Research Report

Comparison of Multidimensional Item Response Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Ability Distributions

Shelby J. Haberman

Matthias von Davier

Yi-Hsuan Lee

August 2008

ETS RR-08-45

Comparison of Multidimensional Item Response Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Distributions

Shelby J. Haberman, Matthias von Davier, and Yi-Hsuan Lee
ETS, Princeton, NJ

August 2008

Abstract

Multidimensional item response models can be based on multivariate normal ability distributions or on multivariate polytomous ability distributions. For the case of simple structure in which each item corresponds to a unique dimension of the ability vector, some applications of the two-parameter logistic model to empirical data are employed to illustrate how, at least for the example under study, comparable results can be achieved with either approach. Comparability involves quality of model fit as well as similarity in terms of parameter estimates and computational time required. In both cases, numerical work can be performed quite efficiently. In the case of the multivariate normal ability distribution, multivariate adaptive Gauss-Hermite quadrature can be employed to greatly reduce computational labor. In the case of a polytomous ability distribution, use of log-linear models permits efficient computations.

Key words: Log penalty, 2PL model, general diagnostic model

Acknowledgments

The authors thank Andreas Oranje, Sandip Sinharay, and Dan Eignor for their helpful comments.

Multidimensional item response models are well-known in the psychometric literature but relatively little used in practice (Reckase, 2007). In this report, simple-structure multidimensional two-parameter logistic (2PL) models are considered in which each item is associated with one coordinate of the ability vector (Zhang, 2004). This restriction simplifies analysis to a considerable degree relative to approaches in which the relationship of items to coordinates of the ability vector is not specified. Two distinct models are considered for the distribution of the ability vector. In the first case, the ability vector is assumed to be a multivariate normal random vector with mean 0 and with a covariance matrix that has all diagonal elements equal to 1, so that each coordinate has variance 1, and with unknown off-diagonal elements that are the correlations of the coordinates of the ability vector. In the second case, the ability vector is assumed to have polytomous coordinates, a choice that may be able to reduce the computational burden associated with multidimensional model, but one that may seem less familiar than assuming a normal distribution. In the polytomous case, the realizations of each coordinate of the ability vector are from a discrete and finite set of real valued ability levels. Unidimensional models of this type are sometimes referred to as discrete latent trait models (Heinen, 1996) or located latent class models (Formann, 1992). In the case of multidimensional discrete latent traits, the term *diagnostic models* (von Davier, 2005) is employed. Each of the coordinate sets of ability levels may be different, and, for a given coordinate, it is common to use evenly spaced integers, so that the set of levels for a coordinate might be the two-member set $\{-1, +1\}$ or the three-member set $\{-1, 0, +1\}$. Sets used will often have four or more elements. In both models for the ability vector, algorithms are provided for computation of maximum likelihood. These algorithms are sufficiently efficient so that complete data from an assessment can be analyzed rapidly enough for practical use.

By means of the expected log penalty criterion (Gilula & Haberman, 1994), the two cases are compared in terms of their effectiveness at describing the observed data. In addition, the two cases are compared in terms of reliability of ability parameter estimates provided by the models. Approaches used do not assume that any model examined is valid, and comparisons involve measurement of the quality of prediction of response patterns rather than test of goodness of fit.

The basic conclusion suggested by the example studied is that the choice of latent-variable distribution has remarkably limited effect. This conclusion is consistent with a previous one-dimensional analysis (Haberman, 2005a), although it is possible that other examples can be found in which larger differences between model performance are evident.

In section 1, the multivariate two-parameter logistic (2PL) model under study is introduced. Section 2 considers application to a multivariate normal ability distribution. Section 3 considers application to multivariate polytomous ability distributions. Section 4 illustrates application of results to a Praxis™ administration. Section 5 provides conclusions based on the empirical results observed.

1 The Multidimensional 2PL Model

In the general model under study, a test is considered with $q \geq 2$ right-scored items. A sample of $n \geq 2$ examinees is used in analysis of the data. For examinee i, $1 \leq i \leq n$, for item j, $1 \leq j \leq q$, X_{ij} is 1 if the response to item j is correct, and X_{ij} is 0 otherwise. The q-dimensional vectors $\mathbf{X}_i$ with coordinates X_{ij}, $1 \leq j \leq q$, are independent and identically distributed for examinees i from 1 to n, and the set of possible values of $\mathbf{X}_i$ is denoted by Γ.

The basic 2PL model under study assumes that an r-dimensional random ability vector $\boldsymbol{\theta}_i$ with coordinates θ_{ik}, $1 \leq k \leq r$, is associated with each examinee i. The pairs $(\mathbf{X}_i, \boldsymbol{\theta}_i)$, $1 \leq i \leq n$, are independent and identically distributed, and, for each examinee i, the response variables X_{ij}, $1 \leq j \leq q$, are conditionally independent given $\boldsymbol{\theta}_i$. Let

$$P(h; y) = \exp(hy)/[1 + \exp(y)]$$

for h and y real.

To each item j, $1 \leq j \leq q$, corresponds an ability coordinate $\upsilon(j)$, $1 \leq \upsilon(j) \leq r$. For an unknown item discrimination a_j and an unknown real parameter γ_j, if $\boldsymbol{\omega}$ is a d-dimensional vector with coordinates ω_k, $1 \leq k \leq r$, then the conditional probability that $X_{ij} = h$ given $\boldsymbol{\theta}_i = \boldsymbol{\omega}$ is $P(h; a_j\omega_{\upsilon(j)} - \gamma_j)$. Provided that the discrimination a_j is positive, the item difficulty for item j is then $\gamma_j/a_j = b_j$. If r is 1, then one has a one-dimensional 2PL model, for

$$P(x_{ij}; a_j\omega_1 - \gamma_j) = \frac{\exp[x_{ij}a_j(\omega_1 - b_j)]}{1 + \exp[a_j(\omega_1 - b_j)]}.$$

If, in addition, all a_j are equal, then one has a one-dimensional one-parameter logistic (1PL) model. This model may also be termed a Rasch model. If $r > 1$, then the 2PL model is multidimensional. For $r > 1$, the assumption is made that, for $1 \leq k \leq r$, the set $\upsilon^{-1}(k)$ of items j, $1 \leq j \leq q$, with $\upsilon(j) = k$ is nonempty, so that each coordinate θ_{ik} of $\boldsymbol{\theta}_i$ corresponds to at least one item. If a_j is constant for j in $\upsilon^{-1}(k)$, $1 \leq k \leq r$, then one has a multidimensional 1PL model.

In all cases under study, a restrictive model is used for the distribution of the ability vector $\boldsymbol{\theta}_i$. In section 2, $\boldsymbol{\theta}_i$ is assumed to have a multivariate normal distribution with mean 0 and with a covariance matrix that has all diagonal elements equal to 1, so that each θ_{ik} is assumed to have variance 1. In section 3, the distribution of $\boldsymbol{\theta}_i$ is assumed to have all mass on a known finite set Ω, which represents the possible values of a multidimensional discrete ability vector.

2 The Multivariate Normal Case

In the multivariate normal case, the assumption is made that $\boldsymbol{\theta}_i$ has a multivariate normal distribution $N(\mathbf{0}, \mathbf{D})$. Here $\mathbf{0}$ is the r-dimensional vector with all coordinates 0, and $\mathbf{D}$ is an r-by-r positive-definite symmetric matrix with elements $d_{kk'}$, $1 \leq k \leq r$, $1 \leq k' \leq r$, such that each diagonal element $d_{kk} = 1$, and $d_{kk'}$, $k \neq k'$, is the unknown correlation of θ_{ik} and $\theta_{ik'}$. The assumption that the mean of $\boldsymbol{\theta}_i$ is $\mathbf{0}$ and the variance d_{kk} of each θ_{ik} is 1 is imposed to permit identification of the item parameters a_j and b_j for each item j from 1 to q. For comparison with the polytomous case presented in section 3, let d^{km} be row k and column m of $\mathbf{D}^{-1}$ for $1 \leq k \leq m \leq r$, let $|\mathbf{D}|$ denote the determinant of $\mathbf{D}$, and let δ_{km} be 1 for $k = m$ and 0 otherwise. Then the density $p_{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_i$ at a vector $\boldsymbol{\omega}$ with coordinates ω_k, $1 \leq k \leq r$, satisfies

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{\omega}) = -\frac{r\log(2\pi) + \log(|\mathbf{D}|)}{2} - \sum_{k=1}^{r}\sum_{m=1}^{k}[(1 - \delta_{km}/2)d^{km}]\omega_k\omega_m. \tag{1}$$

