Educational and Psychological Measurement

http://epm.sagepub.com/

Interrater Agreement Evaluation: A Latent Variable Modeling Approach

Tenko Raykov, Dimiter M. Dimitrov, Alexander von Eye and George A. Marcoulides

Educational and Psychological Measurement published online 6 July 2012
DOI: 10.1177/0013164412449016

Educational and Psychological
Measurement
XX(X) 1–20
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164412449016
http://epm.sagepub.com

Interrater Agreement Evaluation: A Latent Variable Modeling Approach

Tenko Raykov[1], Dimiter M. Dimitrov[2], Alexander von Eye[1], and George A. Marcoulides[3]

Abstract

A latent variable modeling method for evaluation of interrater agreement is outlined. The procedure is useful for point and interval estimation of the degree of agreement among a given set of judges evaluating a group of targets. In addition, the approach allows one to test for identity in underlying thresholds across raters as well as to identify possibly aberrantly evaluating judges. A measure of interrater agreement is proposed, which is related to popular indexes of interrater reliability for observed variables and composite reliability. The outlined method also permits the examination of underlying common sources of ratings variability, provides a useful complement to the literature on interrater agreement with manifest measures, and relaxes some of its assumptions. The procedure is illustrated with numerical data.

Keywords

interrater agreement, latent variable modeling, reliability, raters, targets

In behavioral, educational, social, and biomedical research, a study design is oftentimes used where several ''raters'' (or ''judges'')—such as psychiatrists, counselors, teachers, clinicians, evaluators, or observers—are asked to evaluate a group of

[1] Michigan State University, East Lansing, MI, USA
[2] George Mason University, Fairfax, VA, USA
[3] University of California at Riverside, Riverside, CA, USA

Corresponding Author:
Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443A Erickson Hall, East Lansing, MI 48824, USA
Email: raykov@msu.edu

'targets,' for example, subjects, students, teachers, clients, objects, or patients. In this context, a researcher is typically interested in estimating the interrater agreement (IRA) as the degree to which there is consistency in the ratings (cf. Shrout & Fleiss, 1979). A large body of literature is currently available that deals with this and related topics in various circumstances characterized by different inferential aims (e.g., Burke & Dunlap, 2002; Congdon & McQueen, 2000; Dimitrov, 2012; LeBreton & Senter, 2008; Lindell, 2001; Raymond, Harik, & Clauser, 2011; Shoukri, 2011; von Eye & Mun, 2005).[1]

A useful approach to evaluating IRA would be to adopt a widely employed assumption across a number of empirical contexts in the behavioral and social sciences. Specifically, each rater may be presumed to be using an underlying continuous dimension along which he or she implicitly evaluates each target. However, the rater can provide only ordinal information about the location of each rated target on that continuum, with such information resulting from a process of coarse measurement by the judge of the extent to which the characteristic of concern may be possessed by that target (e.g., Skrondal & Rabe-Hesketh, 2004). Thereby, the rater states only one of a limited set of several possible integer numbers—the pertinent individual rating.

The assumption of the existence of an underlying latent dimension (continuum) along which rater evaluation proceeds is widely used in the popular latent variable modeling (LVM) framework when dealing with categorical dependent variables (e.g., Muthén, 1984). This assumption, which has proved useful in theoretical and empirical behavioral and social research, opens here an opportunity to capitalize on recent advances in LVM for the purpose of IRA evaluation (see Raykov & Marcoulides, 2012, for examining this assumption in an empirical setting; cf. Jöreskog & Sörbom, 1996.) To our knowledge, an LVM-based approach to IRA evaluation with this assumption has not been published in previous IRA studies. In addition, interval estimations of IRA have been rarely carried out in past research, and only on assumptions that could not be readily considered tenable (cf. Raykov, 2011; Shrout & Fleiss, 1979; Shrout & Lane, in press).

The present article intends to contribute to bridging this gap. The following discussion describes an LVM-based approach to evaluating IRA when multiple targets are evaluated by a fixed set of raters. The outlined procedure can be used for point and interval estimation of the extent to which the raters are consistent in their assessment of the targets. The procedure uses less stringent assumptions compared with methods that are based on manifest variables. The proposed approach can also be used when the goals are (a) to examine a set of raters for possible common sources in their ratings' variability, (b) to test underlying thresholds for identity across raters, and (c) to identify possibly aberrantly evaluating raters.

With these features, the article contributes to a substantial body of literature on IRA evaluation that already contains several prior attempts to devise particular models for IRA. Within the context of manifest variables, Schuster (2001), for example, advanced an interpretation of Cohen's (1960) kappa as a parameter of a log-linear symmetry model. von Eye and Mun (2005) discussed a family of log-linear models of rater agreement that incorporates, among others, Tanner and Young's (1985) equal weight agreement model, the weight-by-response-category agreement model, the symmetry model, models with covariates, weighted kappa (Cohen, 1968), and models with ordinal categories (Schuster & von Eye, 2001; see also Schuster & Smith, 2005; Uebersax, 1993). Alternatively, within the latent variable framework, mixture models (Schuster, 2002) have been discussed as well as latent class models (Schuster & Smith, 2002) that similarly aimed at IRA evaluation. von Eye and Mun (2005) further proposed a structural equation model for the comparison of ratings from two or more groups of raters.

The method outlined in the remaining discussion differs from all these approaches in a number of important respects. First, this is a latent variables approach, and as such it is distinct from manifest variables procedures in that a measure of the degree of agreement is based on underlying latent dimensions here. In addition, the following procedure differs from the latent variable models proposed in the above-mentioned work by Schuster and colleagues (Schuster & Smith, 2002; Schuster & von Eye, 2001) in that (a) the present one uses groups of raters instead of individuals and (b) the raters are considered random variables here. This article's method is also distinct from von Eye and Mun's (2005) structural equation modeling approach in that the former goes beyond comparing groups of raters by proposing measures of the degree of agreement and tools to also evaluate individual raters. Finally, the following procedure differs from all those alternative approaches in that this method makes assumptions about underlying processes that are continuous in nature.

