Statistical Modeling of Expert Ratings on Medical Treatment Appropriateness

JOHN S. UEBERSAX*

This article uses latent structure analysis to model ordered category ratings by multiple experts on the appropriateness of indications for the medical procedure carotid endarterectomy. The statistical method used is a form of located latent class analysis, which combines elements of latent class and latent trait analysis. It assumes that treatment indications fall into distinct latent classes, with each latent class corresponding to a different level of appropriateness. The appropriateness rating of a treatment indication by a rater is assumed determined by the latent class membership of the indication, rating category thresholds of the rater, and random measurement error. The located latent class model has two alternative forms: a normal ogive form, which derives from the assumption of normally distributed measurement error, and a logistic approximation to the normal form. The approach has the following advantages for the analysis of ordered category ratings by multiple experts: (1) it assesses whether different raters base ratings on the same or different criteria; (2) it assesses rater bias—the tendency of some raters to make higher or lower ratings than others; (3) it characterizes rater differences in rating category definitions; (4) it provides theoretically based methods for combining the ratings of different raters; and (5) it provides a description of the distribution of the latent trait. The data examined are appropriateness ratings on 848 indications for carotid endarterectomy made by nine medical experts. The located latent class approach provides unique insights concerning the data. It identifies what appears to be a set of clear nonindications for carotid endarterectomy, but a corresponding set of clear indications is not evident. The results indicate that all raters measured a common latent trait of treatment appropriateness, but that some measured the trait better than others. Rater differences in overall bias and rating category definitions are evident. Two methods are used to combine raters' ratings. One uses ratings to calculate a continuous appropriateness score for each indication. The other uses ratings to assign indications to discrete outcome categories, each corresponding to a specific level of appropriateness. The located latent class approach for ordered category measures has possible applications besides the analysis of expert ratings, such as item analysis. Potential extensions of the model are discussed.

KEY WORDS: Expert ratings; Health care; Latent class analysis; Latent trait analysis; Observer agreement; Ordered category data.

1. INTRODUCTION

There is much recent concern about ensuring the appropriateness of medical treatment. One proposed strategy for establishing treatment standards is to have expert panels review and rate the appropriateness of various indications for a given medical procedure.

The use of multiple expert raters raises several important measurement issues. Do different raters have the same or different criteria for treatment appropriateness? Do some raters tend to make higher or lower ratings than others? If ordered category ratings are used, do different raters attach the same meanings to the rating categories? How should the ratings of different raters be combined? Analysis of panel rating data often focuses narrowly on whether or how much raters tend to agree and neglects many of these important questions.

This article uses latent structure methods to study ordered category expert ratings on treatment appropriateness. We specifically examine ratings for the carotid endarterectomy procedure. In carotid endarterectomy, used to prevent stroke, the surgeon inserts a catheter into the carotid artery and manipulates it to remove obstructive plaque. Although this procedure is generally effective, there are concerns about possible side effects and overuse. The data come from a comprehensive study by Brook and colleagues (Park et al. 1986), who had panels of national experts rate the appropriateness of several medical treatments.

Because of space limitations, only a brief synopsis of statistical methods for the analysis of multiple expert rating data with ordered categories is provided. More extensive discussion of this literature was provided by Agresti (1992) and Uebersax (1991, 1992).

A version of the kappa coefficient to assess agreement on ordered category ratings was presented by Cohen (1968). Tanner and Young (1985) discussed log-linear models for agreement with ordered categories; Agresti (1988) and Becker (1989, 1990) described related methods based on Goodman's (1986) association models. Darroch and McCloud (1986) examined quasi-symmetry agreement models, and Agresti and Lang (1993) discussed an approach that combines quasi-symmetry and latent class models.

With dichotomous ratings, Darroch and McCloud's and Agresti and Lang's models produce results equivalent to a special case of the present approach—the special case is equal measurement error across raters and a logistic response function. Unlike the present approach, these other methods do not permit the representation of rater differences in measurement error and do not provide a simple way to combine multiple ratings into a score that measures the trait of interest.

The present approach is an example of *latent structure analysis* (Lazarsfeld and Henry 1968) and contains elements of latent class analysis (Goodman 1974; Haberman 1979), item–response theory (IRT; Lord and Novick 1968), and

* John S. Uebersax is Associate Professor, Department of Public Health Sciences, The Bowman Gray School of Medicine of Wake Forest University, Winston-Salem, NC 27157-1063. The author thanks the editor, an associate editor, two anonymous referees, and Clifford Clogg for valuable suggestions concerning the article and Rolla E. Park for providing the carotid endarterectomy rating data. This research was partly supported by the RAND Corporation.