2.1 Model Parameters

The multivariate normal case can be parametrized so that a version of the stabilized Newton-Raphson algorithm (Haberman, 1988) can be readily applied. The basic requirement involves an appropriate decomposition of $\mathbf{D}$. If r is 1, then $\mathbf{D}$ reduces to the one-by-one matrix with element 1, and $\mathbf{D} = \mathbf{F}\mathbf{F}'$, where $\mathbf{F}$ and its transpose $\mathbf{F}'$ equal $\mathbf{D}$. By use of the Cholesky decomposition (Stewart, 1973, p. 134), it follows that, if $r > 1$, then $\mathbf{D}$ is determined by unique real constants $\tau_{kk'}, 1 \leq k' < k \leq r$, by the decomposition $\mathbf{D} = \mathbf{F}(\boldsymbol{\tau})[\mathbf{F}(\boldsymbol{\tau})]'$. Here $\boldsymbol{\tau}$ is an $r(r-1)/2$-dimensional vector with element $k' + k(k-1)/2$ equal to $\tau_{kk'}$ for $1 \leq k' < k$ and $1 \leq k \leq r$, and $\mathbf{F}(\boldsymbol{\tau})$ is an r-by-r matrix with elements $f_{kk'}(\boldsymbol{\tau})$, $1 \leq k \leq r$, $1 \leq k' \leq r$. The upper triangular elements $f_{kk'}(\boldsymbol{\tau}) = 0$ for $1 \leq k < k' \leq r$. Let

$$\nu_k(\boldsymbol{\tau}) = \left(1 + \sum_{k'=1}^{k-1}\tau_{kk'}^2\right)^{1/2}$$

for $2 \le k \le r$, and let $\nu_1(\boldsymbol{\tau})$ be 1. The diagonal element $f_{kk}(\boldsymbol{\tau}) = 1/\nu_k(\boldsymbol{\tau})$ for each integer k, $1 \le k \le r$, the first column in the lower triangle of $\mathbf{F}(\boldsymbol{\tau})$ satisfies $f_{k1}(\boldsymbol{\tau}) = \tau_{k1}/\nu_k(\boldsymbol{\tau})$ for $1 < k \le r$, and the remaining lower triangle of $\mathbf{F}(\boldsymbol{\tau})$ satisfies

$$f_{kk'}(\boldsymbol{\tau}) = \frac{\tau_{kk'}}{\nu_k(\boldsymbol{\tau})\nu_{k'}(\boldsymbol{\tau})}$$

for $1 < k' < k \le r$. The constants $\tau_{kk'}$ can have any combination of real values. If $r = 2$, then $\tau_{21} = d_{21}/(1 - d_{21})^{1/2}$, where d_{21} is the correlation of θ_{i2} and θ_{i1}. The distribution of $\boldsymbol{\theta}_i$ under the model is the same as the distribution of $\mathbf{F}(\boldsymbol{\tau})\mathbf{Z}$, where $\mathbf{Z}$ is an r-dimensional random vector with independent coordinates Z_k with standard normal distributions, $1 \le k \le r$. Let ϕ be the density function of the standard normal distribution, and let $\boldsymbol{\phi}_r$ be the function on R^r such that $\boldsymbol{\phi}_r(\mathbf{z}) = \prod_{k=1}^{r} \phi(z_k)$ for each r-dimensional real vector with coordinates z_k, $1 \le k \le r$. Thus $\boldsymbol{\phi}_r$ is the density of $\mathbf{Z}$.

Consider the vector $\boldsymbol{\beta}$ with $\nu = 2q + r(r-1)/2$ coordinates β_j, $1 \le j \le u$ such that $\beta_j = a_j$ for $1 \le j \le q$, $\beta_{q+j} = \gamma_j$ for $1 \le j \le q$, and $\beta_{2q+k'+(k-1)(k-2)/2}$ is $\tau_{kk'}$ if $1 \le k' < k \le r$. Let $\boldsymbol{\tau}(\boldsymbol{\beta})$ be the $r(r-1)/2$-dimensional vector with elements β_{2q+h} for $1 \le h \le r(r-1)/2$. Let $\mathbf{R}(\boldsymbol{\beta})$ be the one-by-one identity matrix if r is 1. Otherwise, let $\mathbf{R}(\boldsymbol{\beta})$ be $\mathbf{F}(\boldsymbol{\tau}(\boldsymbol{\beta}))$.

For any q-dimensional vector $\mathbf{x}$ with all coordinates 0 or 1, the probability that $\mathbf{X}_i = \mathbf{x}$ is then

$$p(\mathbf{x}; \boldsymbol{\beta}) = \int p\left(\mathbf{x} | \mathbf{R}(\boldsymbol{\beta})\mathbf{z}; \boldsymbol{\beta}\right) \boldsymbol{\phi}_r(\mathbf{z}) d\mathbf{z}.$$

For the r-dimensional vector $\boldsymbol{\omega}$ with coordinates ω_k, $1 \le k \le r$,

$$p(\mathbf{x}|\boldsymbol{\omega}; \boldsymbol{\beta}) = \prod_{j=1}^{q} P(x_j, \beta_j \omega_{\upsilon(j)} - \beta_{q+j})$$

is the conditional probability that $\mathbf{X}_i = \mathbf{x}$ given that $\boldsymbol{\theta}_i = \boldsymbol{\omega}$. If, for $1 \le k \le r$,

$$s_k(\mathbf{x}; \boldsymbol{\beta}) = \sum_{j=1}^{q} \delta_{k\upsilon(j)} \beta_j x_j,$$

and if

$$V(\mathbf{x}, \boldsymbol{\omega}; \boldsymbol{\beta}) = \prod_{j=1}^{q} \frac{\exp(-\beta_{q+j} x_j)}{1 + \exp(\beta_j \omega_{\upsilon(j)} - \beta_{q+j})},$$

then

$$p(\mathbf{x}|\boldsymbol{\omega}; \boldsymbol{\beta}) = V(\mathbf{x}, \boldsymbol{\omega}; \boldsymbol{\beta}) \exp\left[\sum_{k=1}^{r} s_k(\mathbf{x}; \boldsymbol{\beta}) \omega_k\right].$$

The log likelihood function is then

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}),$$

where

$$\ell_i(\boldsymbol{\beta}) = \log p(\mathbf{X}_i; \boldsymbol{\beta}), \ 1 \le i \le n.$$

If

$$K_{it}(\mathbf{z}) = p(\mathbf{X}_i | \mathbf{R}(\boldsymbol{\beta})\mathbf{z}; \boldsymbol{\beta}) \boldsymbol{\phi}_r(\mathbf{z})$$

for r-dimensional vectors $\mathbf{z}$, then

$$\ell_i(\boldsymbol{\beta}) = \log \int K_{it}(\mathbf{z}) d\mathbf{z}.$$

2.2 The Stabilized Newton-Raphson Algorithm

The likelihood function may be maximized by a simple variation on the stabilized Newton-Raphson algorithm (Haberman, 1974a, 1988). It is also possible to use the EM algorithm (Dempster, Laird, & Rubin, 1977); however, because the Hessian matrix of the log likelihood is not used in computations in this case, the EM algorithm is less helpful for estimation of asymptotic variances. The one major complication is the problem of r-dimensional quadrature. Adaptive Gauss-Hermite integration is appropriate for this problem (Haberman, 2006), although the multidimensional version of adaptive integration is a bit more complex than is the univariate version. Consider use of $s(k)$ quadrature points for dimension k, $1 \le k \le r$. Let v_{kh} and y_{kh}, $1 \le h \le s(k)$, be defined so that

$$\sum_{e=1}^{s(k)} y_e^m v_{kh} = \int y^m \phi(y) dy$$

for $1 \le m \le 2s(k) - 1$. Let $\hat{\boldsymbol{\beta}}$ denote the maximum-likelihood estimate of $\boldsymbol{\beta}$, so that $\ell(\hat{\boldsymbol{\beta}})$ is the supremum ℓ_* of $\ell(\boldsymbol{\beta})$ for all possible ν-dimensional vectors $\boldsymbol{\beta}$. Consider an iteration $t \ge 0$ of the stabilized Newton-Raphson algorithm. Let H be the set of all r-dimensional vectors $\mathbf{h}$ with coordinates $h(k)$, $1 \le h(k) \le s(k)$, $1 \le k \le r$. Thus H has $\prod_{k=1}^{r} s(k)$ elements. Let $\mathbf{y_h}$ be the vector with coordinates $y_{h(k)}$ for $1 \le k \le r$. Then

$$\int \pi(\mathbf{z}) \boldsymbol{\phi}_r(\mathbf{z}) d\mathbf{z} = \sum_{\mathbf{h} \in H} \pi(\mathbf{y_h}) \prod_{k=1}^{r} v_{kh(k)}$$

whenever $\pi(\mathbf{z})$ is a polynomial such that no power of a coordinate z_k exceeds $2s(k) - 1$.