Background, Notation, and Assumptions

In the remainder of this article, we assume that n targets are evaluated by each of r raters using an ordinal scale with possible scores 1, 2, . . ., m ($n, m > 1$; $r > 2$). That is, each rater is asked to evaluate a certain characteristic for each target and can use thereby only one of these m possible values. Below, we will also refer to the raters as ''evaluators'' or ''judges,'' to the objects of measurement as ''targets'' or ''objects,'' and on a few occasions to IRA as ''inter-observer agreement'' (cf. Shavelson & Webb, 1991). For example, a set of teachers may be evaluated by a set of inspectors, whereby the teacher's performance in a prespecified area of professional activity (e.g., degree of being successful in engaging students in discussion during class) is graded using the rating 1, 2, . . ., or m, independently of how any other teacher is evaluated on the same characteristic by the same or another judge. In this setting, the goal is to obtain point and interval estimates of an IRA index, that is, the extent to which the judges are consistent in their ratings. This IRA index should reflect correspondingly the amount of overlap among observers in their evaluations of the examined objects.

As indicated earlier, this article adopts an LVM-based approach to IRA evaluation. The method outlined below is motivated by the assumption that a set of judges would be consistent in their ratings of targets to the extent to which they might be using a common underlying ‘‘metric’’ for the purpose of target assessment. As part of this assumption, that metric may be conceptualized as a common underlying continuum along which the targets may be thought of as being approximately positioned. This assumption is testable within the proposed LVM-based approach. The following method is also readily extended to the case when more than a single latent dimension may be underlying the raters’ evaluations, which will be addressed in a later section.

To develop an LMV-based approach to IRA estimation, we view the raters as separate observed random variables that are formally ‘‘taken’’ or measured on each of the targets. Because of the coarse evaluation involved, these variables represent categorical manifest measures, which are denoted here $y_1, \ldots, y_r$ and viewed as elements of the vector $\mathrm{y} = (y_1, \ldots, y_r)'$ (priming denotes transposition in this article). Furthermore, as typically done in applications of LVM and indicated earlier, we assume the existence of a normal latent variable underlying each categorical observed variable (Muthén, 1984; Muthén & Muthén, 2010; see also Raykov & Marcoulides, 2012, for testing of this assumption). Let us designate these underlying variables by y_j^*, respectively ($j = 1, \ldots., r$), and collect them in the vector $\mathrm{y}^* = (y_1^*, \ldots, y_r^*)'$. The latent variables in y^* are of intrinsic interest to measure and obtain observations on, but because of their ‘‘hidden’’ (latent) nature, only their coarse evaluation is possible by the raters when assessing the target characteristic(s) of concern in an empirical study. That is, while of real interest are the individual values y_{ij}^*, only their corresponding coarse observations y_{ij} are available ($i = 1, \ldots, n$; $j = 1, \ldots, r$). For simplicity of notation, we use in the remainder the same symbol y for random variables and their realizations and will point out this distinction where confusion may arise.

To use the LVM framework with categorical observed variables for IRA estimation, it is next noted that the observed ratings (scores) produced by the raters can be viewed as resulting from the relationship of their underlying latent variable realization to an associated set of thresholds (e.g., Skrondal & Rabe-Hesketh, 2004). Specifically, denoting these thresholds by $\tau_{j1}, \tau_{j2}, \ldots, \tau_{j,m-1}$ for the jth rater, this relationship is

$$\begin{aligned} y_j &= 1, \text{if } -\infty < y_j^* \leq \tau_{j1}, \\ &= 2, \text{if } \tau_{j1} < y_j^* \leq \tau_{j2}, \\ &\quad \ldots \\ &= m, \text{if } \tau_{j,m-1} < y_j^* < \infty, \end{aligned} \tag{1}$$

where, $j = 1, 2, \ldots, r$). As often conducted in applications of LVM with categorical manifest variables, a confirmatory factor analysis (CFA) model can be fitted to the underlying variables y^* (Muthén, 1984). In the context of IRA evaluation, it may be argued that oftentimes it would be of interest to fit a single-factor model. Indeed, it is readily realized that rater agreement may well result from the fact that the underlying variables y^* have a dominant source of shared variability, which would then be represented by their common factor (see the Discussion and Conclusion section for alternatives). That is, the following testable model

$$y^* = \Lambda\eta + \varepsilon, \tag{2}$$

where Λ is the $r \times 1$ matrix of factor loadings, η the factor, and ε the $r \times 1$ vector of unique factors, can be employed as a useful starting point when studying IRA. As usual, the unique factors in Model (2) are assumed uncorrelated among themselves and with the common factor η (e.g., Muthén, 1984). We stress that Model (2) is testable when empirical data are available using LVM. If this model fits the data well, the single factor η can be straight-forwardly interpreted as the common source of variability in the rater-specific dimensions $y^*_1, \ldots, y^*_r$ that underlie the process of target evaluation by the judges. In the remainder, we assume that Model (2) is plausible, unless indicated otherwise (see the Discussion and Conclusion section).

Adopting Model (2) permits us also to address an issue that does not seem to have received the deserved attention in the extant literature on IRA. Specifically, when evaluating IRA, a natural question to posit initially is whether there are commonalities in the specific ''cutoffs'' that raters use to evaluate the targets. This question translates here to that of identity across judges of each of the $m - 1$ thresholds in Equation (1), that is, across the r observed random variables involved. In our view, it is important to assume that these thresholds are at least comparable if not the same for all raters, since otherwise it may be hard to speak of IRA to begin with. The proposed LVM approach is applicable whether the thresholds are the same or not across the raters, and in addition allows testing for their identity across raters as a by-product (see below). Moreover, Model (2) permits one to identify judges that provide possibly aberrant evaluations compared with the other raters. Specifically, a comparison of each factor loading in Model (2) with each of the remaining $(r - 1)$ factor loadings for the judges makes it possible to identify potentially ''outlying'' raters. In addition, the method does not need to assume that each judge ''draws'' to the same extent from the underlying latent dimension when evaluating the targets. That is, the rater-specific loadings $\lambda_1, \ldots, \lambda_r$ in the single-factor Model (2) need not be assumed equal across judges, and in fact their differences permit a considerable degree of flexibility in the model when applied in empirical research, while the identity of these loadings for all or a subset of the raters is similarly testable with the present approach. (Equality of any of these factor loadings is testable using essentially the same approach as that for testing threshold identity; see below.) Furthermore, the procedure outlined next does not assume equal residual variances across the raters, that is, the variances of the rater-specific residual terms $\varepsilon_1, \ldots, \varepsilon_r$ need not be equal (cf. Shrout & Fleiss, 1979; Shrout & Lane, in press; the equality of these variances is similarly testable with essentially the same method as that used for threshold identity testing below).

