BIOMETRICS 49, 823–835
September 1993

A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement

John S. Uebersax*

The RAND Corporation, 1700 Main Street, Santa Monica, California 90406-2138, U.S.A.

and

William M. Grove

Department of Psychology, University of Minnesota,
Minneapolis, Minnesota 55455, U.S.A.

SUMMARY

This article presents a latent distribution model for the analysis of agreement on dichotomous or ordered category ratings. The model includes parameters that characterize bias, category definitions, and measurement error for each rater or test. Parameter estimates can be used to evaluate rater performance and to improve classification or measurement with use of multiple ratings. A simple maximum likelihood estimation procedure is described. Two examples illustrate the approach. Although considered in the context of analyzing rater agreement, the model provides a general approach for mixture analysis using two or more ordered-category measures.

1. Introduction

Consider a forest researcher who wants to survey the health of a tree species. Because the area to survey is large, the plan is to use several field workers, each covering part of the entire area. Tree status is assessed with a simple scale of "good health," "moderate health," and "diseased or unhealthy."

Since different raters will supply ratings, it is important that they understand and use rating categories similarly; if not, there should at least be some way to account for their differences. Therefore the researcher might conduct a pilot study where a certain number of trees are rated by all raters, and use this information to compare, evaluate, and improve ratings.

The above typifies a class of rater agreement problems that often arise in biometry, social science, and—if we extend "ratings" to include diagnostic tests—medicine. We consider here useful statistical methods one might apply to such problems. Although the present concern is rating agreement, the approach is applicable in general to mixture estimation using ordered-category measures.

Analysis of agreement data is sometimes limited to the calculation of omnibus indices, such as the kappa coefficient (Cohen, 1960). Recently, however, there has been emphasis on *modeling* such data (Agresti, 1992; Uebersax, 1992). Tanner and Young (1985a, 1985b) discussed log-linear models for agreement data. Agresti (1988) and Becker (1989, 1990) discussed extensions of the log-linear approach, based on *association models*, and Darroch and McCloud (1986) described a useful approach based on *category distinguishability* and *quasisymmetry*.

There is also now extensive literature on *latent class models* of agreement (see, for example, Alvord et al., 1988; Baker, Freedman, and Parmar, 1991; Dawid and Skene, 1979; Dillon and Mulani, 1984; Espeland and Handelman, 1989; Gelfand and Solomon, 1975; Uebersax and Grove, 1990; Walter and Irwig, 1988), including special versions for ordered-category ratings (Clogg, 1979; Uebersax, 1993). Agresti and Lang (1993) recently combined latent class and quasisymmetry models to analyze rating agreement. We do not consider the advantages and disadvantages of these approaches here, but this is discussed by Uebersax (1992).

Another new development is *latent trait models* for agreement data. Uebersax (1988) discussed a

---

* *Current address*: Department of Public Health Sciences, The Bowman Gray School of Medicine of Wake Forest University, Medical Center Boulevard, Winston-Salem, North Carolina 27157-1063, U.S.A.

Key words: Agreement; Finite mixture model; Item response theory; Latent class analysis; Latent structure analysis; Latent trait analysis; Observer variation; Ordered category data; Rasch model; Reproducibility of results.

latent trait model for dichotomous ratings; similar models were considered in unpublished work by Darroch, and by Kraemer (1979), Quinn (1989), and Uebersax and Grove (1989). Everitt (1988), Everitt and Merette (1990), and Henkelman, Kay, and Bronskill (1990) presented related models which allow for ordered category as well as dichotomous ratings, but which are computationally demanding and require integration over several variables.

The model here can be viewed as a special case of the Everitt and Henkelman et al. models. The present model is simpler and more easily estimated—for example, it requires integration over only one variable—but fits actual data very well. The simplification makes it easier to use alternative models to test hypotheses about raters and ratings. The present approach also includes an explicit parameterization of rater bias, category definitions, and measurement error, leading to many useful applications.

In the psychometric literature, models similar to that here have been discussed for dichotomous items. Papers by Bock and Aitkin (1981) and Mislevy (1984) are especially profitable reading.

Section 2 presents the model and considers parameter estimation, identifiability, and the statistical evaluation and comparison of models. Section 3 discusses model applications. Section 4 gives two examples that illustrate use of the approach. The final section considers possible limitations and extensions.

2. Model

The term "rater" is used here in a general sense that includes any method for assigning rating levels. One can distinguish three basic designs used to collect agreement data. The first is a *fixed panel design*, where the same raters rate each case in a sample. The second is a *varying panel design*, where each case is rated by a separate randomly selected rater panel. The third is a *replicate measurement design*, where the same rating procedure is applied two or more times to each case. The type of design affects the form of the model. We mainly consider fixed panel and replicate measurement designs here. For dichotomous ratings, the varying panel and replicate measurement designs are mathematically equivalent. A more general discussion of the varying panel design is deferred until another occasion.

2.1 *Fixed Rater Panel*

Consider a continuous *latent trait*, denoted by θ. The latent trait is the quality that ratings assess, for example, symptom severity. We assume a population of cases and, in the present discussion, that the population contains two *case types* (we leave implicit extension to more than two case types). Case types correspond to different subgroups of cases—for example, cases with and without a disease. We do not observe a case's type directly, although, as we shall see, we may be able to estimate it from the case's ratings.

We assume normal distributions (usually overlapping) of latent trait levels for case types $c = 1$ and $c = 2$. The distributions are defined by probability density functions $g_1(\theta)$ and $g_2(\theta)$, respectively. We further define

$$f_1(\theta) = \lambda_1 g_1(\theta) \quad \text{and} \quad f_2(\theta) = \lambda_2 g_2(\theta).$$

The terms λ_1 and λ_2 denote the population prevalences of the two types, so that $\lambda_1 + \lambda_2 = 1$.

The overall probability density function of case trait levels is

$$f(\theta) = f_1(\theta) + f_2(\theta).$$

The function $f(\theta)$ defines a *finite mixture distribution* (Everitt and Hand, 1981; Titterington, Smith, and Makov, 1985), where λ_1 and λ_2 are the *mixing proportions* and $g_1(\theta)$ and $g_2(\theta)$ are the *component density functions*.