Journal of the American Statistical Association
June 1993, Vol. 88, No. 422, Applications and Case Studies

Rasch (1980) modeling. Latent class models for multiple rater data have received increasing attention. Gelfand and Solomon (1975), Walter and Irwig (1988), Espeland and Handelman (1989), and Uebersax and Grove (1990), among others, have discussed such models for dichotomous ratings. Dawid and Skene (1979) and Dillon and Mulani (1984) proposed similar models for unordered polytomous ratings, and Clogg (1979) considered simple latent class models for ordered category ratings.

Uebersax (1988) and Uebersax and Grove (1989, 1993) discussed a latent distribution model for multiple rater ordered category data. Quinn (1989) and Henkelman, Kay, and Bronskill (1990) derived similar approaches from signal detection theory. These models regard the characteristic that ratings assess as a continuous latent trait. But there are reasons why one may instead want to view the trait as having discrete levels.

First, the trait may actually be discrete; for example, a disease with distinct stages. Second, some applications require classification of cases into specific outcome categories. Third, discrete latent trait models can avoid potentially restrictive distributional assumptions; for example, that the trait is normally distributed. Finally, discrete models usually require less computation.

The approach here draws on recent advances in the integration of latent class and latent trait models (Dayton and Macready 1988; Formann 1985, 1992; Kelderman and Macready 1990; Lindsay, Clogg, and Grego 1991; Mislevy and Verhelst 1990; Rost 1988). Rost's models are especially relevant, though they differ in some ways from the approach here.

Rost's approach is based on the polytomous Rasch model, whereas the approach here uses the polytomous IRT model (Samejima 1969). Both approaches use located thresholds to define response probabilities. With the polytomous IRT model, response probabilities are determined by the proportions of a pdf that fall between successive thresholds. With the polytomous Rasch model, response probabilities are given by the probability of a case exceeding threshold $t + 1$ given that it exceeds threshold t. For further discussion of both approaches, see Andrich (1978).

Section 2 describes the analytic approach that is applied to the data in Section 3. Section 4 discusses implications for further research.

2. MODEL

2.1 Latent Class Rating Model

The reader will benefit from a basic knowledge of latent class analysis, such as that provided by Goodman (1974); see also McCutcheon (1987) for a less technical introduction.

Let N cases be rated by R raters on a scale with I ordered categories. The general latent class rating model assumes C case subtypes, or *latent classes.* As given by Clogg (1979), the model with $R = 3$ raters is

$$\pi_{ijk} = \sum_{c=1}^{C} \pi_c \pi_{ijk|c} \tag{1}$$

and

$$\pi_{ijk|c} = \pi_{i|c1}\pi_{j|c2}\pi_{k|c3}. \tag{2}$$

The parameters are

1. *unconditional joint probabilities,* π_{ijk} $(i, j, k = 1, \ldots, I)$, the probabilities that a randomly sampled case is assigned rating category i by rater 1, category j by rater 2, and category k by rater 3
2. *latent class prevalences,* π_c $(c = 1, \ldots, C)$, the probabilities that a randomly sampled case belongs to each latent class
3. *conditional joint probabilities,* $\pi_{ijk|c}$ $(i, j, k = 1, \ldots, I; c = 1, \ldots, C)$, the conditional probabilities that a case is assigned categories i, j, and k, by raters 1, 2, and 3, given membership in latent class c
4. *conditional rating probabilities,* $\pi_{i|c1}$, $\pi_{j|c2}$, $\pi_{k|c3}$ $(c = 1, \ldots, C; i, j, k = 1, \ldots, I)$, where, for instance, $\pi_{i|c1}$ is the probability of rater 1 assigning rating level i given a case belonging to latent class c.

The many independent basic parameters—$(C - 1)$ prevalences and $RC(I - 1)$ conditional rating probabilities—complicate estimation. Clogg suggested simplifying constraints. For example, when $I = C$, one might require $\pi_{1|1r} = \cdots = \pi_{C|Cr}$ for each r. Such constraints are arbitrary, however, and the requirement of equal numbers of latent classes and manifest categories is limiting. A more flexible and theoretically based method for parameter constraint is desirable.

2.2 Located Latent Classes

Let the latent trait that is the basis of ratings define a unidimensional continuum, and let θ denote a given latent trait level. Also, let each rater r have thresholds τ_{ir} $(i = 2, \ldots, I)$ on this continuum. We define threshold τ_{ir} as the trait level above which rater r applies category i or higher.