To apply adaptive quadrature, consider an iteration $t \geq 0$. At the start of the iteration, let $\boldsymbol{\beta}_t$ be an approximation for the maximum-likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. The standard formula in calculus for change of variables permits $\ell_i(\boldsymbol{\beta})$ to be approximated by a function

$$\ell_{it}(\boldsymbol{\beta}) = \log L_{it}(\boldsymbol{\beta}),$$

where

$$L_{it}(\boldsymbol{\beta}) = |\mathbf{W}_{it}|^{-1} \sum_{\mathbf{h} \in H} [K_{it}(\mathbf{u}_{it\mathbf{h}})/\phi_r(\mathbf{y}_\mathbf{h})] \prod_{k=1}^{r} v_{kh(k)},$$

$$\mathbf{u}_{it\mathbf{h}} = (\mathbf{W}'_{it})^{-1}\mathbf{y}_\mathbf{h} + \mathbf{z}_i,$$

$|\mathbf{W}_{it}|$ is the determinant of $\mathbf{W}_{it}$, $\mathbf{W}_{it}$ is an r-by-r matrix with coordinates $w_{itkk'}$, $1 \leq k \leq r$, $1 \leq k' \leq r$, w_{itkk} is positive for $1 \leq k \leq r$, $w_{itkk'} = 0$ for $1 \leq k < k' \leq r$, $\mathbf{z}_{it}$ is an approximation to the location of the maximum over $\mathbf{z}$ in R^r of $K_{it}(\mathbf{z})$, and $\mathbf{W}_{it}\mathbf{W}'_{it} = -\nabla^2 K_{it}(\mathbf{z}_{it})$ for the Hessian matrix $\nabla^2 K_{it}(\mathbf{z}_{it})$ of K_{it} at $\mathbf{z}_{it}$. Note that $|\mathbf{W}_{it}|$ is the product of the w_{itkk} for $1 \leq k \leq r$ (Rao, 1973, p. 23).

With the starting value $\boldsymbol{\beta}_t$, one step of the stabilized Newton-Raphson algorithm is applied to $\ell_{St} = \sum_{i=1}^{n} \ell_{it}$ to yield a new approximation,

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \alpha_t \boldsymbol{\zeta}_t.$$

To define α_t and $\boldsymbol{\zeta}_t$, let κ and $\kappa^* < 1/2$ be given positive constants, let $|\mathbf{z}|$ be $\max_{1 \leq j \leq \nu} |z_j|$ for a ν-dimensional vector $\mathbf{z}$ with coordinates z_j, $1 \leq j \leq \nu$, let $\nabla \ell_{St}$ be the gradient of ℓ_{St}, let $\nabla^2 \ell_{St}$ be the Hessian matrix of ℓ_t, let $\mathbf{I}$ be the ν-by-ν identity matrix, let

$$\boldsymbol{\Lambda}_t = -\nabla^2 \ell_t(\boldsymbol{\beta}_t) + c_t \mathbf{I}$$

be positive definite, let

$$\boldsymbol{\zeta}_t = \boldsymbol{\Lambda}_t^{-1} \nabla \ell_t(\boldsymbol{\beta}_t),$$

let $|\boldsymbol{\zeta}_t| < \kappa$, and let $\alpha_t > 0$ satisfy

$$\ell_{St}(\boldsymbol{\beta}_{t+1}) - \ell_{St}(\boldsymbol{\beta}_t) > \alpha_t \kappa^* \boldsymbol{\zeta}'_t \nabla \ell_{St}(\boldsymbol{\beta}_t). \tag{2}$$

Here c_t is 0 if this choice satisfies the conditions that $\boldsymbol{\Lambda}_t$ is positive definite and $|\boldsymbol{\zeta}_t| < \kappa$. Otherwise, c_t is obtained by letting c_t^* be the maximum absolute value of a diagonal element of $\nabla^2 \ell_t(\boldsymbol{\beta}_t)$ and successively trying $\kappa^* c_t^*$, $(1 + 2^2)\kappa^* c_t^*$, $(1 + 2^2 + 3^2)\kappa^* c_t^*$, and so on. If (2) is satisfied with $\alpha_t = 1$, then α_t is set to 1. In general, α_t is found by use of a rough approximation to the maximum of $\ell_{St}(\boldsymbol{\beta}_t + \alpha\boldsymbol{\zeta}_t)$ for $\alpha > 0$ (Haberman, 1974a, 2006). The choices of $\kappa = 2$ and $\kappa^* = 1/16$ are used in calculations reported in this report.

For the example studied in this paper, use of $s_k = 4$ for each k was quite adequate for a case with $r = 4$, $q = 118$, and 29 or 30 items associated with each coordinate θ_{ik}. The choice of $s_k = 3$ for each coordinate k was also acceptable, and even $s_k = 2$ for each coordinate k was tolerable. These relatively small values are important, for $s_k = 4$ for each k and r equals 4 leads to 256 quadrature points, while $s_k = 3$ for each coordinate leads to 81 quadrature points, and $s_k = 2$ for each coordinate leads to 16 quadrature points. The relatively small number of points required is consistent with existing literature (Schilling & Bock, 2005). The quadrature situation with adaptive quadrature is far better than with the nonadaptive quadrature approach used in the National Assessment of Educational Progress (NAEP). This approach, found in the NAEP BGROUP program, uses 41 points for each coordinate (Sinharay & von Davier, 2005), so that, for $r = 4$, $41^4 = 2,825,761$ quadrature points would result. In practice, for more than two dimensions, NAEP uses the CGROUP program. This program employs a generalization of Laplace approximations for integral evaluation, so that the actual computational labor is much less than suggested by this comparison. Nonetheless, accuracy of the Laplace approach is an issue.

2.3 Estimated Expected Log Penalty

To evaluate the model, consider the expected log penalty

$$H(\mathbf{z}) = -q^{-1}E(\ell_i(\mathbf{z}))$$

per item (Gilula & Haberman, 1994). Consider the minimum I of $H(\mathbf{z})$ for ν-dimensional vectors $\mathbf{z}$ such that $z_j > 0$ for $1 \le j \le q$. Let $H(\boldsymbol{\beta}) = I$. If the 2PL model with multivariate normal ability vector is correct, then $\boldsymbol{\beta}$ is defined as in the model definition and I is the entropy per item of the vector $\mathbf{X}_i$. If $\boldsymbol{\beta}$ is uniquely defined, then $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}$ with probability 1 as the sample size n goes to ∞, whether or not the model holds, and $\hat{I} = (nq)^{-1}\ell_*$ converges to I. Let

$$\mathbf{Z} = E(-\nabla^2\ell_i(\boldsymbol{\beta})),$$

let

$$\mathbf{Y} = E(\nabla\ell_i(\boldsymbol{\beta})[\nabla\ell_i(\boldsymbol{\beta})]'),$$

and let tr denote a trace. Then $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a normal random vector with mean 0 and covariance matrix $\mathbf{Z}^{-1}\mathbf{Y}\mathbf{Z}^{-1}$. If the model holds, then $\mathbf{Y} = \mathbf{Z}$ and the covariance matrix is $\mathbf{Y}^{-1}$. The scaled difference $n^{1/2}(\hat{I} - I)$ converges in distribution to a normal random variable with mean 0 and variance equal to the variance of $q^{-1}\ell_i(\boldsymbol{\beta})$. The expected value of $\hat{I}$ is less than I. As n approaches ∞, $2nq[I - E(\hat{I})]$ converges to $\psi = \mathrm{tr}(\mathbf{Z}^{-1}\mathbf{Y})$. In addition, if $\mathbf{X}_0$ is independent of $\mathbf{X}_i$ for $1 \leq i \leq n$ and $\mathbf{X}_0$ has the same distribution as $\mathbf{X}_i$, then the conditional expectation $\tilde{I}_0$ of the log penalty $-q^{-1}\ell_0(\hat{\boldsymbol{\beta}})$ given $\mathbf{X}_i$, $1 \leq i \leq n$, for prediction of $\mathbf{X}_0$ satisfies the condition that $nq(\tilde{I}_0 - I)$ converges in distribution to a random variable with expectation ψ, and ψ is ν if the model holds. More generally, ψ is estimated by $\hat{\psi} = \mathrm{tr}(\hat{\mathbf{Z}}^{-1}\hat{\mathbf{Y}})$, where

$$\hat{\mathbf{Z}} = -n^{-1}\nabla^2\ell(\hat{\boldsymbol{\beta}})$$

and

$$\hat{\mathbf{Y}} = n^{-1}\sum_{i=1}^{n}\nabla\ell_i(\hat{\boldsymbol{\beta}})[\nabla\ell_i(\hat{\boldsymbol{\beta}})]'.$$

Thus $\tilde{I}_0$ may be approximated by $\hat{I}_0 = \hat{I} + \hat{\psi}/(nq)$. In practice, $\nabla\ell_i$ is approximated by use of adaptive Gaussian quadrature. A simplified approximation that assumes the model is correct is $\hat{I}_{a0} = \hat{I} + \nu/(nq)$. This approximation is the Akaike information criterion (AIC) divided by the number of items times twice the sample size (Akaike, 1974).