Now consider N cases, each rated by $R \geq 2$ raters on a scale with C ordered categories (we could allow different numbers of rating categories per rater, but do not here). We refer to rating categories by number, beginning with 1 for the lowest category and using successive integers for the others, and also number raters in an arbitrary order.

The *rating probability function* $p_j(k|\theta)$ gives the probability of rating category k ($k = 1, \ldots, C$) being assigned by rater j ($j = 1, \ldots, R$) for a case with latent trait level θ. Let x_{ij} denote the rating level that rater j assigns to case i. The vector $\mathbf{x}_i = \{x_{i1}, \ldots, x_{iR}\}$ describes the pattern of responses by all raters to case i. The probability of pattern $\mathbf{x}_i$ given a case with trait level θ is $\prod_j p_j(x_{ij}|\theta)$; note that this assumes independence of ratings conditional on latent trait level (conditional independence). Let π_i denote the probability of $\mathbf{x}_i$ given a randomly sampled case. Then

$$\pi_i = \int_{-\infty}^{\infty} f(\theta) \prod_{j=1}^{R} p_j(x_{ij}|\theta)\, d\theta. \tag{1}$$

We parameterize the rating probability functions with a threshold model similar to Rasch (1980) and item response (Lord and Novick, 1968; Samejima, 1969) modeling. Each rater j is assumed to have a threshold t_{jk} associated with each rating category k ($k \geq 2$). A case's apparent trait level must exceed threshold t_{jk} for rater j to use rating category k or above.

A case's apparent trait level is assumed to vary about its true level. Variation in apparent trait level, which we equate with *measurement error*, is assumed normally distributed. Under this assumption, the probability of a case's apparent trait level exceeding a given threshold is given by the normal cumulative distribution function. However, the normal cumulative distribution function is closely approximated by a logistic ogive, and the latter is computationally advantageous (Lord and Novick, 1968, p. 400). We accordingly define

$$\Psi_{jk}(\theta) = \{1 + \exp[1.7\alpha_j(t_{jk} - \theta)]\}^{-1}, \tag{2}$$

for $j = 1, \ldots, R$ and $k = 2, \ldots, C$.

The term α_j corresponds to measurement error for rater j; with use of the constant 1.7, $1/\alpha_j^2$ approximates the rater's measurement error variance. Each function $\Psi_{jk}(\theta)$ defines the probability that a case with trait level θ will exceed rater j's threshold for category k. Since we want to know the probabilities of a case's apparent trait level falling in each of the C intervals defined by the rater's thresholds, we define

$$p_j(k|\theta) = \begin{cases} 1 - \Psi_{j2}(\theta) & k = 1, \\ \Psi_{jk}(\theta) - \Psi_{j,k+1}(\theta) & 1 < k < C, \\ \Psi_{jC}(\theta) & k = C. \end{cases} \tag{3}$$

Equations (1), (2), and (3) define the main elements of the model.

The model simplifies with dichotomous ratings. Let $x_{ij} = 0$ for a negative rating and $x_{ij} = 1$ for a positive rating. We then define for each rater $\Psi_j(\theta) = \{1 + \exp[1.7\alpha_j(t_j - \theta)]\}^{-1}$, where t_j is rater j's threshold for a positive rating. The probability of rating pattern $\mathbf{x}_i$ for a randomly observed case becomes

$$\pi_i = \int_{-\infty}^{\infty} f(\theta) \prod_{j=1}^{R} \Psi_j(\theta)^{x_{ij}}[1 - \Psi_j(\theta)]^{1-x_{ij}}\, d\theta. \tag{4}$$

We term (4) with the added restriction $\alpha_1 = \cdots = \alpha_R = \alpha$ the *Rasch rating model.* The components of this model are illustrated in Figure 1.

The fixed rater panel design is probably the most common in practice. Details concerning the model for replicate measurement, of more specialized interest, are given in the Appendix.

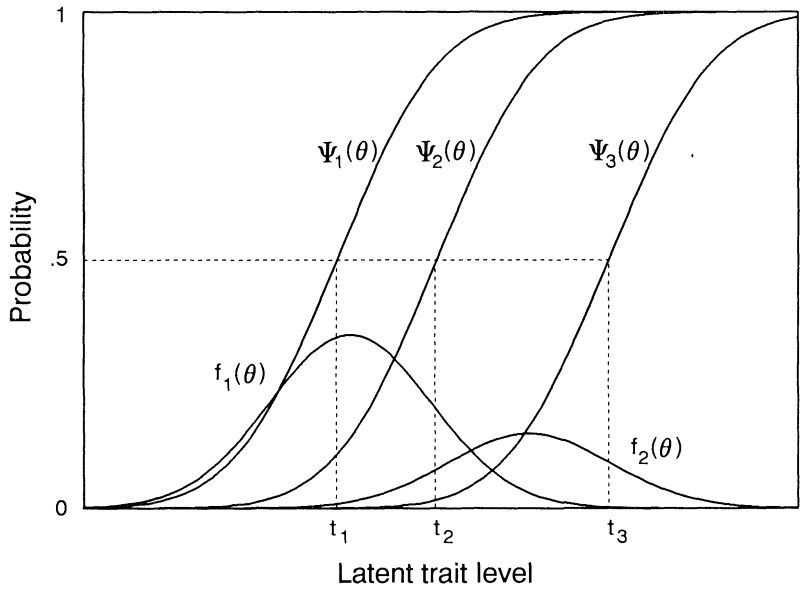

Figure 1. Components of the Rasch rating model with three raters.

2.2 *Parameter Estimation*

The parameters for the model are summarized as follows: (i) the mean and standard deviation of latent distribution $g_1(\theta)$, μ_1 and σ_1; (ii) the mean and standard deviation of latent distribution $g_2(\theta)$, μ_2 and σ_2; (iii) the mixing proportions λ_1 and λ_2; (iv) a measurement error parameter α_j for each rater; and (v) a threshold t_{jk} for each combination of rater j and rating category k ($k \geq 2$).