A Latent Variable Modeling–Based Index of Interrater Agreement

Typically when working with observed categorical variables, of actual interest are the individual values on their underlying normal variables, y^*_{ij} ($i = 1, \ldots, n$; $j = 1, \ldots, r$). For this reason, once having fitted the single-factor Model (2) and found it plausible as a means of data description, for the following purposes one may consider these values y_{ij}^* as if they were available. (This consideration is not essential for the developments next, but is useful in describing the idea behind the following proposed IRA coefficient.) Then the IRA index of this article, defined next in the section, can be obtained independently using one of four possible approaches applied on the underlying values y_{ij}^* ($i = 1, \ldots, n$; $j = 1, \ldots, r$). In the following subsections, we elaborate in turn on each of these approaches, which we show lead to the same IRA index. (In the remainder of this section, for convenience of notation we drop the individual subindex i in order to simplify notation, since we will not need specific reference to it, unless indicated otherwise.)

Interrater Agreement as Proportion Common Variance

One possible way to conceptualize IRA in the currently considered setting is as the proportion of variance in the average underlying rating value that is common to all raters. Accordingly, we define IRA as the proportion of variance in the average underlying rating,

$$y_\bullet^* = \left(y_1^* + \cdots + y_r^*\right)/r, \tag{3}$$

which is accounted for by the common factor η of the pertinent rater variables $y_1^*, \ldots, y_r^*$ (see Equation 2). To explicate this definition formally, first we notice that under Model (2), this average rating is

$$\begin{aligned} y_\bullet^* &= \left(y_1^* + \cdots + y_r^*\right)/r \\ &= (\lambda_1\eta + \cdots + \lambda_r\eta)/r + (\varepsilon_1 + \cdots \varepsilon_r)/r. \end{aligned} \tag{4}$$

Then the IRA index proposed in this article, denoted ρ_1, is defined as follows ($Var(\cdot)$ designates variance in the sequel):

$$\begin{aligned} \rho_1 &= Var[(\lambda_1\eta + \cdots + \lambda_r\eta)/r]/Var[(\lambda_1\eta + \cdots + \lambda_r\eta)/r + (\varepsilon_1 + \cdots + \varepsilon_r)/r] \\ &= \frac{\left(\sum_{j=1}^{r} \lambda_j\right)^2 \sigma_\eta^2}{\left(\sum_{j=1}^{r} \lambda_j\right)^2 \sigma_\eta^2 + \sum_{j=1}^{r} \sigma_{\varepsilon_j}^2}, \end{aligned} \tag{5}$$

where σ_η^2 is the factor variance whereas $\sigma_{\varepsilon_j}^2$ are the unique variances (disturbance term or residual variances; $j = 1, \ldots r$).

We observe that since the IRA index (5) instrumentally depends on the underlying latent or error-free dimension η, this index can be also interpreted as the extent of "true agreement" among the raters. This interpretation is further corroborated by noting that the observed individual ratings y_{ij} provided by the judges cannot be usually considered to be very precise (perfect) measures of examined target characteristics in the behavioral and social sciences ($i = 1, \ldots, n; j = 1, \ldots, r$).

Interrater Agreement as "Scale Reliability"

An alternative and related approach to obtaining the IRA index in Equation (5) is to formally use expressions for reliability of the "scale score," $z = y_1^* + \cdots + y_r^*$ (cf. Raykov, 2012). Since under Model (2) this score is

$$\begin{aligned} z &= y_1^* + \cdots + y_r^* \\ &= (\lambda_1\eta + \cdots + \lambda_r\eta)/r + (\varepsilon_1 + \cdots \varepsilon_r), \end{aligned} \tag{6}$$

we may view for our aims the first term in the right-hand side of Equation (6) as a "true score," and its second term as an "error score" within the well-known classical test theory decomposition of the "observed score" z. (In this section, we use apostrophes when referring verbally to z, in order to indicate that z is not the usual scale score—as it is not actually observable—but is conceptually treated as such here for the purpose of an alternative justification of the proposed IRA index 5.

With this in mind, an application of the scale reliability formula (e.g., Raykov & Marcoulides, 2011) to the expression in Equation (6) leads to

$$\begin{aligned} \rho_2 &= Var(\lambda_1\eta + \cdots + \lambda_r\eta)/Var(\lambda_1\eta + \cdots + \lambda_r\eta + \varepsilon_1 + \cdots + \varepsilon_r) \\ &= \frac{\left(\sum_{j=1}^{r}\lambda_j\right)^2 \sigma_\eta^2}{\left(\sum_{j=1}^{r}\lambda_j\right)^2 \sigma_\eta^2 + \sum_{j=1}^{r}\sigma_{\varepsilon_j}^2} \\ &= \rho_1. \end{aligned} \tag{7}$$

An implication of the equality $\rho_1 = \rho_2$ shown here is that the proposed IRA index (5) can be alternatively defined as a reliability coefficient of the "scale score" z.