The AIC, which is used widely in model selection, balances the gain in log likelihood of a model (the improved model fit) against the cost in terms of parameters being estimated. Therefore, a model that fits the data better but needs a much larger number of parameters than competing models with just slightly lower estimated expected log penalty may not fare as well when evaluated by means of the AIC. The Gilula-Haberman criterion $\hat{I}_0$ generally leads to results similar to those obtained with the AIC criterion, although appreciable differences can arise when the model fits the data rather poorly. When sample sizes are large, $\hat{I}$, $\hat{I}_0$, and $\hat{I}_{a0}$ are normally very similar (Gilula & Haberman, 2001). This situation is helpful when the EM algorithm is employed, for estimation of $\hat{I}_0$ is less readily accomplished in this case than in the case of the stabilized Newton-Raphson algorithm.

2.4 Estimated Ability Parameters

The ability parameter $\boldsymbol{\theta}_i$ can be defined and approximated even if the underlying model is not accurate (?, ?). Let $\boldsymbol{\theta}_i$ be defined as a random vector such that the conditional distribution of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i = \mathbf{x}$ is the same as the conditional distribution of a random vector $\boldsymbol{\theta}_i^*$ given the random vector $\mathbf{X}_i^*$ with values in Γ, where $\boldsymbol{\theta}_i^*$ has a multivariate normal distribution with mean 0 and covariance matrix $\mathbf{R}(\boldsymbol{\beta})[\mathbf{R}(\boldsymbol{\beta})]'$ and the conditional probability that $\mathbf{X}_i^* = \mathbf{x}$ in Γ given $\boldsymbol{\theta}_i^* = \boldsymbol{\omega}$ is $p(\mathbf{x}|\boldsymbol{\omega};\boldsymbol{\beta})$. Thus the conditional density $p(\boldsymbol{\omega}|\mathbf{x};\boldsymbol{\beta})$ at $\boldsymbol{\omega}$ of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i = \mathbf{x}$ is given by Bayes's theorem to be

$$p_{\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\omega}|\mathbf{x};\boldsymbol{\beta}) = \frac{p(\mathbf{x}|\boldsymbol{\omega};\boldsymbol{\beta})\boldsymbol{\phi}_r([\mathbf{R}(\boldsymbol{\beta})]^{-1}\boldsymbol{\omega})}{|\mathbf{R}(\boldsymbol{\beta})|p(\mathbf{x};\boldsymbol{\beta})}.$$

If the model actually holds, then the definition of $\boldsymbol{\theta}_i$ in the model definition is consistent with the definition applied here. The information per item provided by $\boldsymbol{\theta}_i$ is

$$\Delta = I - I_{\boldsymbol{\theta}}.$$

Alternatively, $q\Delta$ is the information that $\mathbf{X}_i$ provides concerning $\boldsymbol{\theta}_i$. Here $I_{\boldsymbol{\theta}}$ is the expected value per item of the log penalty $-\log p(\mathbf{X}_i|\boldsymbol{\theta}_i;\boldsymbol{\beta})$ from use of the conditional probability approximation $p(\mathbf{X}_i|\boldsymbol{\theta}_i)$ for the conditional probability given $\boldsymbol{\theta}_i$ for the observed value of $\mathbf{X}_i$. One has

$$I_{\boldsymbol{\theta}} = q^{-1}\sum_{j=1}^{q} I_{j\boldsymbol{\theta}},$$

where

$$I_{j\boldsymbol{\theta}} = -E(\log P(X_{ij};\beta_j\theta_{v(j)} - \beta_{q+j})).$$

An application of Bayes's theorem shows that $I_{\boldsymbol{\theta}}$ may be estimated by

$$\hat{I}_{\boldsymbol{\theta}} = -(nq)^{-1}\sum_{i=1}^{n}[p(\mathbf{X}_i;\hat{\boldsymbol{\beta}})]^{-1}\int \boldsymbol{\phi}_r(\boldsymbol{\omega})p(\mathbf{X}_i;\mathbf{R}(\hat{\boldsymbol{\beta}})\boldsymbol{\omega};\hat{\boldsymbol{\beta}})\log p(\mathbf{X}_i;\mathbf{R}(\hat{\boldsymbol{\beta}})\boldsymbol{\omega};\hat{\boldsymbol{\beta}})d\boldsymbol{\omega}.$$

It follows that Δ has estimate $\hat{\Delta} = \hat{I} - \hat{I}_{\boldsymbol{\theta}}$.

The conditional expectation $\tilde{\boldsymbol{\theta}}_i = E(\boldsymbol{\theta}_i|\mathbf{X}_i)$ of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i$, the EAP estimate of $\boldsymbol{\theta}_i$ (Bock & Aitkin, 1981), is found from Bayes's theorem to be $E_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i;\boldsymbol{\beta})$, where

$$E_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x};\boldsymbol{\beta}) = \int \boldsymbol{\omega}p_{\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\omega}|\mathbf{x};\boldsymbol{\beta})d\boldsymbol{\omega}.$$

Although the expectation $E(\boldsymbol{\theta}_i) = E(\tilde{\boldsymbol{\theta}}_i)$ of $\boldsymbol{\theta}_i$ is the zero vector $\mathbf{0}$ under the model, the expectation need not be $\mathbf{0}$ if the model does not hold. The estimated conditional expectation of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i$ is $\hat{\boldsymbol{\theta}}_i = E_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i;\hat{\boldsymbol{\beta}})$. Computations may be performed by use of adaptive quadrature. One may employ $\hat{\boldsymbol{\theta}}_i$ as an estimate of $\boldsymbol{\theta}_i$. The expectation $E(\boldsymbol{\theta}_i)$ is then estimated by the average $\bar{\boldsymbol{\theta}} = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i$.

The covariance matrix of $\tilde{\boldsymbol{\theta}}_i$ is

$$\operatorname{Cov}(\tilde{\boldsymbol{\theta}}) = E([\tilde{\boldsymbol{\theta}}_i - E(\boldsymbol{\theta}_i)][\tilde{\boldsymbol{\theta}}_i - E(\boldsymbol{\theta}_i)]').$$

The corresponding estimate is

$$\widehat{\operatorname{Cov}}(\tilde{\boldsymbol{\theta}}) = n^{-1}\sum_{i=1}^{n}[\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}][\hat{\boldsymbol{\theta}}_i - \bar{\boldsymbol{\theta}}]'.$$

The conditional covariance matrix $\widetilde{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X})$ of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i$ may be used to assess the accuracy with which the data $\mathbf{X}_i$ determine $\boldsymbol{\theta}_i$. One has $\widetilde{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X}) = \operatorname{Cov}_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i;\boldsymbol{\beta})$, where

$$\operatorname{Cov}_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x};\boldsymbol{\beta}) = \int [\boldsymbol{\omega} - E_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x};\boldsymbol{\beta})][\boldsymbol{\omega} - E_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x};\boldsymbol{\beta})]' p_{\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\omega}|\mathbf{x};\boldsymbol{\beta})d\boldsymbol{\omega}.$$

The estimate of $\widetilde{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X})$ is then $\widehat{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X}) = \operatorname{Cov}_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i;\hat{\boldsymbol{\beta}})$. The expected conditional covariance matrix $E(\widehat{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X}))$ is then estimated by

$$\overline{\operatorname{Cov}}(\boldsymbol{\theta}|\mathbf{X}) = n^{-1}\sum_{i=1}^{n}\widehat{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X}).$$

For any nonzero r-dimensional vector $\mathbf{c}$, the reliability of $\mathbf{c}'\tilde{\boldsymbol{\theta}}_i$ is

$$\rho^2(\mathbf{c}) = \frac{\mathbf{c}'\operatorname{Cov}(\tilde{\boldsymbol{\theta}})\mathbf{c}}{\mathbf{c}'E(\widehat{\operatorname{Cov}}_i(\boldsymbol{\theta}|\mathbf{X}))\mathbf{c} + \mathbf{c}'\operatorname{Cov}(\tilde{\boldsymbol{\theta}})\mathbf{c}}.$$

The reliability of $\mathbf{c}'\hat{\boldsymbol{\theta}}_i$ is approximately the same in large samples, and the estimated reliability is then

$$\hat{\rho}^2(\mathbf{c}) = \frac{\mathbf{c}'\widehat{\operatorname{Cov}}(\tilde{\boldsymbol{\theta}})\mathbf{c}}{\mathbf{c}'\overline{\operatorname{Cov}}(\boldsymbol{\theta}|\mathbf{X})\mathbf{c} + \mathbf{c}'\widehat{\operatorname{Cov}}(\tilde{\boldsymbol{\theta}})\mathbf{c}}.$$

3 The Polytomous Case

In the polytomous case, the assumption is made that the distribution of $\boldsymbol{\theta}_i$ is confined to a finite set Ω with M elements. Often, the set of multidimensional ability levels Ω will be a nonempty subset of the Cartesian product $\prod_{k=1}^{r}\Omega_k$ of sets Ω_k, $1 \le k \le r$, where Ω_k is a subset of the real line that contains $c_k > 1$ possible values of θ_{ik}. In typical cases, Ω_k is the set of integers from $-(c_k-1)/2$ to $(c_k-1)/2$ if c_k is odd, and Ω_k is the set of integers $-c_k - 1 + 2d$ for integers d from 1 to c_k if c_k is even. Thus Ω_k is $\{-1, 1\}$ for $c_k = 2$ and $\{-1, 0, 1\}$ for $c_k = 3$. Computations are most rapid if the number of elements of Ω is small. Thus permitting Ω to have fewer than the $\prod_{k=1}^{r} c_k$ elements of $\prod_{k=1}^{r}\Omega_k$ can save computational labor. Of course, such a saving is only appropriate if the ability of the model to predict the joint distribution of the $\mathbf{X}_i$ is not impaired to a substantial degree.