Since $\lambda_1 + \lambda_2 = 1$, only either λ_1 or λ_2 must be estimated. Two constraints are required to fix the scale. Since it is the distance between μ_1 and μ_2 that is important, we can define

$$\mu_1 = -\delta \quad \text{and} \quad \mu_2 = \delta$$

and estimate δ rather than both μ_1 and μ_2; this supplies one scaling constraint. As the second constraint one can fix the value of δ or a σ, α, or t parameter. Another way to fix the scale is to require that thresholds have zero mean and unit variance. To avoid convergence on a trivial solution with reverse ogives, α parameters are required to be nonnegative.

Let NT denote the number of threshold parameters and NE denote the number of measurement error parameters. For fixed panel designs $NT = R(C - 1)$; $NE = R$ if measurement error is allowed to vary across raters, and $NE = 1$ if measurement error is assumed constant. For a model with two normal latent distributions, the number of estimated parameters is NT + NE + 3, if σ_1 and σ_2 are allowed to differ, and NT + NE + 2 if $\sigma_1 = \sigma_2$ is assumed.

Parameters can be estimated by maximum likelihood. Let s index each of $S = C^R$ possible response patterns. The overall log-likelihood is

$$\ln L = \sum_s n_s \ln(\pi_s), \tag{5}$$

where π_s is the probability of rating pattern $\mathbf{x}_s$, calculated with equation (1), and n_s is the observed frequency of this pattern. The maximum likelihood parameter estimates are those that maximize the log-likelihood function. Patterns for which $n_s = 0$ do not contribute to the log-likelihood and need not be considered; this is helpful when there are many raters and rating categories.

Several algorithms can be used for estimation. For the examples here, a direct search optimization routine (Chandler, 1969) was used. A potential drawback of direct search algorithms is their slowness; to achieve good convergence (for example, a change in the log-likelihood of less than 10E-8) may require several thousand iterations. However, this is still feasible with a microcomputer.

Estimation based on the EM algorithm is also possible (Bock and Aitkin, 1981; Mislevy, 1984), although the slow convergence of the EM algorithm is well known. More efficient estimation might combine algorithms—for example, using an EM or direct search algorithm initially, and then a Newton–Raphson or similar algorithm once near-convergence occurs.

Integration is done numerically, with discrete approximation of θ. For the examples in Section 4, the latent trait continuum was represented by 101 equidistant points (quadrature points) from -10 to $+10$, inclusive. This is probably higher resolution and a wider trait range than is required—with these data, use of 26 quadrature points over the range -6 to $+6$ produces very similar results. For accurate representation of latent distributions, σ_1 and σ_2 must not be too small; a good strategy is to fix the smaller to 1, though it may require experimentation to determine which is the smaller. An upper constraint on α parameters (e.g., 10) should be used to guard against the possibility of their tending to infinity.

For a unique solution the number of estimated parameters cannot exceed $S - 1$. In some cases parameters may not be identified even though this criterion is met. This may occur for two reasons. First, R, C, and the latent distribution parameterization may be a nonidentifiable combination; the model is then not identified, regardless of observed data. Second, a potentially identifiable model may be nonidentifiable because of an unusual pattern of observed data. Also, parameters can be nearly nonidentifiable, in which case the G^2 fit index (see section below) may not follow the theoretical chi-square distribution.

A remarkable form of *partial identifiability* applies to the Rasch rating model (Lindsay, Clogg, and Grego, 1991). One can express Rasch models in log-linear form (Cressie and Holland, 1983; Kelderman, 1984). Further, such models are *necessarily* quasisymmetrical (Darroch and McCloud, unpublished manuscript). As a result, the Rasch rating model can fit data no better than the model of unconstrained quasisymmetry. With dichotomous ratings, unconstrained quasisymmetry entails $2R - 1$ parameters. If the Rasch rating model has more parameters than this, the α parameter and threshold parameters will be uniquely identified, but not the latent trait distribution parameters. This means that for the Rasch rating model $R \geq 4$ is required when one assumes $\sigma_1 = \sigma_2$ and $R \geq 5$ is required when $\sigma_1 \neq \sigma_2$ is allowed.

Identifiability is easily verified by evaluating the rank of the observed information matrix (−1 times the matrix of second derivatives of ln L relative to estimated parameters). This is convenient, since this matrix is often calculated to estimate standard errors (see Section 4.2). The matrix is of less than full rank for nonidentifiable models and/or data. The ratio of the largest to the smallest eigenvalue is an index of proper matrix conditioning; weak identifiability is indicated when this number is very large, say above 10,000. One can also detect nonidentifiability by testing different start values and seeing if they result in different solutions with identical fit.

Nonidentifiability can usually be resolved by simplifying the model or introducing plausible constraints. Identifiability has not been problematic with data so far examined. However, it may become more important if the model is extended to include more than two component distributions or multiple latent trait dimensions.

2.3 *Model Fit*

The Pearson chi-square statistic X^2 or the likelihood ratio chi-square statistic G^2 can be used to evaluate model fit. The X^2 statistic is calculated as $X^2 = \sum_s (n_s - \hat{m}_s)^2/\hat{m}_s$ and the G^2 statistic is calculated as $G^2 = 2 \sum_s n_s \ln(n_s/\hat{m}_s)$, where $\hat{m}_s = N\hat{\pi}_s$ and $\hat{\pi}_s$ denotes the probability of the sth rating outcome calculated with maximum likelihood parameter estimates. Both statistics are asymptotically distributed as χ^2, with degrees of freedom equal to $(S - 1)$ minus the number of estimated parameters.

The difference in fit between two nested models can be assessed with the difference G^2 statistic, calculated as the difference in G^2 for the two models. The difference G^2 is also asymptotically distributed as χ^2, with degrees of freedom equal to the difference in the number of estimated parameters for the two models.

3. Applications

The purpose of analyzing ratings varies from study to study. Sometimes a researcher has a set of potential raters rate an initial sample, and uses the results to select raters with desirable characteristics. Another common situation is to analyze ratings made on an initial sample and to use this information to try to improve subsequent ratings. In some cases, the goal is to better interpret panel ratings that have already been made.

