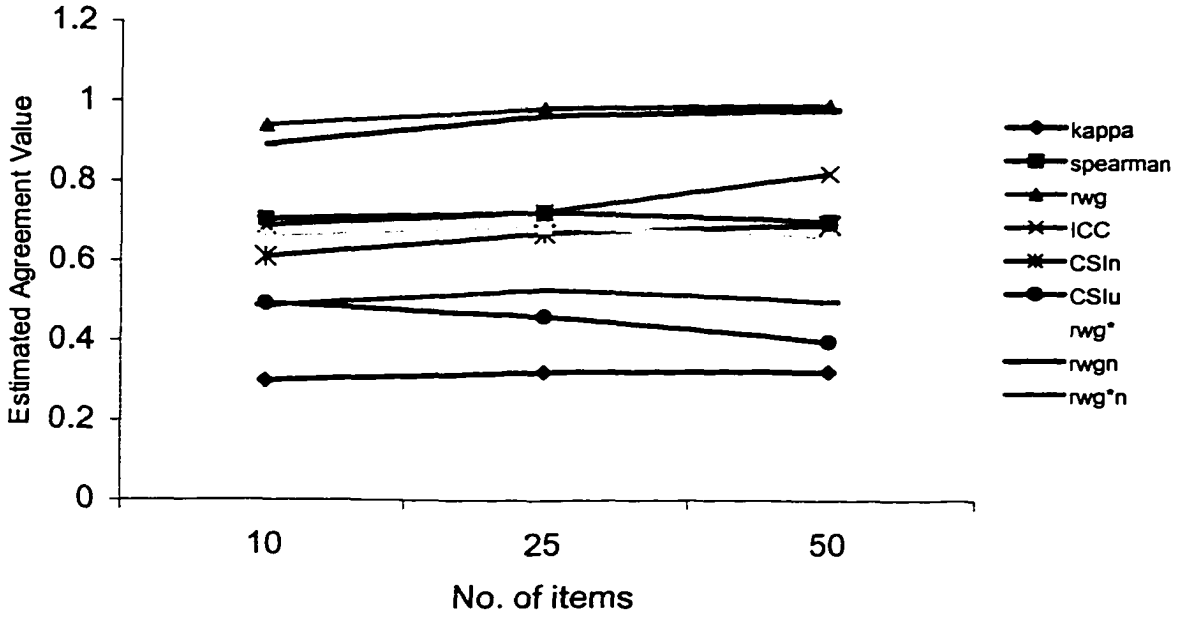


Figure 12. Uniform True Distribution, 50% Consensus

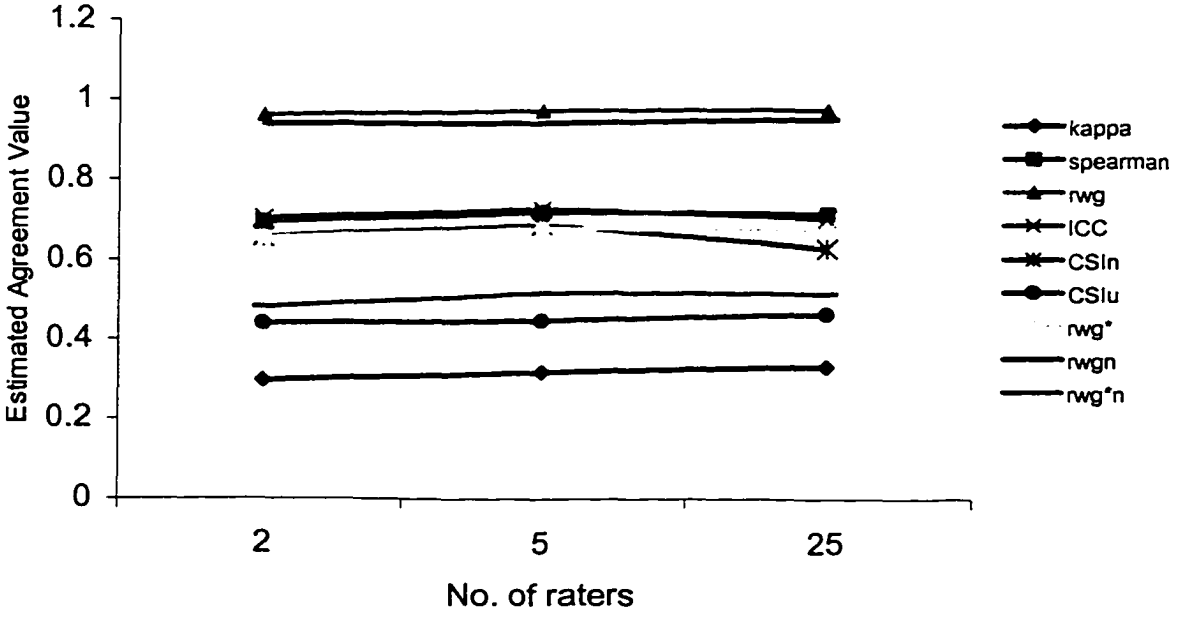


Figure 13. Uniform True Distribution, 20% Consensus

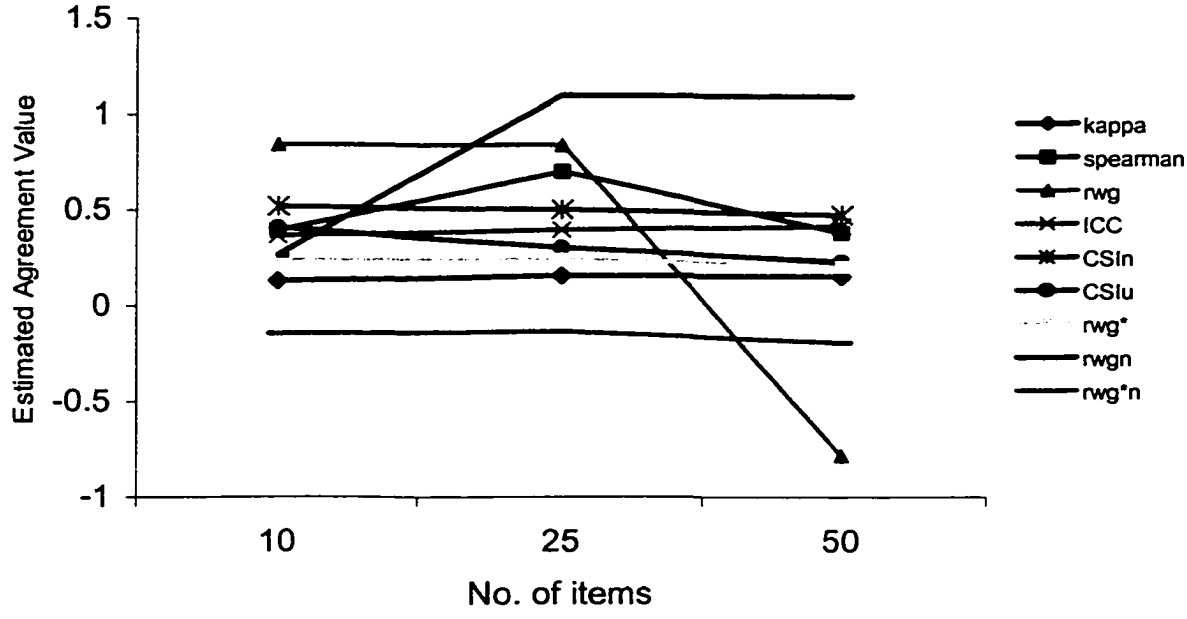


Figure 14. Uniform True Distribution, 20% Consensus