Each latent class is assumed to correspond to a true latent trait level β_c; latent classes are numbered such that $\beta_1 < \beta_2 < \cdots < \beta_C$. Because of measurement error, apparent trait levels for members of latent class c may vary from β_c. For rater r, measurement error is assumed normally distributed with variance σ_{cr}^2.

Let $\Phi_{cr}(\theta)$ denote the cdf of apparent trait levels of latent class c for rater r. The probability that a randomly observed member of the latent class will have an apparent trait level x_c that exceeds rater r's threshold for rating level i is

$$\Pr[x_c > \tau_{ir}] = 1 - \Phi_{cr}(\tau_{ir}). \tag{3}$$

We leave details on a *normal ogive* model based on (3) implicit. Instead, following Birnbaum (Lord and Novick 1968, pp. 399–400), we replace $\Phi_{cr}(\theta)$ with a two-parameter logistic function, $\Psi_{cr}(\theta)$, which closely approximates the former and has desirable computational properties

$$\Pr[x_c > \tau_{ir}] = 1 - \Psi_{cr}(\tau_{ir}), \tag{4}$$

where

$$\Psi_{cr}(\tau_{ir}) = \{1 + \exp[-1.7\alpha_r(\tau_{ir} - \beta_c)]\}^{-1} \tag{5}$$

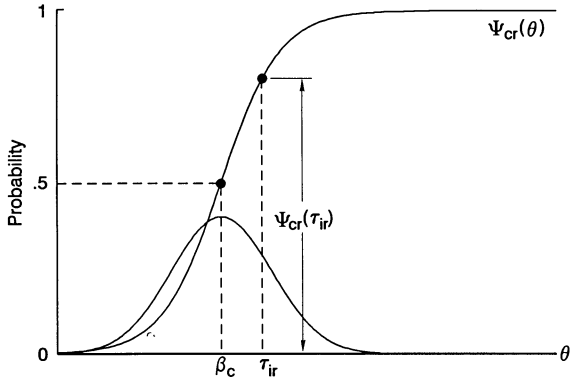

Figure 1. Located Latent Class Model. For each rater r, apparent trait levels of members of latent class c follow a normal pdf around true level β_c. Logistic function $\Psi_{cr}(\theta)$ approximates the apparent trait level cdf; 1 − $\Psi_{cr}(\tau_{ir})$ estimates the probability that a member of latent class c exceeds τ_{ir}, rater r's threshold for category i.

and α_r is a measurement error parameter for rater r. With the -1.7 constant, (5) approximates a normal cdf with variance $1/\alpha_r^2$. A special case of (5) assumes equal measurement error (EME) across raters; that is, $\alpha_1 = \cdots = \alpha_R = \alpha$. The logistic version of the located latent class model is illustrated in Figure 1.

Note that, unlike the logistic latent class models of, for example, Formann (1985, 1992) and Rost (1988), the logistic model here is viewed as an approximation to the normal ogive model, which requires slightly more computation.

Rater thresholds and logistic function parameters supply the conditional rating probabilities for Equation (2). Specifically,

$$\begin{aligned} \pi_{i|cr} &= \Psi_{cr}(\tau_{2r}) && i = 1 \\ &= \Psi_{cr}(\tau_{i+1,r}) - \Psi_{cr}(\tau_{ir}) && 1 < i < I \\ &= 1 - \Psi_{cr}(\tau_{Ir}) && i = I. \end{aligned} \tag{6}$$

We refer to (5) and (6) as the *basic model.* Identification requires two constraints on the combined set of β_c, α_r, and τ_{ir} parameters; also, only $C - 1$ prevalences need to be estimated, because $\Sigma_c \pi_c = 1$. The total number of estimated parameters is, therefore, $IR + 2C - 3$.

The parameters of the located latent class model quantify three important rater characteristics: *bias, category widths,* and *rating precision.* A rater's mean threshold provides an index of general bias—the tendency to make higher or lower ratings overall. The distance between adjacent thresholds corresponds to a category's width or definition for a rater.