One can distinguish three factors that cause raters to disagree. The first is *bias*, which we equate with a rater's thresholds being generally higher or lower than other raters' thresholds, resulting in generally higher or lower ratings. The second factor is rater differences in *category widths*, the distances between adjacent thresholds. The third is imprecision associated with the rating process, or what we have already termed *measurement error*. Model parameters provide a means to describe and quantify each of these factors.

3.1 *Describing Rater Performance*

A rater's mean threshold provides an index of overall bias. By portraying mean thresholds graphically, one can show raters their relative bias. Similarly, if threshold locations for each rater are plotted, raters can see when they have unusually wide or narrow category definitions and adjust thresholds accordingly.

An obvious use of model parameters is to select from a set of raters those whose ratings seem most consistent with the latent trait. Earlier we noted that $1/\alpha_j^2$ closely approximates measurement error variance. Let σ_θ^2 denote the variance of the latent trait (calculated from δ, λ_1, λ_2, σ_1, and σ_2). The quantity $\sigma_\theta/(\sigma_\theta^2 + 1/\alpha_j^2)^{1/2}$ then estimates the *latent correlation* between true trait level and apparent trait level for rater j. The latent correlation provides a convenient index of measurement error. Higher values indicate that a rater's judgments are mainly guided by the trait. Lower values indicate a significant contribution of random error, or that the rater is evaluating cases according to unique criteria. Usually, one would prefer raters with high latent correlations and α's. Of course, this is situation-dependent—some applications may benefit from maintaining rater diversity.

3.2 *Hypothesis Testing*

Rater differences in bias, thresholds, and measurement error can be statistically assessed by comparing the fit of unconstrained and various constrained models.

Threshold differences. One constrained model holds thresholds constant across raters—that is, that $t_{1k} = t_{2k} = \cdots = t_{Rk}$ $(k = 2, \ldots, C)$. We term this the *identical thresholds model*. If this model fits well, and if the unconstrained model does not fit significantly better, one would conclude that raters do not differ either in terms of bias or category widths.

We can separate these effects with two other constrained models. We term the first the *simple bias*

model. This assumes that interthreshold distances are the same for each rater, but that threshold locations may vary from rater to rater by a fixed amount. In other words, let Δ_j be a constant for rater j; the simple bias model states that, for any two raters j and j', $t_{j2} - t_{j'2} = t_{j3} - t_{j'3} = \cdots = t_{jC} - t_{j'C} = \Delta_j - \Delta_{j'}$. The simple bias model is nested within the unconstrained model, and the identical thresholds model is nested within the simple bias model.

We can also require that bias be constant across raters, but permit raters to have different category widths. To do so we now define Δ_j specifically as the mean of rater j's thresholds. The *equal bias model* assumes that $\Delta_1 = \cdots = \Delta_R$. The equal bias model is nested within the unconstrained model, and the identical thresholds model is nested within the equal bias model.

Comparison of the simple bias and identical thresholds models tests the effect of differential rater bias. One can alternatively test this effect by comparing the unconstrained model and the equal bias model. For both comparisons the difference G^2 statistic has $R - 1$ degrees of freedom. Comparison of the unconstrained and simple bias models tests differential category widths. This effect can also be tested by comparing the equal bias model and identical thresholds model. For both of these comparisons, the difference G^2 statistic has $(R - 1) \times (C - 2)$ degrees of freedom.

Differential measurement error. We term the requirement that $\alpha_1 = \cdots = \alpha_R$ the *equal measurement error* constraint. Comparison of a model with this constraint and the corresponding unconstrained model tests the effect of differential measurement error. The associated difference G^2 statistic has $R - 1$ degrees of freedom.

Constrained models can be used to test other hypotheses about thresholds, such as equal or symmetrical category widths within raters.

3.3 *Measurement and Classification with Multiple Ratings*

An advantage of multiple ratings is that they permit improved trait measurement. Let $\mathrm{E}(\theta|\mathbf{x}_s)$ denote the expected level of θ given observed rating pattern $\mathbf{x}_s$. Then

$$\mathrm{E}(\theta|\mathbf{x}_s) = \frac{\int_{-\infty}^{\infty} \theta f(\theta) \prod_j p_j(x_{sj}|\theta)\, d\theta}{\int_{-\infty}^{\infty} f(\theta) \prod_j p_j(x_{sj}|\theta)\, d\theta}, \tag{6}$$

and $\mathrm{E}(\theta|\mathbf{x}_s)$ can be taken as a *latent trait score* for a case with pattern $\mathbf{x}_s$. This score should provide a better estimate of the latent trait than, for example, assigning integers to rating categories and averaging across raters because it takes differences in category widths and rater bias into account.

A related application is case classification using multiple ratings. Let $\pi_{s,c}$ denote the joint probability that a case belongs to case type c ($c = 1, 2$) and receives rating pattern $\mathbf{x}_s$. We calculate $\pi_{s,c}$ by

$$\pi_{s,c} = \int_{-\infty}^{\infty} f_c(\theta) \prod_{j=1}^{R} p_j(x_{sj}|\theta)\, d\theta. \tag{7}$$

We can also calculate the conditional probability that a case belongs to type c given observed rating pattern $\mathbf{x}_s$ as $\pi_{c|s} = \pi_{s,c}/\pi_s$.

The $\pi_{s,c}$ terms correspond to what Lazarsfeld and Henry (1968) termed *recruitment probabilities.* To maximize correct classifications, one classifies each case as type 1 or type 2 according to whether $\pi_{s,1}$ or $\pi_{s,2}$ is larger. It is relatively simple to estimate various proportions of cases correctly and incorrectly classified (see following section) by this method. The relative value of correct and incorrect classifications can also be considered to classify cases to maximize expected utility.

One could extend this approach to develop *adaptive rating strategies.* For example, each case could be initially rated by two raters. Only if the case cannot be classified with sufficient accuracy would other—perhaps more expert—raters be used.

