Biometrics 62, 261–268
March 2006

DOI: 10.1111/j.1541-0420.2005.00408.x

A Mixed-Effects Regression Model for Longitudinal Multivariate Ordinal Data

Li C. Liu* **and Donald Hedeker**

Institute for Health Research and Policy, University of Illinois at Chicago,
1747 W. Roosevelt Road, Room 558, M/C 275, Chicago, Illinois 60608, U.S.A.
**email:* lqi1@uic.edu

Summary. A mixed-effects item response theory model that allows for three-level multivariate ordinal outcomes and accommodates multiple random subject effects is proposed for analysis of multivariate ordinal outcomes in longitudinal studies. This model allows for the estimation of different item factor loadings (item discrimination parameters) for the multiple outcomes. The covariates in the model do not have to follow the proportional odds assumption and can be at any level. Assuming either a probit or logistic response function, maximum marginal likelihood estimation is proposed utilizing multidimensional Gauss–Hermite quadrature for integration of the random effects. An iterative Fisher scoring solution, which provides standard errors for all model parameters, is used. An analysis of a longitudinal substance use data set, where four items of substance use behavior (cigarette use, alcohol use, marijuana use, and getting drunk or high) are repeatedly measured over time, is used to illustrate application of the proposed model.

Key words: Item response theory models; Maximum marginal likelihood; Multilevel data; Multiple outcomes; Nonproportional odds assumption; Ordinal data; Random-effects models; Two-parameter model.

1. Introduction

Ordinal responses are common in many areas of research. For instance, smoking status might be measured using categories "nonsmoker," "light smoker," and "heavy smoker." Additionally, many kinds of data have a hierarchical or clustered structure. For example, subjects can be observed nested within clusters (i.e., families, schools), or repeatedly measured across time, and so the use of ordinal regression models that ignore the correlation between subjects within a cluster or the correlation between the repeated measurements of a subject is problematic. For multilevel ordinal data (either clustered or longitudinal), mixed-effects regression models have become increasingly popular (Hedeker and Gibbons, 1994). These developments have been mainly in terms of logistic and probit regression models. In particular, because the logistic-based proportional odds model described by McCullagh (1980) is a common choice for analysis of ordinal data, many of the multilevel models for ordinal data are generalizations of this model.

The psychometrics and educational testing literatures report an enormous body of research relating to mixed-effects models for subject-specific comparisons of multiple categorical responses, referred to as item response theory (IRT) models (Lord, 1980). Here, item responses are nested within subjects, forming a two-level nesting structure. Samejima (1969), Masters (1982), and Tutz (1990) discuss more general ordinal IRT models. These models typically do not include covariates, other than terms for the specific ordinal items.

In some situations, the clustered data structure could have three levels of nesting. For instance, in a longitudinal study with multivariate outcomes, multiple item responses (level 1) are nested within measurement occasions (level 2), which are nested within subjects (level 3). For three-level discrete data, methods have been developed primarily for univariate binary responses (Gibbons and Hedeker, 1997; Ten Have, Kunselman, and Tran, 1999). Also, in the IRT field, little work has been done on generalizing models for multivariate data with three-level nesting structures. Fox and Glas (2001) introduce an IRT model for binary outcomes in a strictly clustered setting (i.e., items are nested within students and students are nested within schools). For three-level longitudinal multivariate data where multiple random effects are needed at the individual level, methods have been developed for continuous outcomes (Roy and Lin, 2000, 2002).

In this article, we extend the three-level mixed-effects model for dichotomous clustered data of Gibbons and Hedeker (1997) to allow for multiple ordinal outcomes in longitudinal settings. More specifically, at the multiple outcomes level, a separate random-effects standard deviation, or item discrimination parameter, is included for each ordinal outcome. These are akin to factor loadings in a factor analysis model. At the subject level, multiple random subject effects are included to account for the longitudinal design. This model is developed for both the probit and logistic response functions. To motivate the model, we use the threshold concept, in which it is assumed that a latent unobservable continuous response underlies each of the observed ordinal outcomes, and the latent continuous response follows a linear regression model incorporating random effects. A maximum marginal likelihood (MML) solution, via Fisher scoring, is described using

multidimensional Gauss–Hermite quadrature to numerically integrate over the distribution of random effects. In modeling the cumulative logits for the observed ordinal responses, both a proportional odds and a nonproportional odds model are described. Both models are applied to a data set from the Aban Aya prevention study, where multiple ordinal items of substance use behavior are repeatedly measured over time in a sample of inner-city youth.

2. Multilevel Representation of IRT Model

To prepare for the three-level IRT model for ordinal responses, we first consider a basic IRT model for binary outcomes. Suppose that there are N subjects responding to m questions (items). In multilevel terms, this corresponds to a two-level data structure with repeated observations at level 1 nested within subjects at level 2. The Rasch model (Rasch, 1961) specifies that the probability of correct response by subject $i\,(i = 1, 2, \ldots, N)$ on item $j\,(j = 1, 2, \ldots, m)$ depends on a subject's ability θ_i and item difficulties α_j through the equation

$$p_{ij} = \Pr(Y_{ij} = 1 \,|\, \theta_i) = \frac{1}{1 + \exp[-(\theta_i - \alpha_j)]}\,, \qquad (1)$$

which can be recast in terms of the logit of the response as $\text{logit}(p_{ij}) = \theta_i - \alpha_j$. The distribution of abilities θ_i in the population of subjects is typically assumed to be standard normal. An extended logistic model for dichotomous responses is the two-parameter model described in Lord (1980), which specifies that the conditional probability of a correct response is given by

$$\Pr(Y_{ij} = 1 \,|\, \theta_i) = \frac{1}{1 + \exp[-\tau_j(\theta_i - \alpha_j)]}\,, \qquad (2)$$

where τ_j is the slope parameter for item j (i.e., item discrimination), and α_j is again the difficulty parameter for item j (i.e., item difficulty). As noted by Bock and Aitkin (1981), it is convenient to let $\beta_j = -\tau_j\alpha_j$ and write

$$\Pr(Y_{ij} = 1 \,|\, \theta_i) = \frac{1}{1 + \exp[-(\tau_j\theta_i + \beta_j)]}\,, \qquad (3)$$

or, in the logit form $\text{logit}(p_{ij}) = \tau_j\theta_i + \beta_j$. Comparing equations (1) and (3), it is clear that the Rasch model is essentially a restricted version of the two-parameter model where all item discriminations (τ_j's) are equal. For simplicity, in what follows, the two-parameter model will be referred to as the "IRT model" while the one-parameter Rasch model will simply be termed the "Rasch model."