Let ρ_r denote the correlation between cases' true and apparent trait levels for rater r; we term this the *latent correlation.* We define $\rho_r^2 = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_{er}^2)$, where σ_θ^2 denotes the latent trait variance. Because $1/\alpha_r^2$ approximates σ_{er}^2, we can approximate ρ_r with

$$\rho_r' = \sigma_\theta/(\sigma_\theta^2 + 1/\alpha_r^2)^{1/2}. \tag{7}$$

Because ρ_r' reflects how much apparent trait levels are determined by the latent trait rather than by measurement error, it provides an index of *rating precision.*

Measurement error is interpretable as reflecting either random noise or a unique trait that influences the ratings of a particular rater. Latent correlations assess agreement or consensus among raters as their tendency to be influenced by a common latent trait—in other words, whether raters are measuring the same thing. This is conceptually different than traditional approaches to agreement, which focus more on whether raters tend to assign a case to the same rating category; in applications where manifest agreement is the main concern, other approaches should be considered.

2.3 Some Submodels

Adding restrictions to the basic model can reduce the number of estimated parameters. Restricted models also allow statistical tests of rater differences.

We define

$$\tau_{ir} = \Delta_r + \delta_{ir}, \tag{8}$$

where Δ_r is the mean of rater r's thresholds and δ_{ir} is the deviation of threshold τ_{ir} from Δ_r (so that $\Sigma_i \delta_{ir} = 0$). One possibility is that category widths are the same across raters, but that raters differ on overall bias. To accommodate this we impose on (8) the requirement $\delta_{i1} = \cdots = \delta_{iR} = \delta_i$ for $i = 2, \ldots, I$. We term the result the *simple bias* (SB) *model.* These constraints eliminate the need to estimate $(R - 1)(I - 2)$ parameters, reducing the number of estimated parameters to $I + 2(R + C) - 5$.

Another restriction requires $\Delta_1 = \cdots = \Delta_R = \Delta$; that is, equal bias across raters. Adding this restriction to the SB model requires thresholds to be equal across raters, so we term the result the *identical thresholds* (IT) *model.*

The SB model is nested within the basic model, and the IT model is nested within the SB model. One can thus statistically assess rater differences in category widths by comparing the basic and SB models and assess rater differences in bias by comparing the SB and IT models.

The restricted models here are similar to those discussed by Muraki (1990) in the context of rating scale analysis with ordered category measures. Other models and submodels can be created with various parameter constraints. For example, one might require constant or symmetrical category widths within raters (see Tanner and Young 1985 for related discussion).

2.4 Estimation

Let f_{ijk} denote the observed frequency of cases with rating pattern (i, j, k). We assume the standard multinomial sampling model. The log-likelihood (L) of observed results is $L = \sum \sum \sum f_{ijk} \ln(\pi_{ijk})$, where π_{ijk} is obtained from Equations (1) and (2) and the located latent class model.

Maximum likelihood estimates (MLE's) are easily obtained with numerical optimization algorithms. For the analyses here a direct search routine (Chandler 1969) was used. Simplex or gradient-based algorithms probably would also be suitable. Starting values are not crucial but should be reasonable—for example, starting values for successive thresholds should be in ascending order. Several runs with different starting values are recommended to avoid local maximum solutions. A FORTRAN program that implements the approach is available from the author.

For a unique solution, the number of estimated parameters must be less than or equal to $IR - 1$. For complete assurance of local identifiability, one can evaluate the rank of the observed information matrix (-1 times the matrix of second derivatives of L with respect to model parameters), where derivatives are approximated with finite differences. Asymptotic estimated standard errors of parameter estimates are obtained as the square roots of the diagonal elements of the inverse of the observed information matrix.

Identifiability does not generally pose an obstacle to the approach's effective use. If necessary, one can usually add plausible constraints to achieve identification. Certain instances of nonidentifiability, though, are noteworthy.

With dichotomous ratings and the EME assumption, model (5) is equivalent to a Rasch model with a discrete trait distribution. Thus results of Lindsay, Clogg, and Grego (1991) apply. Specifically, $C \leq (R + 1)/2$ is required for identification of all parameters; otherwise, α and τ_{ir} parameters are identified but β_c and π_c parameters are not. Similar partial identifiability appears to exist with $I = 4$, $R = 2$, $C = 4$, and the EME assumption. The phenomenon may occur for other designs as well and needs further study.

With $R = 2$ raters in general, individual α_r terms cannot be estimated; the situation is analogous to factor analysis of two continuous measures, where both will emerge equally correlated with the common factor. Finally, α_r parameters may sometimes tend to infinity. One should therefore impose an upper limit (e.g., 10) on their value. More complex ways of handling this were discussed by Bock, Gibbons, and Muraki (1988).