For each $\boldsymbol{\omega}$ in Ω, the probability $p_{d\boldsymbol{\theta}}(\omega)$ that $\boldsymbol{\theta}_i = \boldsymbol{\omega}$ is assumed positive, and it is assumed that the $p_{d\boldsymbol{\theta}}(\omega)$ satisfy a log-linear model

$$\log p_{d\boldsymbol{\theta}}(\boldsymbol{\omega}) = \lambda + T_0(\boldsymbol{\omega}) + \sum_{g=1}^{G} \tau_{dg} T_g(\boldsymbol{\omega})$$

for known constants $T_g(\boldsymbol{\omega})$,$0 \leq g \leq G < M$, and unknown parameters λ and τ_{dg}, $1 \leq g \leq G$. Given the $T_g(\boldsymbol{\omega})$ and the τ_{dg}, λ is determined by the requirement that the sum of the $p_{d\boldsymbol{\theta}}(\boldsymbol{\omega})$, $\boldsymbol{\omega}$ in Ω, must be 1. To provide any possibility that the τ_{dg}, $1 \leq g \leq G$, can be identified, it is assumed that no real constants u_g, $1 \leq g \leq G$, exist such that some u_g is not zero and $\sum_{g=1}^{G} u_g T_g(\boldsymbol{\omega})$ has the same value for all $\boldsymbol{\omega}$ in Ω. Even with these constraints on G and on the $T_g(\boldsymbol{\omega})$, the τ_{dg}, $1 \leq g \leq G$, cannot be identified unless $2q + G$ is less than $2^q - 1$ (Haberman, 2005a), and, in practice, identification of parameters is much more difficult unless G and the $T_g(\boldsymbol{\omega})$, $0 \leq g \leq G$, $\boldsymbol{\omega}$ in Ω, are carefully selected.

The basic log-linear model to consider is analogous to the multivariate normal distribution applied in the continuous case. One considers a log-linear model with no main effects and with only linear-by-linear interactions, so that, for $\bar{\omega}_k$ the arithmetic mean of the elements of Ω_k,

$$\log p_{d\boldsymbol{\theta}}(\boldsymbol{\omega}) = \lambda + \sum_{k=1}^{r} \sum_{m=1}^{k} \eta_{km} (\omega_k - \bar{\omega}_k)(\omega_m - \bar{\omega}_k). \tag{3}$$

With no restrictions imposed on the η_{km}, this model has $G = r(r+1)/2$ independent parameters. Comparison of (1) and (3) shows that $\log p_{\boldsymbol{\theta}}$ and $\log p_{d\boldsymbol{\theta}}$ have a very similar form, especially in the typical case in which $\bar{\omega}_k = 0$.

More general use of polynomials can be considered. For $1 \leq k \leq r$, let O_{kh}, $0 \leq h < c_k$, be the orthogonal polynomial of degree h that corresponds to the elements of Ω_k and to some positive weighting function w_k on Ω_k, so that

$$\sum_{\omega_k \in \Omega_k} w_k(\omega_k) O_{kh}(\omega_k) O_{km}(\omega_k) = \delta_{hm}$$

for $0 \leq h \leq m < c_k$. Let Ξ be a nonempty set of vectors $\boldsymbol{\xi}$ with integer elements $\xi(k)$, $0 \leq \xi(k) < c_k$, for $1 \leq k \leq r$. Assume that no vector in Ξ has all coordinates 0. Then Ξ defines a log-linear model

$$\log p_{d\boldsymbol{\theta}}(\boldsymbol{\omega}) = \lambda + \sum_{\boldsymbol{\xi} \in \Xi} \zeta_{\boldsymbol{\xi}} \prod_{k=1}^{r} O_{k\xi(k)}(\omega_k). \tag{4}$$

Models of this kind have a long history in the literature on log-linear models (Haberman, 1974b) and variations have begun to appear with general diagnostic models (Xu & von Davier, 2007). The model specified by (4) is equivalent to the model specified by (3) if Ξ consists of all vectors $\boldsymbol{\xi}$ with either two coordinates equal to 1 and all other coordinates 0 or with one coordinate equal to 2 and all other coordinates 0. General diagnostic models have applied (4) with Ξ consisting of all vectors $\boldsymbol{\xi}$ that correspond to the model of (3) together with all additional vectors $\boldsymbol{\xi}$, which have all coordinates but one equal to 0 and one nonzero coordinate with a value between two specified positive integers.

3.1 Model Parameters

As in the multivariate normal case, the polytomous case can be parametrized so that a version of the stabilized Newton-Raphson algorithm (Haberman, 1988) can be readily applied. Alternatively, polytomous discrete cases can be specified as multidimensional discrete latent trait models and estimated with the EM algorithm, for example using the software *mdltm* (von Davier, 2005).

For the log likelihood to be maximized, consider the vector $\boldsymbol{\beta}_d$ with $\nu_d = 2q + G$ coordinates β_{dj}, $1 \le j \le u$ such that $\beta_{dj} = a_j$ for $1 \le j \le q$, $\beta_{d(q+j)} = \gamma_j$ for $1 \le j \le q$, and $\beta_{d(2q+g)}$ is τ_{dg} for $1 \le g \le G$. Let

$$\chi(\boldsymbol{\beta}_d) = \sum_{\boldsymbol{\omega}\in\Omega} \exp\left[T_0(\boldsymbol{\omega}) + \sum_{g=1}^{G} \beta_{d(2q+g)} T_g(\boldsymbol{\omega})\right],$$

and let

$$p_{d\boldsymbol{\theta}}(\boldsymbol{\omega};\boldsymbol{\beta}_d) = [\chi(\boldsymbol{\beta}_d)]^{-1} \exp\left[T_0(\boldsymbol{\omega}) + \sum_{g=1}^{G} \beta_{d(2q+g)} T_g(\boldsymbol{\omega})\right].$$

For any q-dimensional vector $\mathbf{x}$ with all coordinates 0 or 1, the probability that $\mathbf{X}_i = \mathbf{x}$ is then

$$p_d(\mathbf{x};\boldsymbol{\beta}_d) = \sum_{\boldsymbol{\omega}\in\Omega} p_d(\mathbf{x}|\boldsymbol{\omega};\boldsymbol{\beta}_d) p_{d\boldsymbol{\theta}}(\boldsymbol{\omega};\boldsymbol{\beta}_d).$$

For the r-dimensional vector $\boldsymbol{\omega}$ with coordinates ω_k, $1 \le k \le r$,

$$p_d(\mathbf{x}|\boldsymbol{\omega};\boldsymbol{\beta}_d) = \prod_{j=1}^{q} P(x_j, \beta_{dj}\omega_{\upsilon(j)} - \beta_{d(q+j)})$$

is the conditional probability that $\mathbf{X}_i = \mathbf{x}$ given that $\boldsymbol{\theta}_i = \boldsymbol{\omega}$. If, for $1 \le k \le r$,

$$s_{dk}(\mathbf{x};\boldsymbol{\beta}_d) = \sum_{j=1}^{q} \delta_{\upsilon(j)k} \beta_{dj} x_j,$$

and if

$$V_d(\mathbf{x}, \boldsymbol{\omega}; \boldsymbol{\beta}_d) = \prod_{j=1}^{q} \frac{\exp(-\beta_{q+j} x_j)}{1 + \exp(\beta_{dj}\omega_{v(j)} - \beta_{d(q+j)})},$$

then

$$p_d(\mathbf{x}|\boldsymbol{\omega}; \boldsymbol{\beta}_d) = V_d(\mathbf{x}, \boldsymbol{\omega}; \boldsymbol{\beta}_d) \exp\left[\sum_{k=1}^{r} s_{dk}(\mathbf{x}; \boldsymbol{\beta}_d)\omega_k\right].$$

The log likelihood is then

$$\ell_d(\boldsymbol{\beta}_d) = \sum_{i=1}^{n} \ell_{di}(\boldsymbol{\beta}_d),$$

where

$$\ell_{di}(\boldsymbol{\beta}_d) = \log p_d(\mathbf{X}_i; \boldsymbol{\beta}_d),\ 1 \le i \le n.$$

For the maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_d$, $\ell_d(\hat{\boldsymbol{\beta}}_d)$ is the supremum ℓ_{d*} of $\ell_d(\boldsymbol{\beta}_d)$ for all ν_d-dimensional vectors $\boldsymbol{\beta}_d$.