The IRT model can be written in a matrix representation by specifying $\boldsymbol{D}_i$ as an $m \times m$ matrix of dummy codes indicating the items. For instance, suppose that there are four items, then

$$\underset{\boldsymbol{\Lambda}_i}{\begin{bmatrix}\text{logit}_{i1}\\ \text{logit}_{i2}\\ \text{logit}_{i3}\\ \text{logit}_{i4}\end{bmatrix}} = \underset{\boldsymbol{X}_i}{\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}} \underset{\boldsymbol{\beta}}{\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{bmatrix}} + \underset{\boldsymbol{D}_i}{\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}} \underset{\boldsymbol{T}}{\begin{bmatrix}\tau_1\\ \tau_2\\ \tau_3\\ \tau_4\end{bmatrix}} [\theta_i]\,. \qquad (4)$$

This IRT model is simply a random intercept model that allows the random-effects variance terms to vary across items, whereas the Rasch model is an ordinary random intercept model. Notice that $\boldsymbol{\beta}$ is simply the fixed-effects parameter vector and $\boldsymbol{T}$ is the random-effects standard deviation vector that can also be viewed as the factor loadings of the items on the unidimensional underlying subject ability θ_i. Goldstein (1995) refers to this type of two-level model as a multilevel model with complex variance structure.

Unlike traditional IRT models, the multilevel model formulation easily allows multiple covariates at either level (i.e., items or subjects). This and other advantages of casting IRT models as multilevel models are described by Adams, Wilson, and Wu (1997). More recently, Rijmen et al. (2003) provide a comprehensive overview and bridge between IRT models, multilevel models, mixed models, and generalized linear mixed models (GLMMs). As they point out, the Rasch model, and variants of it, belong to the class of GLMMs. However, the more extended two-parameter model is not within the class of GLMMs because the predictor is no longer linear, but includes a product of parameters.

3. The Three-Level IRT Model for Ordinal Responses: Proportional Odds Model

Regression models for discrete responses are often motivated and described using the "threshold concept" (Bock, 1975). For this, it is assumed that a continuous latent variable y underlies the observed ordinal response Y, and the value of Y is determined by a series of increasing thresholds $\gamma_1 < \gamma_2 < \cdots < \gamma_{C-1}$, with $\gamma_0 = -\infty$, $\gamma_1 = 0$, and $\gamma_C = \infty$ (C as the number of ordered categories).

Suppose, in a longitudinal study with multiple ordinal outcomes, m item responses are repeatedly measured across time for N subjects. Let $i\,(i = 1, 2, \ldots, N)$ denote the level-3 subject, let $j\,(j = 1, 2, \ldots, n_i)$ denote the level-2 time point, or occasion, for subject i, and let $k\,(k = 1, 2, \ldots, n_{ij};\ n_{ij} \le m)$ denote the level-1 item on occasion j for subject i. The mixed-effects regression model for the underlying latent variable y_{ijk} can be written as

$$y_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{w}'_{ijk}\boldsymbol{\delta}_i + \boldsymbol{d}'_{ijk}\boldsymbol{T}_{(2)}\theta_{ij} + \epsilon_{ijk}, \qquad (5)$$

where $\boldsymbol{x}_{ijk}$ is the $p \times 1$ covariate vector, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed regression coefficients, $\boldsymbol{w}_{ijk}$ is the design vector for the r random subject effects (level 3), and $\boldsymbol{\delta}_i$ is the $r \times 1$ vector of level-3 random subject effects (e.g., random intercept and slope), which follow a multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma}_{(3)})$. Similar to the two-parameter IRT model in equation (4), $\boldsymbol{d}_{ijk}$ is the $m \times 1$ indicator vector for the repeated items, and $\boldsymbol{T}_{(2)}$ is the random-effects standard deviation vector for the level-2 subject ability θ_{ij}, which follows a standard normal distribution. The residuals ϵ_{ijk} are assumed to follow a normal or logistic distribution with mean 0 and variance σ^2_ϵ, for either a probit or logistic regression. Since the level-3 subscript i, the level-2 subscript j, and the level-1 subscript k are present for the $\boldsymbol{x}$, $\boldsymbol{w}$, and $\boldsymbol{d}$ vectors, not all level-3 units (subjects) are assumed to have the same number of level-2 units (occasions), and not all level-2 units are assumed to have the same number of level-1 units (items). Though at a particular time point, subjects may have the same number of item responses, it is not the requirement of the model.