3.4 *Estimating Sensitivity, Specificity, and Predictive Validity*

A potential advantage of latent structure modeling of agreement is that it may permit estimation of rating accuracy in the absence of a definitive criterion. Four common rating accuracy indices are *sensitivity* (*Se*), *specificity* (*Sp*), *positive predictive validity* (*Pv*+), and *negative predictive validity* (*Pv*−). These indices, applicable when there are two case types (e.g., positives and negatives) and dichotomous ratings, are defined as conditional probabilities. Sensitivity is the probability of a positive rating given a positive case. Specificity is the probability of a negative rating given a negative case. Positive predictive validity is the probability of a positive case given a positive rating. Negative predictive validity is the probability of a negative case given a negative rating.

Uebersax (1988) showed how the parameters of the dichotomous latent mixture agreement model can be used to estimate these indices for replicate measurement or varying panel designs. Table 1

Table 1
Formulas for indices of estimated rating accuracy given fixed panel model

Accuracy index	Formula
Sensitivity (Se')	$\int_{-\infty}^{\infty} g_2(\theta)\Psi_j(\theta)\, d\theta$
Specificity (Sp')	$\int_{-\infty}^{\infty} g_1(\theta)[1 - \Psi_j(\theta)]\, d\theta$
Positive predictive validity ($Pv+'$)	$\dfrac{\int_{-\infty}^{\infty} f_2(\theta)\Psi_j(\theta)\, d\theta}{\int_{-\infty}^{\infty} f(\theta)\Psi_j(\theta)\, d\theta}$
Negative predictive validity ($Pv-'$)	$\dfrac{\int_{-\infty}^{\infty} f_1(\theta)[1 - \Psi_j(\theta)]\, d\theta}{\int_{-\infty}^{\infty} f(\theta)[1 - \Psi_j(\theta)]\, d\theta}$

Note: As in Figure 1, $f_2(\theta)$ corresponds to the positive case type (i.e., $\mu_2 > \mu_1$).

shows comparable formulas for Se', Sp', $Pv+'$, and $Pv-'$ for fixed panel designs. We add primes to emphasize that these estimate rating accuracy relative to latent case type; for example, Se' is the probability of a positive rating given a case that belongs to the "positive" latent type 2. How well these values correspond to true rating accuracy depends on how closely the inferred case types correspond to true positive and negative status.

From these indices other measures of rating accuracy can be obtained. For example, the proportion of cases that are *false positives* is estimated as $\lambda_1(1 - Sp')$, the proportion of *false negatives* as $\lambda_2(1 - Se')$, and the total proportion of misclassified cases by the sum of these two numbers. The maximum possible proportion of correct ratings is estimated by the sum of the area under $f_1(\theta)$ from $-\infty$ to y and the area under $f_2(\theta)$ from y to ∞, where y is the latent trait level where $f_1(\theta)$ and $f_2(\theta)$ intersect—that is, where $f_1(y) = f_2(y)$.

The formulas in Table 1 are for individual raters, but we can extend the approach to estimate the accuracy of various aggregate decision rules. For example, consider a rule that requires unanimous positive ratings by R raters to classify a case as positive. We would then define a new rating probability function

$$p^*(\theta) = \prod_{j=1}^{R} \Psi_j(\theta). \tag{8}$$

The formulas in Table 1—with $p^*(\theta)$ used in place of $\Psi_j(\theta)$—estimate Se', Sp', $Pv+'$, and $Pv-'$ for this rule.

The formulas in Table 1 can be used with ordered-category ratings by recoding the ratings to dichotomies.

4. Examples

4.1 *Example* 1

Henkelman et al. (1990) presented (in their Figure 1) data for 298 cases examined for liver metastases with three imaging techniques. The three techniques were magnetic resonance imaging, radionuclide scintigraphy, and computed tomography; we refer to these as Test 1, Test 2, and Test 3, respectively. Results of each test were translated to an ordered-category scale with five levels of increasing evidence of metastases. Table 2 shows the frequency of various results.

Henkelman et al. fit a multidimensional mixture model to these data by iteratively reweighted least squares. We show here that it is possible to represent the data more simply. We also demonstrate how, by statistically comparing the fit of nested models, we can test useful hypotheses.

With sparse data such as these, X^2 and G^2 may not be appropriate to statistically test model fit. We therefore temporarily collapse the three middle categories to create a less sparse $3 \times 3 \times 3$ table. To this we fit a model with two normal distributions with $\sigma_1 = \sigma_2$; as a further precaution in testing this model's fit, cells are combined so that no expected frequency is less than 1. Values of $X^2 = 9.93$ and $G^2 = 11.93$ with 12 degrees of freedom (df) result; the latent trait finite mixture approach therefore appears appropriate for these data.

We now consider several models with the full data. First are two simple models that assume equal measurement error across tests—that is, $\alpha_1 = \alpha_2 = \alpha_3$. Model M_1 assumes a single normal distribution. Model M_2 assumes two normal distributions with $\sigma_1 = \sigma_2 = 1$. We also consider two variations of Model M_2. Model M_3 is otherwise the same as M_2, but permits different measurement error across tests. Model M_4 adds to Model M_2 the assumption of identical thresholds across tests.

Table 2
Cross-classification of results of three diagnostic tests from Henkelman, Kay, and Bronskill (1990)

Test 2 rating level	Test 1 rating level: 1, Test 3 rating level					2, Test 3 rating level					3, Test 3 rating level				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	36	20	2	0	0	10	7	2	0	0	1	3	0	0	0
2	22	14	3	1	0	7	7	1	4	0	1	0	0	0	0
3	3	2	1	0	0	0	0	0	0	1	0	0	0	0	0
4	1	0	0	1	0	3	0	1	0	1	0	1	1	1	1
5	3	0	0	0	1	1	0	0	1	1	0	0	1	0	4

Test 2 rating level	Test 1 rating level: 4, Test 3 rating level					5, Test 3 rating level				
	1	2	3	4	5	1	2	3	4	5
1	2	1	0	1	0	1	1	0	0	1
2	0	3	1	2	0	0	2	0	1	1
3	0	1	1	0	1	0	0	0	0	0
4	1	1	0	0	1	1	1	0	3	9
5	0	1	0	2	5	3	1	1	11	66

The results are summarized in Table 3. It is not unusual that, as with Models M_2 and M_3, a decrease in G^2 is accompanied by an increase in X^2—this is possible because maximum likelihood estimation minimizes G^2 but not X^2. However, the large change in X^2 relative to the small change in G^2 for these models—probably the effect of the sparse data—appears to support our initial caution concerning these statistics. Simulation studies (for example, Agresti and Yang, 1986) suggest that the difference G^2 test can be used even with relatively sparse data.