2.5 Case Classification and Scoring

Parameter estimates can be used to assign each case to its most likely latent class. Recalling earlier definitions, we now define joint probability $\pi_{ijkc} = \pi_c \pi_{ijk|c}$. We also define the conditional probability of membership in latent class c given rating pattern (i, j, k) as $\pi_{c|ijk} = \pi_{ijkc}/\pi_{ijk}$. We can then assign a case to the latent class for which $\pi_{c|ijk}$ is highest, or, with the same result, to the latent class for which π_{ijkc} is highest.

In some applications it is useful to assign cases a *latent trait score.* As described by Clogg (1988), a simple method for this is to assign a case n with ratings (i, j, k) the score

$$s_n = \sum_{c=1}^{C} \hat{\pi}_{c|ijk} \hat{\beta}_c, \tag{9}$$

where $\hat{\pi}_{c|ijk}$ and $\hat{\beta}_c$ are the MLE's of $\pi_{c|ijk}$ and β_c.

3. EXPERT RATINGS ON APPROPRIATENESS OF CAROTID ENDARTERECTOMY

Brook and colleagues (Park et al. 1986) had nine medical experts rate a large number of possible indications for carotid endarterectomy on a scale of 1 for highly inappropriate to 9 for highly appropriate. We consider results for 848 indications rated by all raters.

If one is willing to view the data as interval level, then they can be analyzed with traditional methods. For example, correlations between each rater's ratings and the sum of all other raters' standardized ratings, similar to item–total correlations in classical test theory, can be calculated to estimate rater precision. An alternative would be to factor analyze ratings and estimate raters' precision by their correlations with the first common factor. With the latter approach, rater–factor correlations are closely analogous to ρ_r terms. Rater bias could also be assessed by treating ratings as repeated measures in an analysis of variance (ANOVA).

The problem is that use of integer labels for rating categories does not ensure that raters actually view them as equally spaced. Traditional methods can considerably underestimate rating precision because unequal interval widths are simply absorbed as measurement error.

Analysis of the data can be divided into three steps: (1) model choice, (2) parameter interpretation, and (3) combination of ratings to derive summary measures.

3.1 Model Choice

To facilitate analysis the ratings were collapsed to five rating levels by reassigning levels (1, 2) = 1, (3, 4) = 2, 5 = 3, (6, 7) = 4, and (8, 9) = 5. A further reason for doing this is because one rater avoided the even-numbered rating categories.

To verify the assumption of a unidimensional latent trait, the polychoric correlation matrix between pairs of raters was constructed using PRELIS (Joreskog and Sorbom 1988) and analyzed by principal components. The first eigenvalue was 7.41, and all remaining eigenvalues were less than .43. The dominant first eigenvalue supports the assumption of a unidimensional latent trait.

The extremely sparse data invalidates standard likelihood ratio chi-squared (G^2) and Pearson chi-squared (X^2) model fit tests. For initial examination of the data, the recoded ratings were therefore dichotomized by assigning levels (1, 2) = 1 and (3, 4, 5) = 2. Even with this recoding the data are somewhat sparse, but they provide a better basis for assessing model fit and potentially valid difference G^2 tests.

Several models are considered (see Table 1). Models H1 and H2 are the basic model with three and four latent classes. Models H2a and H2b add to H2 the EME and IT restrictions. Note that with dichtomous ratings the simple bias model is the same as the basic model, because there is only one threshold per rater.

H2 fits noticeably better than H1, while adding only two parameters. H2 fits the data with $X^2 = 527.58$, 488 df, and $X^2/\text{df} = 1.08$, which suggests that our basic approach is

Table 1. Results of Some Models and Submodels Applied to Dichotomized Ratings on Appropriateness of Carotid Endarterectomy

Model	Description	G^2	X^2	df
H1	3 latent classes	393.82	562.10	490
H2	4 latent classes	324.96	527.58	488
H2a	H2 + EME	420.78	767.39	496
H2b	H2 + IT*	1131.46	2588.92	496

NOTE: Constraints abbreviated as follows: EME = equal measurement error across raters; IT = identical thresholds model.

* One measurement error parameter tended to infinity and was fixed at 10; the df reported for the model view the parameter as estimated.

Table 2. Results of Some Models and Submodels Applied to Five-Level Ratings on Appropriateness of Carotid Endarterectomy

Model	Description	Number of parameters	G^2	AIC	Schwarz index
M1	5 latent classes	52	4082.44	11652	11898
M1a	M1 + ELC	49	4129.08	11692	11925
M1b	M1 + ELC + SB	25	4484.69	12000	12119
M1c	M1 + ELC + SB + EME	17	4711.31	12211	12291
M2	7 latent classes + ELC	51	4052.98	11620	11862

NOTE: Constraints abbreviated as follows: ELC = equally spaced latent classes, SB = simple bias model, EME = equal measurement error across raters.

reasonable. Results of other models show that adding latent classes beyond four improves fit only slightly.