Interrater Agreement as a Correlation

A third justification of the proposed IRA index (5) is provided by the following demonstration that this index equals also the degree of linear interrelationship between (a) the common factor η representing the common sources of variance in the underlying rating variables $y_1^*, \ldots, y_r^*$, and (b) their sum—the "scale score" z from the previous subsection—or alternatively their average. This degree of interrelationship is captured by the squared correlation coefficient, $[Corr(z, \eta)]^2$, of the variables in (a) and (b), where $Corr(\cdot, \cdot)$ denotes correlation. Indeed, denoting covariance by $Cov(\cdot, \cdot)$, for this squared correlation the following chain of equations hold

$$\rho_3 = [Corr(z, \eta)]^2$$

$$= \left\{ \mathrm{Cov}(z, \eta)/[Var(z)Var(\eta)]^{1/2} \right\}^2$$

$$= Var(\lambda_1\eta + \cdots + \lambda_r\eta)/Var(\lambda_1\eta + \cdots + \lambda_r\eta + \varepsilon_1 + \cdots + \varepsilon_r)$$

$$= \frac{\left(\sum_{j=1}^{r} \lambda_j\right)^2 \sigma_\eta^2}{\left(\sum_{j=1}^{r} \lambda_j\right)^2 \sigma_\eta^2 + \sum_{j=1}^{r} \sigma_{\varepsilon_j}^2}$$

$$= \rho_1 = \rho_2. \tag{8}$$

Thus, the proposed IRA index (5) is also the squared correlation between the sum of the underlying rating variables $y_1^*, \ldots, y_r^*$, on one hand and their common source of variability η, on the other. Because the correlation coefficient is invariant under linear transformation of the variable(s) involved, ρ_3 is the same correlation as that between the average of the underlying rating variables $(y_1^* + \cdots + y_r^*)/r$ and their common source of variability, η. Taking this fact into account, the developments in this section show that the proposed IRA coefficient (5) is a measure of consistency in the ratings $y_1, \ldots, y_r$ (and specifically of their underlying latent variables $y_1^*, \ldots, y_r^*$).

A by-product of the discussion thus far is also that the IRA coefficient ρ_1 equals the R^2 index associated with the regression of the sum z of the underlying rating variables $y_1^*, \ldots, y_r^*$ on their common source of variability, η.

Interrater Agreement as Interrater Reliability

In the present setting of fixed raters, it is informative to revisit the corresponding developments in Shrout and Fleiss (1979; see also Shrout & Lane, in press) with regard to their interrater reliability (IRR) index ICC(3, *r*), under the assumptions made there in its derivation. The latter index was shown by those authors to represent the reliability of the average observed rating.

It can be directly demonstrated that the IRA index (5) of this article equals the ICC(3, *r*) index in Shrout and Fleiss (1979; see also Shrout & Lane, in press) when applied on the underlying latent variables $y_1^*, \ldots, y_r^*$. To this end, let us rewrite the equation in Model (2) that pertains to the *i*th target when evaluated by the *j*th rater,

$$y_{ij}^* = \lambda_j\eta_i + \varepsilon_{ij}, \tag{9}$$

in the following useful way:

$$y_{ij}^* = \lambda_\bullet\eta_i + \left(\lambda_j - \lambda_\bullet\right)\eta_i + \varepsilon_{ij}, \tag{10}$$

where $\lambda_\bullet$ is the average factor loading in Model (2), that is, $\lambda_\bullet = (\lambda_1 + \cdots + \lambda_r)/r$ ($i = 1, \ldots, n; j = 1, \ldots, r$).

In the terminology of Shrout and Fleiss (1979), the first term on the right-hand side of Equation (10) can be interpreted as the ''true score'' associated with the ith target, considering formally y^*_{ij} as available ratings; for this reason, we use apostrophes when referring verbally to this ''true score.'' Therefore, at the underlying latent variable level, which is of interest here, we can use formally the following expression for the reliability index ICC(3, r) from Shrout and Fleiss (1979),

$$\mathrm{ICC}(3, r) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2/r}, \tag{11}$$

which they showed in their manifest variable context to represent the reliability of the average observed rating, assuming that the error variances were equal (and denoted σ_e^2 here).

For our purposes in this section, we can formally use Equation (11) with regard to the underlying latent variables, $y^*_1, \ldots, y^*_r$ in order to define IRR of the average underlying rating. We emphasize that there is no need to assume equality of the corresponding error variances within our LVM approach. When these error variances are equal, however, using the same developments as in Shrout and Fleiss (1979) that yield Equation (11), one can show that the IRA coefficient (5) in this article equals their IRR coefficient (11) when applied on the rating underlying latent variables $y_1^*, \ldots, y_r{}^*$.

To this end, we first realize that the terms in the right-hand side of Equation (11) are variance components, which were obtained from an appropriate analysis of variance on observed scores in Shrout and Fleiss (1979; (assuming there are equal error variances; see also Shrout & Lane, in press). Therefore, to use the logic behind their formula (11) within our LVM-based approach, we need to work out the ''true score'' variance and ''error'' variance with regard to the underlying latent variable scores $y^*{}_{ij}$. With this and preceding discussion in mind, this ''true score'' variance is readily found as

$$Var(\lambda_\bullet \eta_i) = \lambda_\bullet^2 Var(\eta_i) = \frac{1}{r^2}\left(\sum_{j=1}^{r} \lambda_j\right)^2 \sigma_\eta^2, \tag{12}$$

where σ_η^2 denotes the variance of the common factor in Model (2). We note that the right-hand side of Equation (12) equals the variance of the average ''rating'' $(y^*_1 + \cdots + y^*_r)/r$ associated with the underlying latent variables $y^*_1, \ldots, y^*_r$. Substituting Equation (12) into Equation (11), under the assumption of equal error variances across $y^*_1, \ldots, y^*_r$ in Model 2, denoted, say, σ_ε^2, we obtain

$$\begin{aligned}\mathrm{ICC} * (3, r) &= \frac{\left(\sum_{j=1}^{r} \lambda_j\right) \sigma_\eta^2 / r^2}{\left(\sum_{j=1}^{r} \lambda_j\right) \sigma_\eta^2 / r^2 + \sigma_\varepsilon^2 / r} \\ &= \frac{\left(\sum_{j=1}^{r} \lambda_j\right) \sigma_\eta^2}{\left(\sum_{j=1}^{r} \lambda_j\right) \sigma_\eta^2 + r\sigma_\varepsilon^2} \\ &= \frac{\left(\sum_{j=1}^{r} \lambda_j\right) \sigma_\eta^2}{\left(\sum_{j=1}^{r} \lambda_j\right) \sigma_\eta^2 + \sum_{j=1}^{r} \sigma_\varepsilon^2} \\ &= \rho_1 = \rho_2 = \rho_3. \end{aligned} \tag{13}$$