It is computationally convenient to represent the random subject effects in standardized form (Gibbons and Bock, 1987). Specifically, let $\boldsymbol{\delta}_i = \boldsymbol{T}_{(3)}\boldsymbol{\theta}_i$, where $\boldsymbol{T}_{(3)}\boldsymbol{T}'_{(3)} = \boldsymbol{\Sigma}_{(3)}$ is

the Cholesky decomposition of the $r \times r$ matrix $\boldsymbol{\Sigma}_{(3)}$ and $\boldsymbol{\theta}_i$ is the vector of standardized level-3 random effects. The reparameterized three-level model is then

$$y_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{w}'_{ijk}\boldsymbol{T}_{(3)}\boldsymbol{\theta}_i + \boldsymbol{d}'_{ijk}\boldsymbol{T}_{(2)}\theta_{ij} + \epsilon_{ijk}. \quad (6)$$

This transformation allows us to estimate the Cholesky factor $\boldsymbol{T}_{(3)}$, which is a lower-triangular matrix, instead of the covariance matrix $\boldsymbol{\Sigma}_{(3)}$. As the Cholesky factor is essentially the square root of the covariance matrix, this then allows more stable estimation of near-zero variance terms.

In models (5) and (6), the regression coefficients $\boldsymbol{\beta}$ do not carry the c subscript, that is, they do not vary across categories. Thus, the relationship between the explanatory variables and the cumulative logits does not depend on c. This assumption of identical odds ratios across the $C-1$ cut-offs is the *proportional odds* assumption of McCullagh (1980).

3.1 *Probit and Logistic Response Functions*

With the mixed-effects regression model for the underlying latent variable y_{ijk}, the probability that $Y_{ijk} = c$, conditional on the random effects, under the probit formulation is given by

$$\Pr(Y_{ijk} = c \,|\, \boldsymbol{\theta}^*) = \Phi[(\gamma_c - z_{ijk})/\sigma_\epsilon] - \Phi[(\gamma_{c-1} - z_{ijk})/\sigma_\epsilon] \,, \quad (7)$$

where $\boldsymbol{\theta}^*$ represents all the random effects in the model ($\boldsymbol{\theta}_i$ and θ_{ij}), $z_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{w}'_{ijk}\boldsymbol{T}_{(3)}\boldsymbol{\theta}_i + \boldsymbol{d}'_{ijk}\boldsymbol{T}_{(2)}\theta_{ij}$, and $\Phi(\cdot)$ represents the standard normal cumulative distribution function (cdf). Without loss of generality, the origin and unit of z may be chosen arbitrarily. For convenience, let $\gamma_1 = 0$ and $\sigma^2_\epsilon = 1$ (for the probit response function). Alternatively, if the logistic response function is assumed, then the logistic cdf $\Psi(\cdot)$ replaces the normal, and the standard variance of the logistic is $\sigma^2_\epsilon = \pi^2/3$. In the development that follows, we will assume the probit response function; however, we will note the necessary modifications if the logistic function is used instead.

3.2 *Maximum Marginal Likelihood Estimation*

Let $\boldsymbol{Y}_i$ be the vector pattern of the ordinal responses from subject i for all of the n_i occasions with n_{ij} items at each occasion. Assuming independence of the responses conditional on the random effects, the conditional likelihood of any pattern $\boldsymbol{Y}_i$, given $\boldsymbol{\theta}^*$, is given by

$$L\big(\boldsymbol{Y}_i \,\big|\, \boldsymbol{\theta}^*\big) = \prod_{j=1}^{n_i}\prod_{k=1}^{n_{ij}}\prod_{c=1}^{C}[\Phi(\gamma_c - z_{ijk}) - \Phi(\gamma_{c-1} - z_{ijk})]^{d_{ijkc}}, \quad (8)$$

where

$$d_{ijkc} = \begin{cases} 1 & \text{if} \quad Y_{ijk} = c, \\ 0 & \text{if} \quad Y_{ijk} \neq c. \end{cases}$$

Then the marginal likelihood of $\boldsymbol{Y}_i$ in the population is expressed as the following integral of the conditional likelihood, $L(\cdot)$, weighted by the prior density $g(\cdot)$

$$h(\boldsymbol{Y}_i) = \int_{\boldsymbol{\theta}^*} L\big(\boldsymbol{Y}_i \,\big|\, \boldsymbol{\theta}^*\big) g(\boldsymbol{\theta}^*)\, d\boldsymbol{\theta}^*, \quad (9)$$

where $g(\boldsymbol{\theta}^*)$ represents the distribution of the random effects $\boldsymbol{\theta}^*$ in the population (the joint distribution of $\boldsymbol{\theta}_i$, a multivariate standard normal and θ_{ij}, a standard normal density).

With the assumption that, conditional on the level-3 effect $\boldsymbol{\theta}_i$, the responses from the n_i occasions in subject i are independent, the marginal probability can be rewritten as

$$h(\boldsymbol{Y}_i) = \int_{\boldsymbol{\theta}_i} \left\{ \prod_{j=1}^{n_i} \int_{\theta_{ij}} L_{ij}(\boldsymbol{\theta}^*) g(\theta_{ij})\, d\theta_{ij} \right\} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i, \quad (10)$$

where

$$L_{ij}(\boldsymbol{\theta}^*) = \prod_{k=1}^{n_{ij}}\prod_{c=1}^{C}[\Phi(\gamma_c - z_{ijk}) - \Phi(\gamma_{c-1} - z_{ijk})]^{d_{ijkc}}. \quad (11)$$

Here, $\boldsymbol{\theta}_i$ has dimension r and θ_{ij} has dimension 1. Therefore, the integration is of dimensionality $r+1$ and is tractable as long as the number of level-3 random effects is no greater than three or four, a condition which is typically satisfied for longitudinal studies.