The difference G^2 for the H2a–H2 comparison is 420.78 − 324.96 = 95.81 with 496 − 488 = 8 df. Therefore, rater differences in measurement error appear statistically significant. The H2b–H2 comparison tests whether raters have equal bias. The difference G^2 for this comparison is 806.50 with 8 df, indicating clear bias differences.

Table 2 shows the results of several models applied to the five-level ratings. One unrestricted model (M1) and several restricted models (M1a–M1c, M2) are shown. The EME and SB constraints have already been discussed. The ELC restriction assumes *equally spaced latent classes;* that is, a constant distance between successive β_c values. This simplification reduces the number of estimated parameters.

The Akaike information criterion (AIC) and Schwarz index, model selection criteria, are potentially helpful with sparse data (Sclove 1987). The AIC is calculated as $-2L + 2p$, where L is the log-likelihood and p is the number of estimated parameters. The Schwarz index is calculated as $-2L + \ln(N)p$, where N is the sample size. For both, smaller values indicate more preferred models.

M2 is the statistically preferred model of Table 2 and many others tested, containing from three to ten latent classes. But many of the results of M2 are closely approximated with the simpler models, so we consider some of these also.

3.2 Parameter Interpretation

Latent Trait Distribution. For these data, latent classes could be viewed either as reflecting underlying conceptual categories that correspond to various levels of treatment appropriateness or as providing a discrete approximation to a continuous trait of appropriateness. The ELC assumption is more defensible if one adopts the second view. Table 3 shows MLE's of β_c and π_c parameters for M1a. Note that the β_c parameters are fixed to satisfy the ELC assumption and supply the constraints for identification.

A bimodal latent trait distribution, very clear with M2 and also seen with M1a in Table 3, suggests that indications may be a mixture of two types: a set of clearly inappropriate indications and a set of indications with varying appropriateness. This has practical implications for setting treatment guidelines with expert panels, because it suggests that many indications can be unequivocally determined inappropriate. A corresponding group of clearly appropriate indications, however, is not apparent.

The bimodal distribution is also seen with M1b and M1c. With M1, π_c values are not bimodal, but β_1 is markedly separated from β_2–β_5. One would, therefore, draw the same conclusion of a distinct group of clear nonindications.

Latent Correlations. Table 4 shows α_r and ρ'_r values for M1. Nearly identical estimates are obtained with M2. The ρ'_r values are reasonably high overall, so that raters appear to be measuring a common latent trait. Some differences, though, are apparent, with Raters 4, 8, and 9 having lower values. As noted in connection with Table 1, rater differences in ρ'_r appear significant from the H2a–H2 comparison.

Latent correlations are slightly lower for Models M1a–M1c. This is presumably because departure of the data from the more rigid assumptions of simpler models increases estimates of measurement error.

Rater Bias. Bias estimates are little affected by model choice; for all pairs of models in Table 2, bias estimates of

Table 3. Latent Class Locations and Prevalence Estimates for Model M1a

Latent class (c)	Location parameter (β_c)*	Prevalence (π_c)
1	−3.00	.408 (.025)
2	−1.50	.188 (.022)
3	.00	.217 (.017)
4	1.50	.114 (.013)
5	3.00	.073 (.010)

NOTE: Standard errors in parentheses.
* Parameters fixed.

Table 4. Measurement Error Parameters (α_r) and Latent Correlations (ρ'_r) for Model M1

Rater (r)	α_r	ρ'_r
1	1.272 (.127)	.93
2	1.180 (.122)	.92
3	1.159 (.094)	.92
4	.693 (.053)	.81
5	1.052 (.086)	.90
6	1.633 (.145)	.96
7	1.230 (.116)	.93
8	.715 (.049)	.82
9	.707 (.057)	.82
Mean	1.071	.89

NOTE: Standard errors in parentheses.

raters correlate .99 or above. Table 5 shows Δ_r estimates and their standard errors for Model M1b. Model H2 produces similar bias estimates, and, as already noted, the H2b–H2 comparison indicates statistically significant bias differences overall.