For Model M_1, $G^2 = 163.68$ with 13 estimated parameters and $124 - 13 = 111$ df. For Model M_2, $G^2 = 111.55$ with $124 - 15 = 109$ df. The difference G^2 test is generally considered not appropriate to compare models that differ only in the number of component distributions (Titterington et al., 1985), but it is impressive that Model M_2 reduces the G^2 statistic by 52 on 2 df.

Comparison of Models M_2 and M_3 tests the effect of differential measurement error. The difference G^2 statistic is equal to .60, with $17 - 15 = 2$ df. We therefore find little evidence that measurement error differs across tests. The maximum likelihood estimates (MLEs) for Model M_2 and their estimated asymptotic standard errors are shown in Table 4. We discuss estimation of standard errors in Example 2. The α value of .575 corresponds to a latent correlation of .87, which seems acceptably high.

Comparison of Models M_2 and M_4 tests the effect of differing thresholds among the tests. The difference G^2 is 12.45 with $15 - 7 = 8$ df ($P > .1$). Threshold differences among the tests are therefore statistically nonsignificant. Other comparisons with models that add the simple bias and equal bias restrictions to M_2 further indicate that the tests do not differ significantly either in terms of overall bias or category widths.

Table 3
Description and fit of four models applied to data in Table 2

Model	Description	df	X^2	G^2
M_1	One normal distribution	111	138.89	163.68
M_2	Two normal distributions	109	112.70	111.55
M_3	Two normal distributions and different measurement error across tests	107	118.87	110.95
M_4	Two normal distributions and identical thresholds	117	123.40	124.00

Table 4
Parameter estimates and estimated standard errors (e.s.e.) *for Model* M_2 *of Table* 3 *applied to data in Table* 2

Parameter	Estimate	Asymptotic e.s.e.	Parameter	Estimate	Asymptotic e.s.e.
δ	2.899	.4331	t_{22}	−2.966	.4670
λ_1	.601	.0311	t_{23}	−.783	.2757
λ_2	.399	.0311	t_{24}	−.301	.2671
α	.575	.0959	t_{25}	1.165	.3309
t_{12}	−2.353	.3960	t_{32}	−2.799	.4472
t_{13}	−.753	.2716	t_{33}	−.778	.2734
t_{14}	−.132	.2649	t_{34}	−.028	.2668
t_{15}	1.023	.3184	t_{35}	1.525	.3529

Note: $\mu_1 = -\delta$, $\mu_2 = \delta$; $\sigma = 1$ (fixed).

Whereas the model of Henkelman et al. used 25 parameters, we are able to represent the data fairly well with as few as 7 parameters. The results here suggest that the tests are very similar in terms of measurement error, bias, and category definitions. Though we do not pursue it here, an interesting question with these data is the comparative accuracy of diagnosis based on single tests and various combinations of tests. This could be examined with straightforward application of the methods in Sections 3.3 and 3.4.

4.2 *Example* 2

For the second example, we analyze ratings of photofluorograms for evidence of tuberculosis by groups of eight diagnosticians. These data were originally reported by Yerushalmy (1956) and previously analyzed by Kraemer (1982) and Uebersax and Grove (1990). The first two columns of Table 5 summarize the frequencies of cases with various numbers of positive ratings.

The data correspond to a varying panel design with dichotomous ratings, so the methods and formulas in the Appendix apply. We consider three models. The first, Model M_1, is a latent class model with two latent classes, which we consider for comparison; the two-latent class model is equivalent to a special case of the two-latent distribution model where $\sigma_1 = \sigma_2 = 0$ and δ is fixed arbitrarily. Model M_2 assumes one case type with normally distributed latent trait levels. Model M_3 assumes two case types with normal distributions and $\sigma_1 = \sigma_2 = 1$. The bottom of Table 5 shows the fit of the three models.

For Model M_1, $G^2 = 528.50$, with three estimated parameters and $8 - 3 = 5$ df; the deviation of this model from the observed data is substantial. For Model M_2, $G^2 = 157.67$, with $8 - 2 = 6$ df. Fit is again poor. Model M_3, however, fits the data very well, with $G^2 = 2.38$ and $8 - 4 = 4$ df. Table 6 shows parameter estimates, estimated rating accuracy, and standard errors for Model M_3.

Table 5
Observed frequencies for Yerushalmy (1956) *tuberculosis data and expected frequencies for three models*

Number of positive ratings	Observed frequency	Expected frequency		
		Model M_1	Model M_2	Model M_3
0	13,560	13,453	13,588	13,561
1	877	1,090	730	870
2	168	45	227	177
3	66	25	115	66
4	42	55	71	36
5	28	80	50	27
6	23	72	37	28
7	39	38	29	37
8	64	9	22	64
	X^2	874.20	178.42	2.37
	G^2	528.50	157.67	2.38
	df	5	6	4

Note: Model M_1, two latent classes; Model M_2, one normal distribution; Model M_3, two normal distributions; $N = 14{,}867$; expected frequencies rounded to nearest integer.

Table 6
Parameter and rater accuracy estimates and estimated standard errors (e.s.e.) *for Model* M_3 *of Table* 5

Parameter	Estimate	Asymptotic e.s.e.	Jackknife e.s.e.
δ	1.942	.0839	.0862
λ_1	.988	.0018	.0018
λ_2	.012	.0018	.0018
α	1.148	.0549	.0563
t	1.132	.1187	.1235
Accuracy index			
Se'	.728	—	.0479
Sp'	.987	—	.0007
$Pv+'$	.409	—	.0355
$Pv-'$	.996	—	.0010

Note: $\mu_1 = -\delta$, $\mu_2 = \delta$; $\sigma = 1$ (fixed).