3.3 *Estimation*

For estimation of the p covariate coefficients $\boldsymbol{\beta}$, $r(r+1)/2$ Cholesky elements of the level-3 variance–covariance structure $\boldsymbol{T}_{(3)}$, m item discrimination parameters $\boldsymbol{T}_{(2)}$, and $C-2$ threshold values $\gamma_c\,(c = 2, \ldots, C-1)$, we differentiate the marginal log likelihood for the patterns from the N level-3 subjects,

$$\log L = \sum_{i}^{N} \log h(\boldsymbol{Y}_i) \,. \quad (12)$$

Let $\boldsymbol{\eta}$ be an arbitrary parameter vector, then we obtain

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}'} = \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i) \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}'} \,. \quad (13)$$

For the threshold values, say γ_l, we obtain

$$\frac{\partial h(\boldsymbol{Y}_i)}{\partial \gamma_l} = \int_{\boldsymbol{\theta}_i} E_i \left\{ \sum_{j=1}^{n_i} e_{ij}^{-1} \int_{\theta_{ij}} \sum_{k}^{n_{ij}} \sum_{c}^{C} d_{ijkc} \times \frac{\phi(\gamma_c - z_{ijk})\delta_{c,l} - \phi(\gamma_{c-1} - z_{ijk})\delta_{c-1,l}}{\Phi(\gamma_c - z_{ijk}) - \Phi(\gamma_{c-1} - z_{ijk})} \times L_{ij}(\boldsymbol{\theta}^*) g(\theta_{ij})\, d\theta_{ij} \right\} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i, \quad (14)$$

where

$$e_{ij} = \int_{\theta_{ij}} L_{ij}(\boldsymbol{\theta}^*) g(\theta_{ij})\, d\theta_{ij} \quad E_i = \prod_{j=1}^{n_i} e_{ij} \quad \delta_{c,l} = \begin{cases} 1 & \text{if} \quad c = l, \\ 0 & \text{if} \quad c \neq l. \end{cases}$$

For the estimates of the other parameters in the model, let $\boldsymbol{\eta}^*$ represent the parameter vector containing $\boldsymbol{\beta}$, $v(\boldsymbol{T}_{(3)})$ (unique elements of the Cholesky factor $\boldsymbol{T}_{(3)}$), and $\boldsymbol{T}_{(2)}$ ($\boldsymbol{T}_{(2)}$ is already a vector). We obtain

$$\frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\eta}^{*\prime}} = \int_{\boldsymbol{\theta}_i} E_i \left\{ \sum_{j=1}^{n_i} e_{ij}^{-1} \int_{\theta_{ij}} \sum_{k}^{n_{ij}} \sum_{c}^{C} d_{ijkc} \times \frac{\phi(\gamma_{c-1} - z_{ijk}) - \phi(\gamma_c - z_{ijk})}{\Phi(\gamma_c - z_{ijk}) - \Phi(\gamma_{c-1} - z_{ijk})} \times L_{ij}(\boldsymbol{\theta}^*) g(\theta_{ij}) \frac{\partial z_{ijk}}{\partial \boldsymbol{\eta}^{*\prime}}\, d\theta_{ij} \right\} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i, \quad (15)$$

where

$$\frac{\partial z_{ijk}}{\partial \boldsymbol{\beta}'} = \boldsymbol{x}'_{ijk} \quad \frac{\partial z_{ijk}}{\partial (v(\boldsymbol{T}_{(3)}))'} = (\boldsymbol{\theta}'_i \otimes \boldsymbol{w}'_{ijk})\boldsymbol{J}'_r \quad \frac{\partial z_{ijk}}{\partial (\boldsymbol{T}_{(2)})'} = \theta_{ij}.$$

In the above formula, $\otimes$ is the Kronecker product, and $\boldsymbol{J}_r$ is the transformation matrix of Magnus (1988) that eliminates the elements above the main diagonal. Note that, for the logistic regression formulation, the logistic function $\Psi(\cdot)$ replaces the normal response function $\Phi(\cdot)$, and the product $\Psi(\cdot) \times (1 - \Psi(\cdot))$ replaces the standard normal density function $\phi(\cdot)$ in the equations.

Fisher's method of scoring can be used to provide the solution to these likelihood equations. For this, provisional estimates for the vector of parameters $\boldsymbol{\Theta}$ on iteration ι are improved by

$$\boldsymbol{\Theta}_{\iota+1} = \boldsymbol{\Theta}_\iota - E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_\iota \partial \boldsymbol{\Theta}'_\iota}\right]^{-1} \frac{\partial \log L}{\partial \boldsymbol{\Theta}_\iota}, \tag{16}$$

where the observed information matrix, or expectation of the matrix of second derivatives, is given by

$$E\left[\frac{\partial^2 \log L}{\partial \boldsymbol{\Theta}_\iota \partial \boldsymbol{\Theta}'_\iota}\right] = -\sum_{i=1}^{N} h^{-2}(\boldsymbol{Y}_i) \frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\Theta}_\iota} \left(\frac{\partial h(\boldsymbol{Y}_i)}{\partial \boldsymbol{\Theta}_\iota}\right)'. \tag{17}$$

At convergence, the large-sample variance–covariance matrix of the parameter estimates is then obtained as the inverse of the information matrix.

3.4 *Computer Implementation*

To solve the above likelihood equations, numerical integration on the transformed $\boldsymbol{\theta}^*$ space may be performed. For this, Gauss–Hermite quadrature is used to approximate the integrals by summations on Q quadrature points for each dimension of the integration. The summation first goes over Q points for the level-2 effect θ_{ij}, then over Q^r points for the r-dimension level-3 effects $\boldsymbol{\theta}_i$. For the standard normal univariate density, optimal points and weights are given in Stroud and Sechrest (1966). For the results in this article, 10 quadrature points and corresponding weights were used for each dimension. At each iteration and for each level-3 subject, the solution first goes over 10 points for the level-2 effect θ_{ij}, substituting the random effect θ_{ij} by the current quadrature point and the evaluation of the standard normal density $g(\theta_{ij})$ by the current weight. Then it goes over 10^r points for the r-dimension level-3 effects $\boldsymbol{\theta}_i$, substituting the vector $\boldsymbol{\theta}_i$ and multivariate density $g(\boldsymbol{\theta}_i)$ by the current vector of points and weights, respectively. Following the summation over the quadrature points and level-3 subjects, parameters are corrected using the relative correction $(\boldsymbol{\Theta}_{\iota+1} - \boldsymbol{\Theta}_\iota)/\boldsymbol{\Theta}_\iota$. The entire procedure is repeated until convergence, which is specified when relative corrections of less than 0.0001 are obtained for all parameters. This procedure was implemented using the GAUSS language (GAUSS 3.6, 2001).