The dichotomized ratings also help address the question of which *pairs* of raters differ significantly. Follow-up analyses examined this, using H2 as the baseline model and then requiring Δ_r to be equal for pairs of raters. Bias differences for all but 7 of the 36 rater pairs were significant at the $p \leq .05/36$ level; the nonsignificant differences were for pairs of raters in the set $\{4, 6, 8, 9\}$ and for Raters 1 and 5.

Category Widths. Threshold estimates for M1, M1a, and M2 are very similar. Rater differences in category widths are evident in Figure 2, which shows the estimates for M2.

A comparison of Raters 8 and 9 is illustrative. For Rater 9 the middle categories, especially category 2, are narrower and categories 1 and 5 are wider; thus Rater 9's judgments appear more polarized. Graphical feedback similar to Figure 2 can help raters avoid overly wide or overly narrow category definitions.

3.3 Combining Ratings

With the method described in Section 2.5, indications can be assigned to the most probable latent class. For M1 the proportions of indications assigned to latent classes 1–5 are .41, .21, .19, .11, and .08.

Alternatively, (9) can be used to measure each indication's appropriateness. Resulting scores are very consistent across models; for any pair of models in Table 2, they correlate .99 or above.

A related question is whether some indications are *unscalable* (Dayton and Macready 1980; Goodman 1975)—that is, whether they obtain a markedly inconsistent pattern of ratings. Such cases may require different treatment. For example, raters might discuss them to identify reasons for disagreement.

One way to examine this is to add a latent class $c = 0$, where $\pi_{i|0r} = 1/I$ for $i = 1, \ldots, I$ and $r = 1, \ldots, R$. When this is done with M1a, for example, the estimated prevalence of the unscalable latent class is 0; therefore, no indications are assigned to this class. But one might modify the procedure to assign an indication to the latent class for which $\pi_{ijk|c}$ is highest. When this is done, five indications are assigned to the unscalable class.

Table 5. Rater Bias Estimates for Model M1b (Simple Bias Model)

Rater (r)	Δ_r
1	−.422 (.064)
2	2.577 (.114)
3	.707 (.050)
4	.452 (.073)
5	−.235 (.066)
6	.449 (.050)
7	1.833 (.081)
8	.541 (.062)
9	.398 (.092)

NOTE: Standard errors in parentheses.

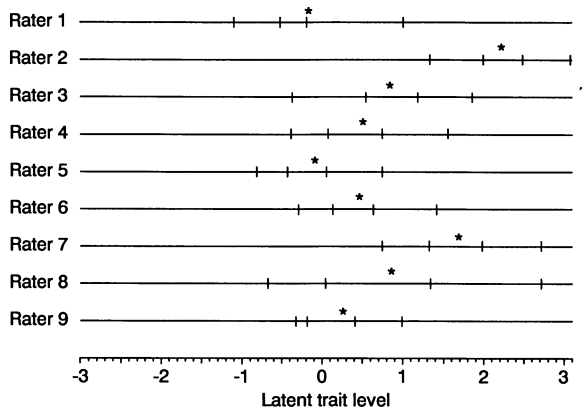

Figure 2. Estimated Threshold Locations for Model M2 of Table 2. The vertical lines show successive rating category thresholds for each rater. The asterisks show mean thresholds.

4. DISCUSSION

It is of some interest that restricted models with fewer parameters can accurately reproduce many results of the basic model. Models with the ELC and SB restrictions appear more than adequate to estimate latent trait scores and characterize tendencies of some raters to make higher or lower ratings, but possibly not to accurately estimate latent correlations.

Several extensions of the present approach appear possible. One would be to consider multiple latent trait dimensions that raters weight differently. Multidimensional latent trait models described by Bock and Aitkin (1981) and Bock et al. (1988) potentially could be adapted for this.

We have assumed complete data. In some applications, however, each rater is assigned only a subset of cases to rate. The approach here can be easily adapted to handle this situation. This involves a slight modification of the likelihood equation so that it considers only nonmissing ratings.

Panel rating studies sometimes use a large number of raters. In such cases it may be useful to model the distribution of certain parameters (e.g., Δ with the simple bias model or α) rather than estimate values for each rater.

The statistical model used here has several possible applications beyond the analysis of expert ratings, such as item analysis and survey data analysis.

[*Received November 1990. Revised August 1992.*]

REFERENCES

Agresti, A. (1988), "A Model for Agreement Between Ratings on an Ordinal Scale," *Biometrics,* 44, 539–548.

——— (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, Vol. 49.

Andrich, D. (1978), "A Rating Formulation for Ordered Response Categories," *Psychometrika,* 43, 561–573.