In Equation (13), ICC*(3, r) designates the IRR coefficient ICC(3, r) in Equation (11) from Shrout and Fleiss (1979) when applied to the average rating for the underlying rater latent variables $y_1^*, \ldots, y_r^*$, that is, using the latter random variables in the role of observed ratings in the Shrout and Fleiss coefficient ICC(3, r). Thus, under the assumption of equal residual variances in Model (2), the proposed IRA index (5) in this article represents the IRR coefficient when considering the rating underlying normal variables $y_1^*, \ldots, y_r^*$ as indicators of their underlying common factor η. We reiterate that the assumption of equal residual variances in Model (2) is not needed for the IRA index proposed in this article, but when that assumption holds, this index equals—as just shown—the IRR coefficient ICC(3, r) in Shrout and Fleiss (1979) applied to these underlying variables (see also Shrout & Lane, in press).

Likewise, one can also show that the IRR index (5) proposed in this article equals the longitudinal reliability coefficient for fixed time points in Laenen, Alonso, and Molenberghs (2009), if applied to the underlying rater variables $y_1^*, \ldots, y_r^*$.

Testing Identity in Rating Thresholds

As indicated earlier, testing for identity of the rater-specific thresholds under the present LVM-based approach can shed further light on the nature of IRA in an empirical study. To test the hypothesis that the set of thresholds in Equation (1) are the same across all raters, one tests for significance the difference in fit of two nested models (e.g., Muthén & Muthén, 2010). Specifically, these are (a) Model (2) and (b) the same model after imposing the constraints of equal thresholds, that is, Model (2) with the restrictions

$$\tau_{j1} = \tau_{j-1,1}; \tau_{j2} = \tau_{j-1,2}; \ldots, \text{ and } \tau_{j,m-1} = \tau_{j-1,m-1} \qquad (j = 2, \ldots, r). \tag{14}$$

The test of the null hypothesis (14) is readily carried out by fitting both models described in (a) and (b) to the analyzed rating data set and evaluating the significance of the (corrected) difference in associated chi-square values (e.g., Muthén & Muthén, 2010; see next section for an illustration).

Interval Estimation of Interrater Agreement

Because of the fact that the proposed IRA index (5) is defined as a proportion of common variance, one can readily construct a confidence interval for this index. To this end, once Model (2) is fitted and found plausible for a given data set, a large-sample standard error for the IRA index (5) can be obtained using the popular delta method in a first step (e.g., Raykov & Marcoulides, 2004). This is achieved by introducing an ''external'' parameter in Model (2), which is defined as the right-hand side of Equation (5) (see the appendixes for Mplus source code and pertinent model constraint accomplishing this aim). Using this standard error and the observation that Equation (5) is a bounded ratio of two variances that cannot be negative or larger than 1, the confidence interval procedure outlined in Raykov and Marcoulides (2011, chap. 7) can be directly applied in a second step via the logit transformation and then its inverse—the logistic transformation—to obtain a confidence interval of the proposed IRA index. This second step can be achieved using the R function ''ci.ira'' presented in Appendix B. The IRA interval estimation procedure is demonstrated in the illustration section.

Identification of Possible "Outlying" Raters

In certain empirical settings, it is possible that one or more of the judges do not evaluate the targets in a manner consistent with the remaining judges (or a certain group of them). In such cases, it is of relevance to be in a position to identify possibly ''outlying'' judges. The IRA evaluation procedure proposed in this article can be straightforwardly used to address this issue as well. Specifically, once Model (2) is fitted to the data and found plausible, one compares the factor loading confidence intervals that can be requested from the software used (e.g., Muthén & Muthén, 2010) across the raters. A judge(s) associated with a confidence interval below (above) the confidence intervals of all remaining evaluators of interest may be considered possibly aberrant in their ratings.[2]

We illustrate next the outlined IRA evaluation procedure with an example.

Illustration on Data

In this section, to demonstrate the utility and applicability of the proposed IRA index (5) and related evaluation procedures, we employ simulated data. (A main reason for using simulated data here is to have a knowledge of all underlying model parameters, and hence the resulting possibility to compare them with the results obtained using the outlined procedure in this article.) Specifically, first we generate data for n = 1,000 ‘‘targets’’ and r = 5 ‘‘raters’’ using a Likert-type scale with m = 4 possible values (1 through 4, say). The data are simulated under the following model (see Equation 2):

$$\begin{aligned} y_1^* &= .70\ \eta + \varepsilon_1, \\ y_2^* &= .75\ \eta + \varepsilon_2, \\ y_3^* &= .80\ \eta + \varepsilon_3, \\ y_4^* &= .85\ \eta + \varepsilon_4, \\ y_5^* &= .90\ \eta + \varepsilon_5, \end{aligned} \tag{15}$$

where $Var(\eta) = 1$ is the variance of a zero-mean normal variable, ε_1 through ε_5 are independent normal zero-mean residual terms with variances .5100, .4375, .3600, .2775, and .1900, respectively, while the thresholds are $\tau_{1,1} = \ldots, \tau_{5,1} = .20$, $\tau_{1,2} = \ldots, \tau_{5,2} = .50$, and $\tau_{1,3} = \ldots, \tau_{5,3} = .80$.[3] (Further details on the simulation procedure can be obtained from the authors on request.)