4. Application

The Aban Aya Youth Project (AAYP) was a longitudinal efficacy trial designed to compare the effects of three multiyear interventions on reducing health-compromising behaviors such as substance use, violence, and unsafe sex. The three interventions, namely, School-Community Program (SC), Social Development Curriculum (SDC), and Health Enhancement Curriculum (HEC), were implemented in grades 5 through 8 in 12 poor, African American metropolitan schools in Chicago and the surrounding suburbs. A randomized block design was used to assign schools to the three interventions after stratification on multiple indicators of risk. The AAYP participants were students in fifth grade classes in the 12 schools (9 inner-city and 3 near-suburban) during the 1994–1995 school year or those who transferred in during the study. Students who transferred out were not followed. Self-report data were collected from the participating students at the beginning of fifth grade (pretest in October of grade 5, $T_1 = 0$, $N = 668$), and posttests in April or May at the end of grades 5 ($T_2 = 0.5$ year, $N = 634$), 6 ($T_3 = 1.5$ years, $N = 674$), 7 ($T_4 = 2.5$ years, $N = 597$), and 8. At the eighth grade, data collection occurred at one of the three time points, $T_5 = 3.0$, 3.25, or 3.5 years, with total $N = 645$. The five waves of follow-ups resulted in a total sample of 1153 African American students, of whom 571 were male students. For details about the design and procedures of AAYP, see Flay et al. (2004).

To illustrate application of the proposed mixed-effects IRT model for longitudinal multivariate ordinal data, substance use behavior, which was represented by four ordinal items (cigarette use, alcohol use, marijuana use, and getting drunk or high) and was measured in all five waves, is used. All four items were rated by students on a 4-point ordinal scale (1 = never, 2 = yes in lifetime but not recently, 3 = yes in lifetime and only once recently, 4 = yes in lifetime and recently more than once). Here, we assume that the four item responses form an underlying substance use construct, that is, the underlying latent variable is substance use behavior.

Three proportional odds logistic models were estimated with the AAYP substance use data. Results are listed in Table 1. In all three models, the fixed-effects parameters include four item intercepts and four item-by-time terms, that is, item slopes. In Model 1, the random intercept IRT model, four item discriminations and one random subject effect, the random intercept, are estimated. In Model 2, the random slope Rasch model, one common item discrimination and two random subject effects (random intercept and slope) are estimated. Model 3 is the proposed IRT model with multiple random subject effects. Here, random subject intercepts and slopes, plus four item discriminations, are estimated.

The results for the fixed-effects parameters from all three models generally agree in terms of indicating that at baseline (grade 5), students are more likely to respond in the lower category (no substance use), as indicated by the negative item estimates $(\hat{\beta}_1, \ldots, \hat{\beta}_4)$. However, over time the probabilities of all substance use items increase as indicated by positive item-by-time estimates $(\hat{\beta}_5, \ldots, \hat{\beta}_8)$. In addition, the items of marijuana use and getting drunk or high have lower intercepts at baseline than cigarette use and alcohol use, indicating cigarette use and alcohol use are more common at baseline. However, marijuana use and getting drunk or high increase faster over time than the other two items, as indicated by larger positive item-by-time terms.

Likelihood ratio tests show that Model 3 fits the data significantly better than both Model 1 ($\chi^2_2 = 30.5$, $p < 0.0001$) and Model 2 ($\chi^2_3 = 168.7$, $p < 0.0001$). This indicates that

Table 1
AAYP substance use data: Parameter estimates (SE) for models with different random-effects parameters

Parameters	Model 1 random intercept IRT model	Model 2 random slope Rasch model	Model 3 random slope IRT model
β_1: Item 1, cigarette use	−2.802 (0.127)*	−2.876 (0.152)*	−2.842 (0.153)*
β_2: Item 2, alcohol use	−2.892 (0.130)*	−3.142 (0.163)*	−2.966 (0.160)*
β_3: Item 3, marijuana use	−6.429 (0.234)*	−5.345 (0.193)*	−6.313 (0.241)*
β_4: Item 4, get drunk and high	−8.423 (0.372)*	−5.647 (0.207)*	−8.154 (0.371)*
β_5: Item 1 by time	0.927 (0.052)*	0.927 (0.069)*	0.953 (0.068)*
β_6: Item 2 by time	1.038 (0.050)*	1.084 (0.072)*	1.071 (0.069)*
β_7: Item 3 by time	1.924 (0.085)*	1.649 (0.082)*	1.924 (0.093)*
β_8: Item 4 by time	2.022 (0.121)*	1.479 (0.087)*	2.010 (0.126)*
σ_I: Intercept variance (Cholesky)	2.150 (0.083)*	2.430 (0.158)*	2.399 (0.156)*
σ_{IS}: Intercept-slope covariance		−0.207 (0.079)*	−0.235 (0.077)*
σ_S: Slope variance (Cholesky)		0.653 (0.056)*	0.612 (0.055)*
τ_1: Item 1 discrimination	1.322 (0.085)*	1.127 (0.055)*	1.108 (0.087)*
τ_2: Item 2 discrimination	0.924 (0.071)*		0.703 (0.075)*
τ_3: Item 3 discrimination	2.451 (0.130)*		2.171 (0.126)*
τ_4: Item 4 discrimination	3.670 (0.207)*		3.338 (0.206)*
γ_2: Threshold	1.528 (0.035)*	1.425 (0.033)*	1.537 (0.036)*
γ_3: Threshold	3.082 (0.060)*	2.870 (0.053)*	3.105 (0.062)*
LogL	−7622.29	−7691.38	−7607.04