Standard errors were estimated by two methods. Estimated asymptotic standard errors were obtained as the square roots of the diagonal elements of the inverse of the observed information matrix. Derivatives were estimated with finite differences.

The second method estimated standard errors using the delete-one jackknife procedure (Efron, 1982); this also provides estimated standard errors for Se', Sp', $Pv+'$, and $Pv-'$. As shown, the asymptotic and jackknife standard error estimates for model parameters are very close.

Finally, we use the parameter estimates of Model M_3 to estimate the accuracy obtainable by requiring various numbers of unanimous positive ratings to classify a case as positive. Table 7 shows accuracy indices for decision rules that require unanimous positive ratings by panels of from one to five raters. As shown, Se' and $Pv-'$ decrease as the criterion for positive classification becomes more stringent, whereas Sp' and $Pv+'$ increase. To illustrate how we might use this information, suppose that, to test a new treatment, one requires patients with a high probability of being positive. One would therefore want a classification procedure with a high $Pv+'$. Table 7 shows that requiring unanimous positive ratings by two raters gives much higher $Pv+'$ than requiring a positive rating by a single rater. Still higher levels of $Pv+'$ result from requiring unanimous positive ratings by three and four raters. Beyond this, however, increases in $Pv+'$ are less impressive and must be weighed against the cost of additional raters.

Table 7
Estimated accuracy of decision rules that require various numbers of unanimous positive ratings to classify a case as positive, using parameter estimates for Model M_3 *of Table* 5

Number of unanimous positive ratings required	Estimated diagnostic accuracy			
	Se'	Sp'	$Pv+'$	$Pv-'$
1	.728	.987	.409	.996
2	.601	.998	.811	.995
3	.523	.999	.918	.994
4	.467	1.000	.954	.993
5	.425	1.000	.970	.993

We consider classification based on unanimous agreement merely as an example. Clearly there are other decision criteria of the form m_1-out-of-m_2, where m_2 is the number of raters and m_1 is the required number of positive (or negative) ratings, that one might consider (Gelfand and Solomon, 1975; Uebersax, 1988).

5. Discussion: Limitations and Extensions

The present model makes some fairly strong assumptions—for example, that latent distributions are normal. Because model fit is statistically tested, there is some assurance that these assumptions will not be accepted when they are very inconsistent with the data. Clearly other distributional forms can be considered.

Begg and Metz (1990) noted several potential limitations of latent distribution agreement models. For example, it may be difficult to identify a two-distribution model when the distributions are similar. A good strategy, therefore, is to initially test a one-distribution model, and if that provides adequate fit, to not consider a two-distribution model. However, one then does not obtain some of the advantages of a two-distribution model, such as a simple method for estimation of Se', Sp', etc.

The latent trait approach evaluates rating precision based on agreement with a *latent consensus.* A rater who tends to disagree with other raters will appear less accurate than others, even though the rater may be more accurate. It is important for the researcher to recognize that the latent trait may not be the same as the trait of interest—it may also reflect inappropriate criteria that raters share. Henkelman et al. (1990) found their approach led to conclusions similar to those obtained from analysis of an external criterion. There is a need for additional validation studies that compare the results of latent trait agreement models with criterion measures.

Agreement studies often have limited sample sizes. In such cases it would probably be best to restrict attention to some of the simpler models discussed, for example, the simple bias model with equal measurement error.

One way to express the present model is $y_j = \theta + \varepsilon_j$, where θ is the latent trait level of a case, y_j is its apparent trait level for rater j, and ε_j is measurement error for rater j. As shown by Bock and Aitkin (1981), the model easily generalizes to multiple latent trait dimensions as $y_j = \sum_m w_{jm}\theta_m + \varepsilon_j$, where θ_m is a case's level relative to dimension m and w_{jm} is rater j's weight for dimension m. For instance, raters might base judgments of disease severity on two different factors (e.g., size and brightness of lesions on an image) but different raters may weight the factors differently. A multidimensional model should still not be difficult to estimate, since the complexity of integration depends only on the number of latent trait dimensions, which would ordinarily be few.

Interestingly, the approach here can be viewed as a two-tiered latent structure model. A case's latent trait level does not tell, at least with certainty, type membership. In effect, latent trait level is a proximal or "manifest" latent variable, whereas type membership is a deeper or "latent" latent variable. Located latent class models, which have received recent attention (for example, Formann, 1992; Lindsay et al., 1991; Uebersax, 1993) associate each latent class with a specific latent trait level, so that there is no distinction between case type and trait level. A systematic integration of located latent class and latent distribution models is a challenge for future research.

Acknowledgements

The authors are indebted to Clifford Clogg, John Darroch, Mark Espeland, R. Mark Henkelman, Daniel Relles, Patrick Shrout, and Stephen Walter for comments on earlier versions of this paper. The authors also thank the editor for helpful suggestions and an anonymous reviewer for valuable contributions.

Résumé

Cet article présente un modèle de distribution latente pour l'analyse d'agrément sur des notations en classes dichotomiques ou ordonnées. Le modèle comprend des paramètres qui caractérisent le biais, les définitions des classes, et l'erreur de mesure pour chaque notateur ou chaque test. On peut utiliser des estimateurs des paramètres pour apprécier la performance de la notation et améliorer la classification ou la mesure en utilisant des évaluations multiples. Une estimation du maximum de vraisemblance simple est décrite. Deux exemples illustrent l'approche. Bien que considéré dans le contexte d'analyse d'agrément de notation, le modèle donne une approche générale pour l'analyse de mélanges utilisant les mesures de deux ou plus classes ordonnées.

References

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics* **44,** 539–548.

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research* **1,** 201–218.

Agresti, A. and Lang, J. B. (1993). Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* **49,** 131–139.

Agresti, A. and Yang, M. (1986). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis* **5,** 9–21.

Alvord, W. G., Drummond, J. E., Arthur, L. O., Biggar, R. J., Goedert, J. J., Levine, P. H., Murphy, Jr., E. L., Weiss, S. H., and Blattner, W. A. (1988). A method for predicting individual HIV infection status in the absence of clinical information. *AIDS Research and Human Retroviruses* **4,** 295–304.