Becker, M. P. (1989), "Using Association Models to Analyse Agreement Data: Two Examples," *Statistics in Medicine,* 8, 1199–1207.

——— (1990), "Quasi-symmetric Models for the Analysis of Square Contingency Tables," *Journal of the Royal Statistical Society,* Ser. B, 52, 369–378.

Bock, R. D., and Aitkin, M. (1981), "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," *Psychometrika,* 46, 443–459.

Bock, R. D., Gibbons, R., and Muraki, E. (1988), "Full-Information Item Factor Analysis," *Applied Psychological Measurement,* 12, 261–280.

Chandler, J. P. (1969), "STEPIT—Finds Local Minima of a Smooth Function of Several Parameters," *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.

——— (1988), "Latent Class Models for Measuring," in *Latent Trait and Latent Class Models,* eds. R. Langeheine and J. Rost, New York: Plenum, pp. 173–205.

Cohen, J. (1968), "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin,* 70, 213–220.

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.

Dayton, C. M., and Macready, G. B. (1980), "A Scaling Model With Response Errors and Intrinsically Unscalable Respondents," *Psychometrika,* 45, 343–356.

——— (1988), "Concomitant-Variable Latent Class Models," *Journal of the American Statistical Association,* 83, 173–178.

Dillon, W. R., and Mulani, N. (1984), "A Probabilistic Latent Class Model for Assessing Inter-Judge Reliability," *Multivariate Behavioral Research,* 19, 438–458.

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.

Formann, A K. (1985), "Constrained Latent Class Models: Theory and Applications," *British Journal of Mathematical and Statistical Psychology,* 38, 87–111.

——— (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.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika,* 61, 215–231.

——— (1975), "A New Model for Scaling Response Patterns: An Application of the Quasi-Independence Concept," *Journal of the American Statistical Association,* 70, 755–768.

——— (1986), "Some Useful Extensions of the Usual Log-Linear Models Approach in the Analysis of Contingency Tables" (with discussion), *International Statistical Review,* 54, 243–309.

Haberman, S. J. (1979), *Qualitative Data Analysis: Vol. 2, New Developments,* New York: Academic Press.

Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990), "Receiver Operator Characteristic (ROC) Analysis Without Truth," *Medical Decision Making,* 10, 24–29.

Joreskog, K. G., and Sorbom, D. (1988), *PRELIS User's Manual,* Chicago: Scientific Software, Inc.

Kelderman, H., and Macready, G. B. (1990), "The Use of Loglinear Models for Assessing Differential Item Functioning Across Manifest and Latent Examinee Groups," *Journal of Educational Measurement,* 27, 307–327.

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, MA: Addison-Wesley.

Mislevy, R. J., and Verhelst, N. (1990), "Modeling Item Responses When Different Subjects Employ Different Strategies," *Psychometrika,* 55, 195–215.

Muraki, E. (1990), "Fitting a Polytomous Item Response Model to Likert-Type Data," *Applied Psychological Measurement,* 14, 59–71.

McCutcheon, A. C. (1987), *Latent Class Analysis,* Beverly Hills, CA: Sage Publications.

Park, R. E., Fink, A., Brook, R. H., Chassin, M. R., Kahn, K. L., Merrick, N. J., Kosecoff, J., and Solomon, D. H. (1986), "Physician Ratings of Appropriate Indications for Six Medical and Surgical Procedures," *American Journal of Public Health,* 76, 766–772.

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 ed.), Chicago: University of Chicago Press.

Rost, J. (1988), "Rating Scale Analysis with Latent Class Models," *Psychometrika,* 53, 327–348.

Samejima, F. (1969), "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychometric Monograph No. 17.*

Sclove, S. (1987), "Application of Model-Selection Criteria to Some Problems in Multivariate Analysis," *Psychometrika,* 52, 333–343.

Tanner, M. A., and Young, M. A. (1985), "Modeling Ordinal Scale Disagreement," *Psychological Bulletin,* 98, 408–415.

Uebersax, J. S. (1988), "Validity Inferences from Interobserver Agreement," *Psychological Bulletin,* 104, 405–416.

——— (1991), "Quantitative Methods for the Analysis of Observer Agreement: Toward a Unifying Model," Paper P-7686, Santa Monica, CA: The RAND Corporation.

——— (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," Note N-3029-RC, Santa Monica, CA: The RAND Corporation.

——— (1990), "Latent Class Analysis of Diagnostic Agreement," *Statistics in Medicine,* 9, 559–572.

——— (1993), "A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement," *Biometrics,* Vol. 49.

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.