Unlike in the multivariate normal case, considerable care is needed in the polytomous case to understand when models really differ. For example, consider a positive constant z_k and a real constant u_k for $1 \le k \le r$. Replace each ω_k in Ω_k by $z_k\omega_k + u_k$, divide each item discrimination a_j by z_k if $v(j) = k$, change each intercept parameter γ_j to $\gamma_j - u_k a_j/z_k$ if $v(j) = k$, and let (3) continue to hold for $\boldsymbol{\omega}$ in Ω with each η_{km} divided by $z_k z_m$. Then the probabilities $p_d(\mathbf{x}; \boldsymbol{\beta}_d)$ are unchanged for $\mathbf{x}$ in Γ. It follows that the selection of Ω_k to consist of evenly spaced integers with mean 0 is equivalent in terms of the resulting model to any selection of Ω_k in which the members of Ω_k are evenly spaced points. Thus $\Omega_k = \{-1, 0, 1\}$ leads to the same model as $\Omega_k = \{1, 1.5, 2\}$.

In addition, the connection with the multivariate normal case is stronger than might at first be apparent. Define the covariance matrix $\mathbf{D}$ and the elements d^{km} of $\mathbf{D}^{-1}$ as in the multivariate normal case. If $\eta_{km} = (1 - \delta_{km}/2)d^{km}$, Ω_k consists of numbers $(-c_k - 1/2 + 2d)z_k$, $1 \le d \le c_k$, where $z_k > 0$, z_k approaches 0 and $c_k z_k$ approaches ∞, then $\boldsymbol{\theta}_i$ converges in distribution to a multivariate normal random vector with mean 0 and with covariance matrix $\mathbf{D}$. The argument required involves use of an auxiliary r-dimensional random vector $\boldsymbol{u}$, which is independent of $\boldsymbol{\theta}_i$ and has independent coordinates u_k with uniform distributions on $(-z_k/2, z_k/2)$. It is a straightforward matter to show that $\boldsymbol{\theta}_i + \boldsymbol{u}$ is a continuous random vector with a density that approaches the multivariate normal density $p_{\boldsymbol{\theta}}$ defined in (1). Application of Scheffé's theorem and the Mann-Wald theorem yield the desired result (Rao, 1973, pp. 122–125). Given the previous observations concerning the effects of linear transformations of the elements of Ω_k for $1 \le k \le r$, the practical consequence of the result is that, for any $\epsilon > 0$, there exists an integer $c > 0$ such that $\ell_{d*} > \ell_* - \epsilon$ whenever each $c_k > c$ and $\Omega = \prod_{k=1}^r \Omega_k$. Thus polytomous models must be competitive with multivariate normal models in terms of model fit for sufficiently large c_k. As evident from the data analysis, polytomous models are attractive even for all c_k equal to 4 or 5, and it is possible to use Ω with somewhat fewer elements than $\prod_{k=1}^r \Omega_k$ with little loss.

3.2 The Stabilized Newton-Raphson Algorithm

The log likelihood may be maximized by a simple variation on the stabilized Newton-Raphson algorithm (Haberman, 1974a, 1988). The EM algorithm can also be employed (von Davier, 2005; Xu & von Davier, 2007). In the polytomous case, no integrals are evaluated, so that adaptive Gauss-Hermite quadrature is not required and calculations are simpler. Nonetheless, some restrictions on the size of G are required to ensure that the model parameters are well-enough identified to permit a reasonable rate of convergence (Haberman, 2005a). Consider an iteration $t \geq 0$. At the start of the iteration, let $\boldsymbol{\beta}_{dt}$ be an approximation for the maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_d$ of $\boldsymbol{\beta}_d$. The stabilized Newton-Raphson algorithm yields a new approximation

$$\boldsymbol{\beta}_{d(t+1)} = \boldsymbol{\beta}_{dt} + \alpha_{dt}\boldsymbol{\zeta}_{dt}.$$

To define α_{dt} and $\boldsymbol{\zeta}_{dt}$, let κ_d and $\kappa_d^* < 1/2$ be given positive constants, let $\nabla\ell_d$ be the gradient of ℓ_d, let $\nabla^2\ell_d$ be the Hessian matrix of ℓ_d, let $\mathbf{I}_d$ be the ν_d-by-ν_d identity matrix, let $c_{dt} \geq 0$, let

$$\boldsymbol{\Lambda}_{dt} = -\nabla^2\ell_d(\boldsymbol{\beta}_{dt}) + c_{td}\mathbf{I}_d,$$

let

$$\boldsymbol{\zeta}_{dt} = \boldsymbol{\Lambda}_{dt}^{-1}\nabla\ell_{dt}(\boldsymbol{\beta}_{dt}),$$

let $|\boldsymbol{\zeta}_{dt}| \leq \kappa$, and let $\alpha_{dt} > 0$ satisfy

$$\ell_d(\boldsymbol{\beta}_{d(t+1)}) - \ell_d(\boldsymbol{\beta}_{dt}) > \alpha_{dt}\kappa_d^*\boldsymbol{\zeta}_{dt}'\nabla\ell_d(\boldsymbol{\beta}_{dt}). \tag{5}$$

As in the multivariate normal case, $c_{dt} = 0$ and $\alpha_{dt} = 1$ are used if all constraints are satisfied. Procedures for finding alternative values are the same, and use of $\kappa_d = 2$ and $\kappa_d^* = 1/16$ appears acceptable.

3.3 Estimated Expected Log Penalty

As in the multivariate normal case, to evaluate the model, consider the expected log penalty

$$H_d(\mathbf{z}) = -q^{-1}E(\ell_{di}(\mathbf{z}))$$

per item. Consider the minimum I_d of $H_d(\mathbf{z})$ for ν_d-dimensional vectors $\mathbf{z}$ such that $z_j > 0$ for $1 \leq j \leq q$. Let $H_d(\boldsymbol{\beta}_d) = I$. If the 2PL model with a polytomous ability vector is correct, then $\boldsymbol{\beta}_d$ is defined as in the model definition and I_d is the entropy per item of the vector $\mathbf{X}_i$. If $\boldsymbol{\beta}_d$ is uniquely defined, then $\hat{\boldsymbol{\beta}}_d$ converges to $\boldsymbol{\beta}_d$ with probability 1 as the sample size n goes to ∞, whether or not the model holds, and $\hat{I}_d = (nq)^{-1}\ell_d(\hat{\boldsymbol{\beta}})$ converges to I_d. Let

$$\mathbf{Z}_d = E(-\nabla^2\ell_{di}(\boldsymbol{\beta}_d)),$$

and let

$$\mathbf{Y}_d = E(\nabla\ell_{di}(\boldsymbol{\beta}_d)[\nabla\ell_{di}(\boldsymbol{\beta}_d)]').$$

Then $n^{1/2}(\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d)$ converges in distribution to a normal random vector with mean 0 and covariance matrix $\mathbf{Z}_d^{-1}\mathbf{Y}_d\mathbf{Z}_d^{-1}$. If the model holds, then $\mathbf{Y}_d = \mathbf{Z}_d$ and the covariance matrix is $\mathbf{Y}_d^{-1}$. The scaled difference $n^{1/2}(\hat{I}_d - I_d)$ converges in distribution to a normal random variable with mean 0 and a variance equal to the variance of $q^{-1}\ell_{di}(\boldsymbol{\beta})$. The expected value of $\hat{I}_d$ is less than I_d. As n approaches ∞, $2nq[I_d - E(\hat{I}_d)]$ converges to $\psi_d = \mathrm{tr}(\mathbf{Z}_d^{-1}\mathbf{Y}_d)$. In addition, if $\mathbf{X}_0$ is independent of $\mathbf{X}_i$ for $1 \leq i \leq n$ and $\mathbf{X}_0$ has the same distribution as $\mathbf{X}_i$, then the conditional expectation $\tilde{I}_{d0}$ of the log penalty $-\ell_{d0}(\hat{\boldsymbol{\beta}}_d)$ given $\mathbf{X}_i$, $1 \leq i \leq n$, for prediction of $\mathbf{X}_0$ satisfies the condition that $nq(\tilde{I}_{d0} - I_d)$ converges in distribution to a random variable with expectation ψ_d, and ψ_d is ν_d if the model holds. More generally, ψ_d is estimated by $\hat{\psi}_d = \mathrm{tr}(\hat{\mathbf{Z}}_d^{-1}\hat{\mathbf{Y}}_d)$, where

$$\hat{\mathbf{Z}}_d = -n^{-1}\nabla^2\ell_d(\hat{\boldsymbol{\beta}}_d)$$

and

$$\hat{\mathbf{Y}}_d = n^{-1}\sum_{i=1}^{n}\nabla\ell_{di}(\hat{\boldsymbol{\beta}}_d)[\nabla\ell_{di}(\hat{\boldsymbol{\beta}}_d)]'.$$

Thus $\tilde{I}_{d0}$ may be approximated by $\hat{I}_{d0} = \hat{I}_d + \hat{\psi}_d/(nq)$. If the model is correct, then the simplified Akaike approximation $\hat{I}_{da0} = \hat{I} + \nu/(nq)$ may be employed.