Baker, S. G., Freedman, L. S., and Parmar, M. K. B. (1991). Using replicate observations in observer agreement studies with binary assessments. *Biometrics* **47,** 1327–1338.

Becker, M. P. (1989). Using association models to analyse agreement data: Two examples. *Statistics in Medicine* **8,** 1199–1207.

Becker, M. P. (1990). Quasisymmetric models for the analysis of square contingency tables. *Journal of the Royal Statistical Society, Series B* **52,** 369–378.

Begg, C. B. and Metz, C. E. (1990). Consensus diagnoses and gold standards. *Medical Decision Making* **10,** 29–30.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46,** 443–459.

Chandler, J. P. (1969). STEPIT—Finds local minima of a smooth function of several parameters. Computer program abstract. *Behavioral Science* **14,** 81–82.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research* **8,** 287–301.

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

Cressie, N. and Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika* **48,** 129–142.

Darroch, J. N. and McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics* **28,** 371–388.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* **28,** 20–28.

Dillon, W. R. and Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research* **19,** 438–458.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Espeland, M. A. and Handelman, S. L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* **45,** 587–599.

Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters* **6,** 305–309.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions.* New York: Chapman and Hall.

Everitt, B. S. and Merette, C. (1990). The clustering of mixed-mode data: A comparison of possible approaches. *Journal of Applied Statistics* **17,** 283–297.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association* **87,** 476–486.

Gelfand, A. E. and Solomon, H. (1975). Analyzing the decision-making process of the American jury. *Journal of the American Statistical Association* **70,** 305–310.

Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990). Receiver operator characteristic (ROC) analysis without truth. *Medical Decision Making* **10,** 24–29.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika* **49,** 223–245.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* **44,** 461–472.

Kraemer, H. C. (1982). Estimating false alarms and missed events from interobserver agreement: Comment on Kaye. *Psychological Bulletin* **92,** 749–754.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86,** 96–107.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, Massachusetts: Addison-Wesley.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika* **49,** 359–381.

Quinn, M. F. (1989). Relation of observer agreement to accuracy according to a two-receiver signal detection model of diagnosis. *Medical Decision Making* **9,** 196–206.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests,* 2nd edition. Chicago: University of Chicago Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph* **17.**

Tanner, M. A. and Young, M. A. (1985a). Modelling agreement among raters. *Journal of the American Statistical Association* **80,** 175–180.

Tanner, M. A. and Young, M. A. (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin* **98,** 408–415.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

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

Uebersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* **88,** 421–427.
Uebersax, J. S. (1992). A review of modeling approaches for the analysis of observer agreement. *Investigative Radiology* **17,** 738–743.
Uebersax, J. S. and Grove, W. M. (1989). Latent structure agreement analysis. RAND Note N-3029-RC. Santa Monica, California: The RAND Corporation.
Uebersax, J. S. and Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine* **9,** 559–572.
Walter, S. D. and Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: A review. *Journal of Clinical Epidemiology* **41,** 923–937.
Yerushalmy, J. (1956). The importance of observer error in the interpretation of photofluorograms and the value of multiple ratings. *International Tuberculosis Yearbook* **26,** 110–124.

Received March 1991; *revised February and May* 1992; *accepted May* 1992.

APPENDIX

Replicate Measurement

We now assume that each case is rated R times by the same rater or procedure. Case i's ratings are summarized by a *summary rating vector* $\mathbf{v}_i = \{v_{i1}, \ldots, v_{iC}\}$, where v_{ik} $(k = 1, \ldots, C)$ is the number of times the case is assigned category k. Note that $\sum_k v_{ik} = R$.

Let π_i now denote the probability of summary rating vector $\mathbf{v}_i$ for a randomly sampled case, obtained as

$$\pi_i = o_i \int_{-\infty}^{\infty} f(\theta) \prod_{k=1}^{C} p(k|\theta)^{v_{ik}}\, d\theta. \tag{A.1}$$

The term $p(k|\theta)$ is the probability of rating level k given trait level θ. The term o_i is the number of different orderings of ratings that result in summary rating vector $\mathbf{v}_i$, given by the multinomial formula

$$o_i = \binom{R}{v_{i1}, \ldots, v_{iC}} = \frac{R!}{v_{i1}! \ldots, v_{iC}!}.$$

For $C > 2$, we obtain $p(k|\theta)$ from equations (2) and (3), but eliminate the j subscript throughout. For dichotomous ratings equation (A.1) simplifies to

$$\pi_i = \binom{R}{n} \int_{-\infty}^{\infty} f(\theta)\Psi(\theta)^n[1 - \Psi(\theta)]^{R-n}\, d\theta, \tag{A.2}$$

where n is the number of positive ratings (i.e., $\mathbf{v}_i = \{R - n, n\}$). Here we define

$$\Psi(\theta) = \{1 + \exp[1.7\alpha(t - \theta)]\}^{-1},$$

where α corresponds to the procedure's measurement error and t is the threshold for a positive rating.

The parameters for the replicate measurement model are the same as for fixed panels, except that there is only one measurement error parameter and one set of rating category thresholds; accordingly, NE = 1 and NT = $C - 1$.

Let s now index each of $S = (C + R - 1)!/[R!(C - 1)!]$ possible unique summary rating vectors. Equation (5) for ln L again applies, but π_s and n_s are now the probability and observed frequency of summary rating vector $\mathbf{v}_s$, and π_s is obtained from equation (A.1). The X^2 and G^2 statistics are calculated as described in Section 2.3.

For varying panels with dichotomous ratings, one might assume that thresholds are normally distributed across raters. If so, the distance between the apparent trait level of a case and the threshold of a randomly selected rater is also normally distributed. This leads to a model identical in form to equation (A.1). There are, however, two qualifications. First, with varying panels the interpretation of measurement error changes, since it now includes threshold variation associated with rater selection. Second, this does not account for possible dependencies between cases due to overlap in their raters. If the rater pool is large relative to panel size and raters are sampled randomly, such dependence can be expected to be negligible. If the rater pool is not large, an alternative would be to view the entire pool as a fixed panel, treating as missing rater × case combinations that do not occur—this assumes rater identities are known.