We start by examining the latent structure associated with the resulting data set, which allows us also to test if the thresholds are the same across the five ‘‘judges.’’ To this end, we first fit the single-factor model with categorical indicators (see Equation 2) and unconstrained corresponding thresholds for equality across ‘‘raters.’’ This model, referred to as full model below, is associated with the following tenable goodness-of-fit-indexes: chi-square (χ^2) = 6.588, degrees of freedom (df) = 5, p value (p) = .253, and root mean square error of approximation (RMSEA) = .018. (The needed Mplus source code is provided in Appendix A.) Adding the corresponding equality constraints (14) for the triple of thresholds across raters (see Note in Appendix A for needed code) leads to a tenable model as well, which is referred to as restricted model below: χ^2 = 27.020, df = 17, p = .058, RMSEA = .024. Next, to test the thresholds for identity across raters, we use the corrected difference in chi-square values, which incorporates appropriately the chi-square fit indexes and degrees of freedom of the full and restricted models (Muthén & Muthén, 2010; Satorra, 2000; see Note 2 in Appendix A for code). The associated test statistic is thereby found to be $\Delta_c\chi^2$ = 20.181, df = 12, p = .064. This result indicates plausibility of the assumption of equal thresholds across judges, which is a correct finding since this identity was built into the data generation process.

Given the plausibility of the hypothesis of equal thresholds across raters, we proceed with the estimation of the IRA coefficient (5) proposed in this article. To this end, we include in the restricted model, the definition of this coefficient (see Equation 5) in the form of an ‘‘external parameter.’’ (The needed Mplus source code is given in Appendix B; note that the pertinent ‘‘model constraint’’ does not affect the fit of the restricted model or its degrees of freedom.) Table 1 contains the

Table 1. Factor Loadings and Interrater Agreement (IRA) Index Estimates, Standard Errors, and 95% Confidence Intervals for the Illustration Example

Parameter	Estimate	*SE*	95% CI
$\lambda1$	.699	.026	[.648, .750]
$\lambda2$	.786	.022	[.743, .829]
$\lambda3$	.799	.021	[.759, .840]
$\lambda4$	.862	.017	[.828, .896]
$\lambda5$	.894	.016	[.864, .925]
ϕ	1.000	—	—
ρ_1	.905	.006	[.892, .916][a]

Note. 95% CI = confidence interval at .95 level; ϕ = Var(η) = latent variance (see Equations 15); — = not applicable (because of pertinent parameter being fixed, for identification purposes). Estimates of residual variances, not presented in this table, are obtained by subtracting from 1 the square of the associated factor loading (Muthén & Muthén, 2010; see also Note 3).

a. See Appendix C for obtaining this confidence interval (using the R function "ci.ira," with substituted IRA estimate and *SE* reported here; see also Note 2).

parameter estimates obtained with the last model, along with their standard errors and 95% confidence intervals (95% CIs). We note that the parameter estimates are close to the true values that are well covered by the associated confidence intervals.

As can be seen from Table 1, the resulting IRA index estimate is $\hat{\rho}_1 = .905$, with a standard error .006. Using the R function ''ci.ira'' in Appendix C, we obtain the large-sample 95% CI [.892, .916] for the IRA index (5). This result suggests that at the underlying common latent dimension, the degree of agreement (true agreement) among the raters in the population of targets of interest could be between the high .80s and low .90s, which one may argue is a reasonably high degree of agreement. Further inspection of the 95% CIs for factor loadings in Table 1 does not suggest any outlying rater (see also Note 2 for strict tests if needed). This finding of a lack of such raters can be expected, since the true factor loadings were relatively compactly positioned in the .70s through .90s region in the data generation process.

Because in this example we know the true parameters used for data simulation (see Equations 15 and following discussion), we can work out the population IRA coefficient by substituting them into its defining Equation (5). This yields the latter IRA index as $\rho_1 = .900$, which is very close to its reported estimate of .905 and covered by the 95% CI [.892, .916] obtained above. As an alternative approach to examining IRA, one may evaluate the IRA in this example using the popular and widely used intraclass correlation (ICC) estimation procedure in Shrout and Fleiss (1979; see also Shrout & Lane, in press), which we stress however is an observed-variable-only approach to IRA evaluation. Applying that procedure to the raw data set here directly, one obtains .845 as the ICC for rater consistency (see Shrout & Lane, in press, for the SPSS estimation input file needed thereby). This ICC estimate of IRA, namely .845, is considerably below the true agreement coefficient obtained here

(ρ_1 = .900) and is also notably below the left endpoint of its 95% CI [.892, .916]. This finding can be explained by the realization that the IRA index (5), as shown earlier, is a correlation between underlying latent variables, namely, the sum of the ''true'' ratings $y_1^*, \ldots, y_r^*$ and their common factor η. Hence, because of (5) being defined as a correlation between (error free) latent variables, whereas the ICC is a correlation between observed variables (e.g., Skrondal & Rabe-Hesketh, 2004), such an underestimation can be expected as a result of the well-known attenuation (deflation) effect of measurement error (e.g., Raykov & Marcoulides, 2011).

Discussion and Conclusion

This article was concerned with the evaluation of IRA in behavioral, educational, social, and biomedical research. A latent variable-based IRA index was proposed, and a procedure for its point and interval estimation that uses widely circulated popular software was outlined. The proposed IRA index was shown to be derivable using different analytic approaches and, at the level of latent variables, consistent with certain IRA indexes available at the manifest level. This IRA index can be interpreted also as a ''true agreement'' among the raters, given that their individual ratings cannot be often considered to be very precise (error free) in the behavioral and social disciplines.

This article contributes to the literature on IRA evaluation by proposing an index developed within an alternative framework in comparison with traditional methods of IRA estimation. The framework is inherently concerned with the underlying possible sources of rater agreement in terms of one or more underlying latent dimensions. This IRA index (5) is also instrumentally related to the query of whether raters use the same thresholds leading up to their typically ordinal data. The possibility to test for this identity is not shared with many extant IRA methods and seems to be one particular benefit of the approach adopted in this article.

Although the preceding discussion in the article was couched nearly exclusively in terms of a single latent variable (factor, dimension) underlying the raters' evaluations, it is readily extended to the case where more than one latent variables represent sources of rater agreement. To this end, one can (a) use the IRA index definition as proportion common variance in the normal variables underlying each individual rater's evaluations (e.g., Rabe-Hesketh & Skrondal, 2008) and then (b) apply the method in Raykov and Shrout (2002) to obtain point and interval estimates of this proportion, that is, the IRA index of this article in that case.