3.4 Estimated Ability Parameters

As in the multivariate normal case, the ability parameter $\boldsymbol{\theta}_i$ can be defined and approximated even if the underlying model is not accurate (?, ?). Let $\boldsymbol{\theta}_{di}$ be defined as a random vector such that the conditional distribution of $\boldsymbol{\theta}_{di}$ given $\mathbf{X}_i = \mathbf{x}$ is the same as the conditional distribution of a random vector $\boldsymbol{\theta}^*_{di}$ given the random vector $\mathbf{X}^*_i$ with values in Γ, where $\boldsymbol{\theta}^*_i = \boldsymbol{\omega}$, $\boldsymbol{\omega}$ in Ω, with probability $p_{d\boldsymbol{\theta}}(\boldsymbol{\omega}; \boldsymbol{\beta}_d)$ and the conditional probability that $\mathbf{X}^*_i = \mathbf{x}$ in Γ given $\boldsymbol{\theta}^*_{di} = \boldsymbol{\omega}$ is $p_d(\mathbf{x}|\boldsymbol{\omega}; \boldsymbol{\beta}_d)$. Thus the conditional probability $p_d(\boldsymbol{\omega}|\mathbf{x}; \boldsymbol{\beta}_d)$ that $\boldsymbol{\theta}_{di} = \boldsymbol{\omega}$ of given $\mathbf{X}_i = \mathbf{x}$ is given by Bayes's theorem to be

$$p_{d\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\omega}|\mathbf{x}; \boldsymbol{\beta}_d) = \frac{p_d(\mathbf{x}|\boldsymbol{\omega}; \boldsymbol{\beta}_d) p_{d\boldsymbol{\theta}}(\omega; \boldsymbol{\beta}_d)}{p_d(\mathbf{x}; \boldsymbol{\beta}_d)}.$$

If the model actually holds, then $\boldsymbol{\theta}_i$ in the model definition has the same distribution as $\boldsymbol{\theta}_{di}$. The information per item provided by $\boldsymbol{\theta}_{di}$ is

$$\Delta_d = I_d - I_{d\boldsymbol{\theta}}.$$

Alternatively, $q\Delta_d$ is the information that $\mathbf{X}_i$ provides concerning $\boldsymbol{\theta}_{di}$. Here $I_{d\boldsymbol{\theta}}$ is the expected value per item of the log penalty $-\log p_d(\mathbf{X}_i|\boldsymbol{\theta}_{di}; \boldsymbol{\beta}_d)$ from use of the conditional probability approximation $p_d(\mathbf{X}_i|\boldsymbol{\theta}_{di})$ for the conditional probability given $\boldsymbol{\theta}_{di}$ for the observed value of $\mathbf{X}_i$. One has

$$I_{d\boldsymbol{\theta}} = q^{-1} \sum_{j=1}^{q} I_{dj\boldsymbol{\theta}},$$

where

$$I_{dj\boldsymbol{\theta}} = -E(\log P(X_{ij}; \beta_{dj}\theta_{dk(j)} - \beta_{d(q+j)})).$$

An application of Bayes's theorem shows that $I_{d\boldsymbol{\theta}}$ may be estimated by

$$\hat{I}_{d\boldsymbol{\theta}} = -(nq)^{-1} \sum_{i=1}^{n} [p_d(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_d)]^{-1} \sum_{\omega \in \Omega} p_d(\mathbf{X}_i|\boldsymbol{\omega}; \hat{\boldsymbol{\beta}}_d) \log p_d(\mathbf{X}_i|\boldsymbol{\omega}; \hat{\boldsymbol{\beta}}_d).$$

It follows that Δ_d has estimate $\hat{\Delta}_d = \hat{I}_d - \hat{I}_{d\boldsymbol{\theta}}$.

The conditional expectation $\tilde{\boldsymbol{\theta}}_{di} = E(\boldsymbol{\theta}_{di}|\mathbf{X}_i)$ of $\boldsymbol{\theta}_{di}$ given $\mathbf{X}_i$ is found from Bayes's theorem to be $E_{\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i; \boldsymbol{\beta}_d)$, where

$$E_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x}; \boldsymbol{\beta}_d) = \sum_{\boldsymbol{\omega} \in \Omega} \boldsymbol{\omega} p_{d\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\omega}|\mathbf{x}; \boldsymbol{\beta}_d).$$

As in the multivariate normal case, although the expectation $E(\boldsymbol{\theta}_{di}) = E(\tilde{\boldsymbol{\theta}}_{di})$ of $\boldsymbol{\theta}_{di}$ is the zero vector $\mathbf{0}$ under the model, the expectation need not be $\mathbf{0}$ if the model does not hold. The estimated conditional expectation of $\boldsymbol{\theta}_{di}$ given $\mathbf{X}_i$ is $\hat{\boldsymbol{\theta}}_{di} = E_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_d)$. One may employ $\hat{\boldsymbol{\theta}}_{di}$ as an estimate of $\boldsymbol{\theta}_{di}$. The expectation $E(\boldsymbol{\theta}_{di})$ is then estimated by the average $\bar{\boldsymbol{\theta}}_d = n^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{di}$.

The covariance matrix of $\tilde{\boldsymbol{\theta}}_{di}$ is

$$\mathrm{Cov}(\tilde{\boldsymbol{\theta}}_d) = E([\tilde{\boldsymbol{\theta}}_{di} - E(\boldsymbol{\theta}_{di})][\tilde{\boldsymbol{\theta}}_{di} - E(\boldsymbol{\theta}_{di})]').$$

The corresponding estimate is

$$\widehat{\mathrm{Cov}}(\tilde{\boldsymbol{\theta}}_d) = n^{-1}\sum_{i=1}^{n}[\hat{\boldsymbol{\theta}}_{di} - \bar{\boldsymbol{\theta}}_d][\hat{\boldsymbol{\theta}}_{di} - \bar{\boldsymbol{\theta}}_d]'.$$

The conditional covariance matrix $\widetilde{\mathrm{Cov}}_{di}(\boldsymbol{\theta}|\mathbf{X})$ of $\boldsymbol{\theta}_{di}$ given $\mathbf{X}_i$ may be used to assess the accuracy with which the data $\mathbf{X}_i$ determine $\boldsymbol{\theta}_{di}$. One has $\widetilde{\mathrm{Cov}}_i(\boldsymbol{\theta}_d|\mathbf{X}) = \mathrm{Cov}_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i; \boldsymbol{\beta}_d)$, where

$$\mathrm{Cov}_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x}; \boldsymbol{\beta}_d) = \sum_{\boldsymbol{\omega}\in\Omega}[\boldsymbol{\omega} - E_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x}; \boldsymbol{\beta}_d)][\boldsymbol{\omega} - E_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{x}; \boldsymbol{\beta})]' p_{d\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\omega}|\mathbf{x}; \boldsymbol{\beta}_d).$$

The estimate of $\widetilde{\mathrm{Cov}}_i(\boldsymbol{\theta}_d|\mathbf{X})$ is then $\widehat{\mathrm{Cov}}_i(\boldsymbol{\theta}_d|\mathbf{X}) = \mathrm{Cov}_{d\boldsymbol{\theta}|\mathbf{X}}(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_d)$. The expected conditional covariance matrix $E(\widehat{\mathrm{Cov}}_{di}(\boldsymbol{\theta}_d|\mathbf{X}))$ is then estimated by

$$\overline{\mathrm{Cov}}(\boldsymbol{\theta}_d|\mathbf{X}) = n^{-1}\sum_{i=1}^{n}\widehat{\mathrm{Cov}}_i(\boldsymbol{\theta}_d|\mathbf{X}).$$

For any nonzero r-dimensional vector $\mathbf{c}$, the reliability of $\mathbf{c}'\tilde{\boldsymbol{\theta}}_{di}$ is

$$\rho_d^2(\mathbf{c}) = \frac{\mathbf{c}'\,\mathrm{Cov}(\tilde{\boldsymbol{\theta}}_d)\mathbf{c}}{\mathbf{c}'E(\widehat{\mathrm{Cov}}_i(\boldsymbol{\theta}_d|\mathbf{X}))\mathbf{c} + \mathbf{c}'\,\mathrm{Cov}(\tilde{\boldsymbol{\theta}}_d)\mathbf{c}}.$$