*p-value < 0.05.

there is significant variation in both the individual intercept and linear trend, and that the discrimination parameters (factor loadings) for the four item responses differ significantly. It should be noted that there are concerns in using the likelihood ratio test for nesting the null hypothesis of variance parameters (because of boundary problems), and that standard application of this test does not reject the null hypothesis often enough (Berkhof and Snijders, 2001). In the present case, this is not the case since both Models 1 and 2 are deemed statistically different from Model 3. From the results in Model 3, a significant negative association between the intercept and slope terms (σ_{IS}) is detected, suggesting that those subjects with the lowest initial substance use level increase the most in substance use across time (e.g., largest positive time trends). This finding could be a result of a "ceiling effect," in that subjects with high initial substance use level cannot exhibit a large increase due to the limited range in the ordinal outcome items. In terms of the item discrimination parameters, the estimates for marijuana use and getting drunk or high (τ_3 and τ_4) are larger than those for cigarette use and alcohol use (τ_1 and τ_2). Since larger values of the discrimination parameter indicate that an item is more related to the underlying latent outcome, the results here suggest that the items of marijuana use and getting drunk or high are better at distinguishing an individual's level on the latent substance use construct. This is not surprising considering the relatively higher intercepts of cigarette use and alcohol use, indicating that use of these two items is more common than marijuana use and getting drunk or high. The item discrimination estimates additionally point out that cigarette and especially alcohol use are not as related to the underlying substance use construct.

So far, the only fixed-effects parameters in the model are the item intercepts and (time) slopes. However, the proposed model allows for a general form for covariates. To illustrate this, gender (1 = female, subject level) and gender-by-time terms (occasion-level) were added to the model. Results show that there are significant gender ($z = -5.66$) and gender-by-time ($z = 3.52$) effects, indicating that female subjects have lower intercepts than male subjects, that is, at grade 5, female students are less likely to use substances than male students. Nevertheless, since the gender-by-time interaction is also significant, the probability for female students to use substances increases faster across time than male students.

5. Nonproportional Odds Extension

The three-level ordinal IRT model that has been described thus far assumes proportional odds for all covariates. Thus, the effects of the covariates are assumed to be identical across all of the cumulative logits. As Peterson and Harrell (1990) point out, it is not hard to find violations of this assumption. In this case, it is more appropriate to allow the covariate effects (e.g., item) to differ across the logits. This can be done by adding to the model in equation (6) an additional set of covariates and a vector of regression coefficients, which vary across logits, namely

$$y_{ijk} = \boldsymbol{x}'_{ijk}\boldsymbol{\beta} + \boldsymbol{w}'_{ijk}\boldsymbol{T}_{(3)}\boldsymbol{\theta}_i + \boldsymbol{d}'_{ijk}\boldsymbol{T}_{(2)}\theta_{ij} + \boldsymbol{u}'_{ijk}\boldsymbol{\alpha}_c + \epsilon_{ijk}. \quad (18)$$

Here $\boldsymbol{u}_{ijk}$ represents the vector of covariates for which proportional odds are not assumed and $\boldsymbol{\alpha}_c$ are their regression coefficients that vary across logits. Because $\boldsymbol{\alpha}_c$ carries the c subscript, the effects of the covariates are allowed to vary across the $C - 1$ cumulative logits. These terms are often referred

Table 2
Parameter estimates (standard errors) for models without proportional odds assumption

Parameters	Model 4 nonproportional odds for items	Model 5 nonproportional odds for items and treatment
β_1: Item 1, cigarette use	−2.933 (0.164)*	−2.880 (0.217)*
β_2: Item 2, alcohol use	−3.070 (0.168)*	−3.019 (0.221)*
β_3: Item 3, marijuana use	−6.006 (0.236)*	−5.947 (0.274)*
β_4: Item 4, get drunk and high	−6.753 (0.283)*	−6.689 (0.310)*
β_5: Item 1 by time	1.035 (0.071)*	1.036 (0.072)*
β_6: Item 2 by time	1.159 (0.072)*	1.159 (0.072)*
β_7: Item 3 by time	1.736 (0.093)*	1.736 (0.093)*
β_8: Item 4 by time	1.619 (0.105)*	1.619 (0.106)*
β_9: Treatment		−0.077 (0.207)
σ_I: Intercept variance (Cholesky)	2.662 (0.168)*	2.666 (0.170)*
σ_{IS}: Intercept-slope covariance	−0.295 (0.082)*	−0.297 (0.082)*
σ_S: Slope variance (Cholesky)	0.676 (0.057)*	0.673 (0.057)*
τ_1: Item 1 discrimination	1.239 (0.110)*	1.231 (0.111)*
τ_2: Item 2 discrimination	0.803 (0.082)*	0.805 (0.083)*
τ_3: Item 3 discrimination	1.527 (0.161)*	1.523 (0.162)*
τ_4: Item 4 discrimination	1.918 (0.188)*	1.913 (0.190)*
γ_2: Threshold 2, cig	1.862 (0.076)*	1.969 (0.084)*
γ_3: Threshold 3, cig	3.702 (0.126)*	3.896 (0.142)*
α_{22}: Threshold 2 difference, Alc	−0.095 (0.108)	−0.097 (0.108)
α_{23}: Threshold 3 difference, Alc	0.225 (0.172)	0.222 (0.173)
α_{32}: Threshold 2 difference, Mar	0.943 (0.096)*	0.939 (0.095)*
α_{33}: Threshold 3 difference, Mar	1.600 (0.162)*	1.594 (0.163)*
α_{42}: Threshold 2 difference, D&H	1.288 (0.099)*	1.274 (0.099)*
α_{43}: Threshold 3 difference, D&H	1.818 (0.186)*	1.813 (0.186)*
α_{52}: Threshold 2 difference, Treatment		0.168 (0.061)*
α_{53}: Threshold 3 difference, Treatment		0.293 (0.104)*
LogL	−7477.29	−7473.61