Another extension of the present IRA estimation procedure is easily obtained for the case of clustered targets, such as teachers nested within schools, patients nested within facilities, respondents nested within cities, students nested within classes, and so forth. Such nesting may be more often the case than initially thought in the IRA literature. A particular benefit stemming from adopting the LVM approach of this article resides in the possibility to address many realistic empirical settings where the objects of (coarse) measurement by a set of judges are clustered in higher order units, such as facilities, schools, classes, cities, and so on. In such circumstances, the assumption of independence of the objects, which underlies the majority of traditional IRA estimation methods, is no longer fulfilled. The resulting violations of this assumption are readily handled with this article's LVM-based procedure, namely by fitting Model (2) as a two-level model while accounting for the clustering effects of the targets within corresponding higher level units. (At the software level, this is readily achieved by using the corresponding two-level model fitting and parameter estimation approach as implemented in Mplus; Muthén & Muthén, 2010.) Similarly, under the assumption of data missing at random (MAR; e.g., Little & Rubin, 2002), the procedure of point and interval estimation of the IRA index (5) is straightforwardly applicable as well. With violations of MAR, inclusion of auxiliary variables ''predictive'' of the missing values is recommendable (e.g., Enders, 2010) , which is readily accommodated within the estimation procedure outlined in this article.

The proposed method is directly applicable also in cases where one is interested in including covariates in the analysis. Specifically, one could regress the latent variable(s) η on the covariates in the pertinent extension of Model (2) used for IRA index estimation to explain individual differences in the common continuum underlying the observed ratings. In addition, one can also examine whether individual raters' evaluations could be affected uniquely by particular covariates, that is, over and above their effect assumed than mediated by the latent variable(s). In the affirmative case (e.g., findings of significant and large/small unique regression parameters for individual raters), such a covariate-extended model may provide a possible explanation for the reasons why particular judges may be rendering inconsistent evaluations relative to the remaining raters.

Limitations of the outlined IRA evaluation approach stem from its requirement for large samples with regard to targets. We stress that there is no restriction with respect to raters—their number can be as low as 2, so long as the Model (2) is identified (which it will be then with the additional assumption of equal factor loadings). The reason is the fact that the underlying modeling method is based on an asymptotic statistical theory (Muthén, 1984). This requirement will be particularly important with a sizable number of judges. Although no specific guidelines are at present available for determining appropriate sample size, because of a multitude of rather complicated issues involved, it may be conjectured that with more than, say, 10 judges, it could be recommendable to have perhaps more than 1,000 targets being evaluated by them. Such evaluations typically occur in the context of large-scale assessment programs where multiple raters evaluate the performance of a large number of subjects, say, in educational assessments (e.g., Congdon & McQueen, 2000; Wolfe, 1996), licensure examinations (e.g., Raymond et al., 2011), and so forth. One may also submit that with fairly large fractions of missing values, the results of the outlined procedure should be interpreted with a great deal of caution. We strongly encourage future research in the direction of developing possible guidelines for sample size in relation to number of judges, targets, and fraction of missing data.

In conclusion, this article offers a readily applicable, LVM-based approach to point and interval estimation of IRA, which extends the arsenal available to empirical social and behavioral researchers involved in the measurement of personal characteristics.

Appendix A

Mplus Source Code for Full and Restricted Model When Testing Threshold Identity (14)

```
TITLE:      TESTING RATER THRESHOLD IDENTITY (EQ. (14)). FULL
            MODEL.
DATA:       FILE = <NAME OF RAW DATA FILE.;
VARIABLE:   NAMES = Y1-Y5;
            CATEGORICAL = Y1-Y5;
ANALYSIS:   ESTIMATOR = WLSMV;
MODEL:      F BY Y1 *(P1)
             Y2-Y5 (P2-P5);
            F@1;
```

Note 1. Add the following 3 lines in the MODEL section, for the restricted model

```
[ Y1$1-Y5$1] (1);
[ Y1$2-Y5$2] (2);
[ Y1$3-Y5$3] (3);
```

Note 2. To test the validity of the threshold constraints, fit the above full model with the added last command SAVEDATA: DIFFTEST = DERIVATIVES.DAT;. Then fit the restricted model adding in its ANALYSIS section only the command DIFFTEST = DERIVATIVES.DAT; (e.g., Muthén & Muthén, 2010; cf. Satorra, 2000), and examine pertinent output part for test statistic, degrees of freedom, and associated *p* value.

Appendix B

Mplus Source Code Used for Point and Interval Estimation of the Proposed Interrater Agreement Index

```
TITLE:      POINT AND INTERVAL ESTIMATION OF PROPOSED IRA INDEX
            (5).
            (SEE NOTE 1 BELOW AND APPENDIX C FOR OBTAINING ITS
            CONFIDENCE INTERVAL.)
DATA:       FILE = <NAME OF RAW DATA FILE>;
VARIABLE:   NAMES = Y1-Y5;
            CATEGORICAL = Y1-Y5;
ANALYSIS:   ESTIMATOR = WLSMV;
```

```
MODEL:     F BY Y1 *(P1)
            Y2-Y5 (P2-P5);
           F@1;
           [ Y1$1-Y5$1] (1);
           [ Y1$2-Y5$2] (2);
           [ Y1$3-Y5$3] (3);
MODEL CONSTRAINT:
           NEW(IRA);
           IRA = (P1 + P2 + P3 + P4 + P5)**2/
            ((P1 + P2 + P3 + P4 + P5)**2 +
             5-P1**2-P2**2-P3**2-P4**2-P5**2);
```

Note 1. Use the estimated standard error for the ''external parameter'' IRA and its associated estimate with the R function ''ci.ira'' in Appendix C to obtain a confidence interval for the IRA index (5) of this article. Because of Model (2) being fitted to the polychoric correlation matrix of the observed ratings (raw data; e.g., Muthén, 1984; Muthén & Muthén, 2010), the sum of the residual variances in (2) is that of the complement to 1 of each squared factor loading, as reflected in the last line of this input file (see also Note 2).