*p-value < 0.05.

to as threshold interactions. Such threshold interaction terms have been considered in fixed-effects models by Peterson and Harrell (1990), and in two-level mixed models by Hedeker and Mermelstein (1998). To obtain the estimates of these nonproportional odds coefficients, we differentiate the marginal log likelihood as in equation (15), including $\partial z_{ijk}/\partial \boldsymbol{\alpha}_c' = \boldsymbol{u}_{ijk}'$ as a multiplicative term in the derivative.

5.1 *Application of the Nonproportional Odds Model*

To illustrate the nonproportional odds version of the model, we allow the substance use items to have different effects on the cumulative logits. That is, instead of restricting the thresholds to be equal across items as in Table 1, we add in item by threshold interactions, the vector of nonproportional odds coefficients $\boldsymbol{\alpha}_c$. As a result, the γ parameters represent the thresholds for item 1, cigarette use, and the $\boldsymbol{\alpha}_c$ parameters represent threshold differences attributable to the remaining items. Results for this partial-proportional odds model are listed in Table 2.

As seen in Model 4, based on a likelihood-ratio test, inclusion of these nonproportional odds coefficients significantly improves model fit, $\chi_6^2 = 259.5$, $p < 0.0001$. The results indicate that, with cigarette use as the reference, the thresholds for marijuana use and getting drunk or high are significantly different (the thresholds for alcohol use are not different from those for cigarette use). In Figure 1, we plot the baseline cumulative logits for all items using both Model 3, the proportional odds model (1a), and Model 4, the nonproportional odds model (1b). Without adding in item-by-time-by-threshold interactions, it is assumed that the spread of the three cumulative logits remains the same over time in both models. In Figure 1a, the spread of the three logits is the same for all items as the result of the proportional odds assumption. The logits for marijuana use and getting drunk or high are positioned higher in the figure, because more subjects are in category 1 (never) for these two items than for cigarette use and alcohol use. In Figure 1b, the spread of the logits for the items of marijuana use and getting drunk or high are significantly different from those for cigarette use due to the nonproportional odds relaxation. As the figure indicates, the thresholds are much closer together for these latter two items.

As mentioned earlier, AAYP was designed to compare interventions aimed at preventing health compromising behaviors in adolescents. Unfortunately, for the substance use behavior, the treatment effect is not observed to be significant in that neither the treatment nor the treatment-by-time terms are significant. In this analysis, which is not shown, the treatment variable is defined as 0 for the control group (HEC) and

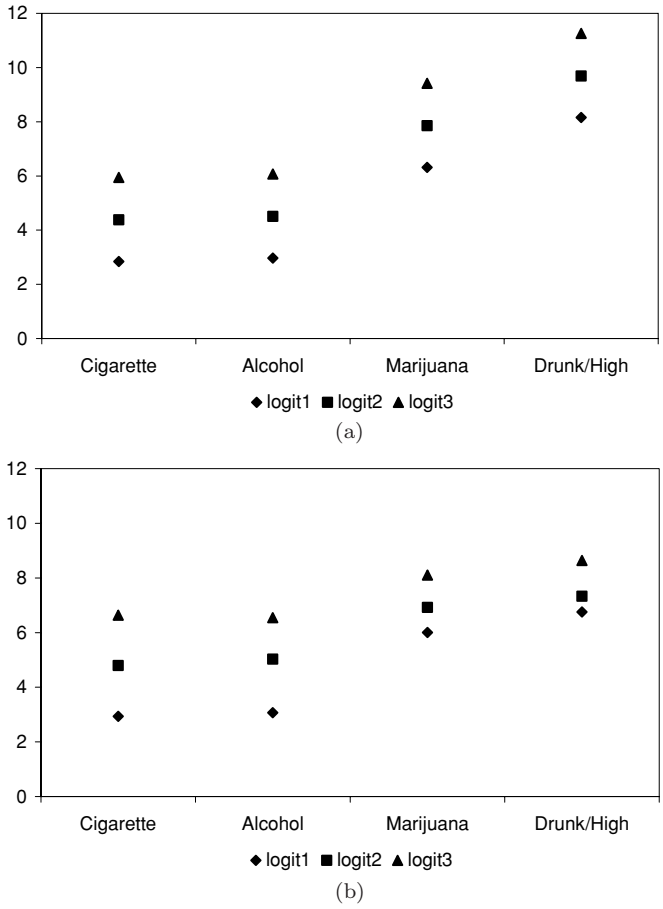

Figure 1. Cumulative logits for all four items at baseline. (a) Model 3: Proportional odds model. (b) Model 4: Nonproportional odds model for items.