Note 2. To obtain confidence intervals for model parameters, in particular underlying factor loadings (e.g., when examining for possible ''outlying'' raters), add in this input file as last the following command: OUTPUT: CINTERVAL;

Appendix C

R Function for Interval Estimation of the Proposed Interrater Agreement Index (5)

```
ci.ira = function(ira, se){
  l = log(ira/(1-ira))
  sel = se/(ira*(1-ira))
  ci_l_lo = l-1.96*sel
  ci_l_up = l + 1.96*sel
  ci_lo = 1/(1 + exp(-ci_l_lo))
  ci_up = 1/(1 + exp(-ci_l_up))
  ci = c(ci_lo, ci_up)
  ci
  }
```

Note. At the *R* prompt (''>''), paste this function (after typing it, say, in a text-only file and then copying it). Once the Mplus input file in Appendix B is submitted to the software, as a following R command use ''ci.ira(ira,se),'' where for ''ira'' the so-obtained estimate of the IRA index is entered and as ''se'' its standard error (see third numerical column in the ''IRA'' line of the ''MODEL RESULTS'' section in the produced output by Mplus with that input file). The above R function then furnishes the sought 95% confidence interval for the IRA index (5). If a confidence interval is sought for this index at another confidence level, use in lieu of 1.96 in the 4th and 5th lines the pertinent quantile of the standard normal distribution—e.g., 1.645 if a 90% confidence interval is required (cf. Raykov & Marcoulides, 2011).

Acknowledgments

We are indebted to R. Brennan and P. Shrout for valuable discussions on interrater reliability and to T. Asparouhov for software simulation procedure details. We are grateful to B. May, J. Piert, and C. Dunbar for posing questions related to the issues discussed in this article and to two anonymous referees for valuable criticism on an earlier version of the article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The literature on interrater agreement (IRA) often emphasizes the distinction between interrater reliability (IRR) and IRA, with the former referring often to relative consistency in ratings provided by multiple judges and the latter to the equivalence of ratings in terms of their absolute value (see, e.g., LeBreton & Senter, 2008). Throughout this article, the term *interrater agreement* is used in the sense of relative agreement or relative consistency in observers' ratings, while the notation IRR is not used as we prefer to avoid the term *reliability* in this context (e.g., Shoukri, 2011).
2. The outlined confidence interval comparison does not furnish a statistical test. To render such, a corresponding procedure in Goldstein (2011) can be used, which amounts to adding and subtracting 1.39 times the standard error to each estimated factor loading before checking the resulting confidence intervals for overlap as a means of testing population loading differences.
3. The values of the error variances used are complements to 1 of the respective squared factor loadings (i.e., $Var(\varepsilon_j) = 1 - \lambda_j^2, j = 1, \ldots, r$) because Model (2) is fitted to the polychoric correlation matrix with its common factor variance assumed 1 for identification (e.g., Muthén & Muthén, 2010). As noted earlier, the equality of the corresponding thresholds across raters is not needed for the method used in this article but we introduce it in the data simulation model so as to allow an application of the earlier discussed threshold identity test in this example.

References

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*, 159-172.

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

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

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*, 163-178.

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Enders, C. A. (2010). *Applied missing data analysis.* New York: Guilford.

Goldstein, H. (2011). *Multilevel modeling*. New York, NY: Wiley.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL8 user's reference guide.* Lincolnwood, IL: Scientific Software International.

Laenen, A., Alonso, A., & Molenberghs, G. (2009). Reliability of a longitudinal sequence of scale ratings. *Psychometrika, 74*, 49-64.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions on inter-rater agreement and inter-rater reliability. *Organizational Research and Methods, 11*, 815-852.

Lindell, M. K. (2001). Assessing and testing interrater agreement in multi-item rating scales. *Applied Psychological Measurement, 25*, 89-99.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data.* New York, NY: Wiley.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, L. K., & Muthén, B. (2010). *Mplus user's guide.* Los Angeles, CA: Muthén & Muthén.

Raykov, T. (2011). Intra-class correlation coefficients in hierarchical designs: Evaluation using latent variable modeling. *Structural Equation Modeling, 18*, 73-90.

Raykov, T. (2012). Scale development using structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472-492). New York, NY: Guilford.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parametric functions in covariance structure models. *Structural Equation Modeling, 11*, 659-675.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Raykov, T., & Marcoulides, G. A. (2012). *On examining the underlying normal variable assumption in latent variable models*. Manuscript submitted for publication.

Raykov, T., & Shrout, P. (2002). Reliability of Scales with General Structure: Point and Interval Estimation Using a Structural Equation Modeling Approach. *Structural Equation Modeling, 9*, 195-212.

Raymond, M. R., Harik, P., & Clauser, B. E. (2011). The impact of statistically adjusting for rater effects on conditional standard errors of performance ratings. *Applied Psychological Measurement, 35*, 235-246.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A festschrift for Heinz Neudecker* (pp. 233-247). London, England: Kluwer Academic.

Schuster, C. (2002). A mixture model to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology, 55*, 289-303.

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*, 384-395.

Schuster, C., & Smith, D. A. (2005). Dispersion-weighted kappa: An integrative framework for metric and nominal scale agreement coefficients. *Psychometrika, 70*, 135-146.

Schuster, C., & von Eye, A. (2001). Models for ordinal agreement data. *Biometrical Journal, 43*, 795-808.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory.* Thousand Oaks, CA: Sage.

Shoukri, M. M. (2011). *Measures of interobserver agreement and reliability*. Boca Raton, FL: Chapman & Hall.

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

Shrout, P. E., & Lane, S. (in press). Reliability. In Cooper, H. (Editor-in-Chief), *APA handbook of research methods in psychology* (Vol. 1). Washington, DC: American Psychological Association.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent linear and mixed models.* Boca Raton, FL: Chapman & Hall.

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association, 80*, 175-180.

von Eye, A. A., & Mun, E. Y. (2005). *Analyzing rater agreement.* Mahwah, NJ: Erlbaum.

Wolfe, E. W. (1996, April). *A report on the reliability of a large-scale portfolio assessment for language arts, mathematics, and science*. Paper presented at the annual meeting of the National Council for Measurement in Education, New York, NY.

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association, 88*, 421-427.