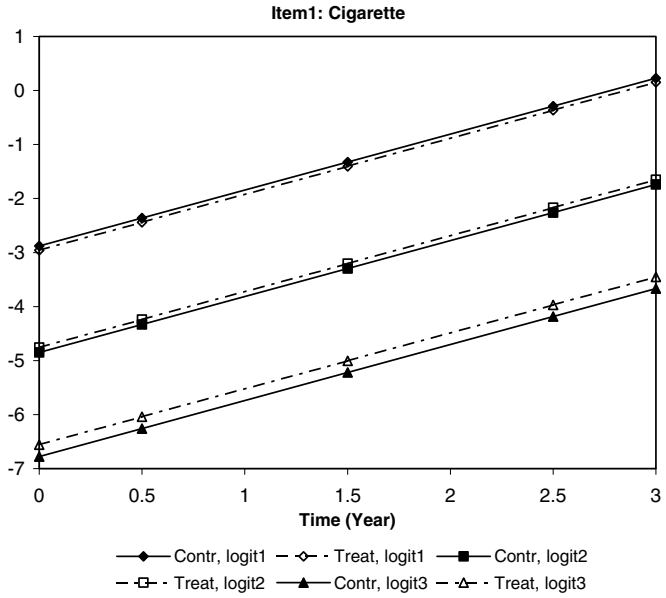

Figure 2. Nonproportional odds model for treatment (Model 5): Cumulative logits across time.

1 for the combination of the two treatment conditions (SDC and SC). To further examine a potential treatment effect and to illustrate application of the nonproportional odds model, we also considered whether the treatment effect was proportional across the cumulative logits. The results for this analysis are listed in Model 5, Table 2. This model indicates that, although the treatment effect is not significant for the first threshold (β_9), there are significant treatment-by-threshold 2 (α_{52}) and treatment-by-threshold 3 (α_{53}) interactions. Here, the second logit compares categories 3 and 4 versus categories 1 and 2, and the third logit compares category 4 versus categories 1, 2, and 3. Therefore, positive estimates for treatment-by-thresholds 2 and 3 interactions indicate higher response probabilities in higher categories (more substance use) for the treatment group relative to the control group. This result is illustrated in Figure 2.

Using the cigarette item as an example, the control-treatment difference for the first logit (comparing categories 2, 3, and 4 versus 1) is not significant across all time points. However, the lines of the second and third logit for the treatment group are above those of the control, indicating that subjects in the treatment group are more likely to respond in categories 3 and 4 across all time points. Note that without item-by-treatment-by-threshold interactions, the pattern of the treatment effect is the same for the other items. While it is unfortunate that this finding suggests an adverse effect of the treatment, our purpose here is to illustrate that our model can handle this kind of situation (i.e., nonproportional odds), which could lead to potentially interesting findings in other circumstances.

6. Discussion

We proposed a three-level mixed-effects IRT model for multivariate ordinal outcomes that includes multiple random subject effects and allows for different item discrimination parameters (item factor loadings). Covariates may be at any level and do not have to follow the proportional odds assumption. By relaxing this assumption for particular model covariates, one can test this assumption using a standard likelihood-ratio test. Parameter estimation is based on maximum marginal likelihood estimation using multidimensional Gauss–Hermite quadrature to numerically integrate over the distribution of random effects.

The analysis of a substance use data set supported use of the proposed model over other simpler models. In particular, the analysis presented indicated that the item factor loadings did vary and that there was appreciable heterogeneity in subjects' trends across time. Also, using a nonproportional odds model for treatment provided a more informative analysis than using a proportional odds model. In these data, it is seen that the treatment has a somewhat adverse effect in terms of the upper categories of the ordinal substance use variables. By allowing for nonproportional odds, one has the ability to find potentially interesting results that the proportional odds model might gloss over or even miss entirely.

As noted, the solution via quadrature can involve summation over a large number of points when the number of random effects is increased. This is because, at a given level, the

number of total quadrature points equals Q^r, where Q is the number of points per random-effects dimension and r is the number of random effects at that level. In the present example, there were two random subject effects and a single random effect for the items, and so the quadrature solution was viable. However, for models with many more random effects, it would be beneficial to employ adaptive quadrature methods that use fewer points per dimension, since the quadrature is adapted to the location and dispersion for each subject (Rabe-Hesketh, Pickles, and Skrondal, 2002).

Acknowledgements

This work was supported by the training grant from the Cancer Education and Career Development Program (5 R25 CA57699-12). The authors would like to thank Professor Brian Flay for use of the example data.

References

Adams, R., Wilson, M., and Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics* **22,** 47–76.

Berkhof, J. and Snijders, T. (2001). Variance component testing in multilevel models. *Educational and Behavioral Statistics* **26,** 133–152.

Bock, R. (1975). *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill.

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

Flay, B., Graumlich, S., Segawa, E., Burns, J., and Holliday, M. (2004). Effects of two prevention programs on high-risk behaviors among African-American youth: A randomized trial. *Archives of Pediatric and Adolescent Medicine* **158,** 377–384.

Fox, J. and Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66,** 269–286.

Gibbons, R. and Bock, R. (1987). Trend in correlated proportions. *Psychometrika* **52,** 113–124.

Gibbons, R. and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics* **53,** 1527–1537.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edition. New York: Halstead Press.

Hedeker, D. and Gibbons, R. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50,** 933–944.

Hedeker, D. and Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research* **33,** 427–455.

Lord, R. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillside, New Jersey: Erlbaum.

Magnus, J. (1988). *Linear Structures*. London: Charles Griffin.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika* **47,** 149–174.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B* **40,** 109–142.

Peterson, B. and Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics* **39,** 205–217.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal* **2,** 1–21.

Rasch, G. (1961). *On General Laws and the Meaning of Measurement in Psychology*. Berkeley: University of California Press.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods* **8,** 185–205.

Roy, J. and Lin, X. (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics* **56,** 1047–1054.

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: Changes in methadone treatment practices. *Journal of the American Statistical Association* **97,** 40–52.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* **17,** 34.

Stroud, A. and Sechrest, D. (1966). *Gaussian Quadrature Formulas*. Upper Saddle River, New Jersey: Prentice Hall.

Ten Have, T., Kunselman, A., and Tran, L. (1999). A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Statistics in Medicine* **18,** 947–960.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology* **43,** 39–55.

Received September 2004. *Revised April* 2005.
Accepted April 2005.