As in the multivariate normal case, the reliability of $\mathbf{c}'\hat{\boldsymbol{\theta}}_{di}$ is approximately the same in large samples, and the estimated reliability is then

$$\hat{\rho}_d^2(\mathbf{c}) = \frac{\mathbf{c}'\widehat{\mathrm{Cov}}(\tilde{\boldsymbol{\theta}}_d)\mathbf{c}}{\mathbf{c}'\overline{\mathrm{Cov}}(\boldsymbol{\theta}_d|\mathbf{X})\mathbf{c} + \mathbf{c}'\widehat{\mathrm{Cov}}(\tilde{\boldsymbol{\theta}}_d)\mathbf{c}}.$$

4 Application to Praxis Data

To illustrate results, data from a Praxis examination were examined. The examination is a multiple-choice right-scored test of content knowledge for certification for elementary school teachers. The test includes 120 items divided into four sections of 30 items apiece. Sections measure knowledge of language arts, mathematics, social studies, and science. For the particular administration studied, two items were not used in scoring due to unsatisfactory performance, one

Table 1
Estimated Expected Log Penalties per Item for Unidimensional Models

Model	Latent variable	Estimated log penalty	Akaike measure	Gilula-Haberman measure
Independent		0.54539	0.54555	0.54555
1PL	Normal	0.50785	0.50811	0.50811
2PL	Normal	0.50115	0.50147	0.50148
3PL	Normal	0.50034	0.50083	
2PL	Polytomous	0.50121	0.50153	0.50154

from the section on language arts and one from the section on social studies. As a consequence, 29 items are used for language arts, 30 for mathematics, 29 for social studies, and 30 for science. Analysis included 6,168 examinees.

Preliminary analysis of the data was based on one scale with 118 items. A summary of results can be found in Table 1. In this analysis, the univariate normal ability distribution was used with a 1PL, 2PL, and 3PL model to obtain a basic perspective on estimated expected log penalties per item. Adaptive quadrature used 9 points. For comparison, a 2PL model was also used with nine ability levels. These levels were the integers -4 to 4, and (3) was used for the ability distribution. A model that assumed that the X_{ij} were all independent was also considered to establish a further baseline. The Gilula-Haberman measure was omitted for the 3PL case due to problems with parameter identification for this model (the Hessian matrix was nearly singular, so that the correction was not approximated in a satisfactory manner).

The preliminary analysis suggests that, relative to the independence model, the normal 1PL model represents an improvement of about 6.86% in the Akaike or Gilula-Haberman measures. The gain from the normal 2PL model is modest, for the improvement over the normal 1PL model is only about 1.30% for these two measures. The gain from the normal 2PL to the normal 3PL is very small, only 0.13% for the Akaike measure. The polytomous 2PL case studied is comparable to the normal 2PL model, for the loss in terms of the Akaike or Gilula-Haberman criterion is only 0.01%.

The choice of 9 points for the polytomous model is not of unusual significance. Use of 7 evenly spaced points rather than 9 in the polytomous case defined by (3) only increases the Akaike and Gilula-Haberman measures by 0.018%. In the other direction, a polytomous model with 11 evenly spaced points defined by (3) leads to Akaike and Gilula-Haberman measures only 0.004% greater than in the normal case.

Use of (4) rather than (3) had relatively limited impact. Nonetheless, it is interesting to note that a model for 9 points that used linear, quadratic, cubic, and quartic components yielded essentially the same Gilula-Haberman measure as did the normal 2PL model and an Akaike measure that was about 0.002% smaller than for the normal 2PL model. Nonetheless, models not based on (3) generally involve somewhat more computational labor than do those based on (3).

For normal models, the choice of the number of quadrature points had relatively little influence. Use of 5 or 7 quadrature points had virtually no effect. Even for 3 quadrature points, the Gilula-Haberman and Akaike criteria increased by only about 0.002% relative to those for 9-point quadrature. Relative to 9-point adaptive quadrature, the extreme case of 2 points only increased the Gilula-Haberman and Akaike criteria by about 0.008%.

Four-dimensional 2PL analysis was then considered for multivariate normal and for polytomous cases. Results are provided in Table 2. The normal case reported used 4 points for each dimension, so that $256 = 4^4$ four-dimensional vectors were involved in the required multidimensional quadratures. Essentially the same results can be obtained for other selections of numbers of points per dimension such as 6 points for the first dimension and 3 points for the remaining three dimensions, so that 162 four-dimensional vectors are required per quadrature. Increases in Akaike and Gilula-Haberman measures of about 0.004% are observed with 3 points per dimension.

A variety of multidimensional polytomous models were explored. A base model used 4 evenly spaced points for each dimension and all possible combinations of these points with a log linear model defined by (3). A second model used 5 evenly spaced points from -2 to 2 with a log linear model defined by (3); however, vectors were excluded whenever the difference between any two coordinates exceeded 2. Thus $(-2, -1, 0, -1)$ was in Ω, but $(-2, -1, 1, -1)$ was not in Ω. In all, Ω contained 211 points. The third polytomous model used 6 evenly spaced points from -5 to 5 and a log linear model defined by (3), but vectors were excluded if any two coordinates differed by more than 4, so that Ω contained 276 points. Thus $(-5, -3, -1, -3)$ was in Ω but $(-5, -3, 1, -3)$ was not in Ω. The last model used 7 evenly spaced points from -3 to 3, and vectors were excluded if any two coordinates differed by more than 4. Thus Ω contained 341 points.

In this example, some gain is achieved by use of a multidimensional analysis. In the normal

Table 2
Estimated Expected Log Penalties per Item for Four-Dimensional 2PL Models

Latent variable	Latent classes per variable	Estimated log penalty	Akaike measure	Gilula-Haberman measure
Multivariate normal		0.49856	0.49889	0.49890
Polytomous	4	0.49947	0.49981	0.49982
Polytomous	5	0.49888	0.49922	0.49923
Polytomous	6	0.49871	0.49905	0.49906
Polytomous	7	0.49861	0.49895	0.49895

case, the four-dimensional model results in a reduction of the Akaike or Gilula-Haberman criterion by 0.514% relative to the one-dimensional model. This percentage change is much smaller than the change from a one-dimensional normal 1PL model to a one-dimensional normal 2PL model, but it is far larger than the change from a one-dimensional normal 2PL model to a one-dimensional normal 3PL model or from a one-dimensional polytomous 2PL model with nine latent classes with probabilities satisfying (3) to a one-dimensional normal 2PL model.

Differences between the normal case and the polytomous case are rather modest, although some details are worth considering. In all cases in Table 2, the normal case is more successful; however, the 7-point model increases the Akaike and Gilula-Haberman criteria by only 0.010 to 0.012%. Even for the 4-point example, the increase in the two criteria is only 0.184%. Use of more general log linear models than the model defined by (3) had little effect. At least for the data under study, model choice is likely to depend on the amount of computation regarded as tolerable and on considerations related to interpretation of test results.

Results are also rather similar in terms of estimated information on $\boldsymbol{\theta}$ and in terms of reliability coefficients for estimated ability coordinates. Table 3 provides a summary of estimates of the information concerning $\boldsymbol{\theta}_i$ provided by $\mathbf{X}_i$ for the models considered. On the whole, the estimates are quite similar, but again the multivariate normal case provides the best result, and the polytomous case with 7 points per dimension has an estimate that is about 1.24% smaller. For comparison, note that the estimated information for θ_i for a one-dimensional normal 2PL model is 1.35963, so that the gain in the four-dimensional case is quite clear.

The estimated reliability coefficients for the four coordinates of $\boldsymbol{\theta}_i$ are quite similar for all 2PL models. Consider Table 4. The composite listed is the sum of the coordinates. For comparison,

Table 3
Estimated Information on $\boldsymbol{\theta}_i$ Provided by $\mathbf{X}_i$

Latent variables	Latent classes per variable	Estimated information
Multivariate normal		2.14457
Polytomous	4	1.99651
Polytomous	5	2.07177
Polytomous	6	2.09598
Polytomous	7	2.11788

Table 4
Estimated Reliability Coefficients
for Ability Estimates for Four-Dimensional 2PL Models

Latent variables	Latent classes per variable	Language arts	Mathematics	Social studies	Science	Composite
Normal		0.86892	0.87644	0.83767	0.88038	0.92495
Polytomous	4	0.86581	0.88014	0.83331	0.87410	0.93046
Polytomous	5	0.86752	0.88098	0.83430	0.97704	0.92986
Polytomous	6	0.86807	0.87820	0.83541	0.88017	0.92903
Polytomous	7	0.86860	0.87940	0.83711	0.87915	0.92852

note that the reliability estimate for the one-dimensional normal case is 0.92620, and the estimated Cronbach alpha for the sum of the 118 item scores is 0.92408. The score sums for the individual sections have respective estimated Cronbach alpha statistics of 0.77088 for language arts, 0.84378 for mathematics, 0.71289 for social studies, and 0.77074 for science. The improved estimated reliability for estimated conditional means of coordinates of $\boldsymbol{\theta}_i$ reflects exploitation of correlations between section scores. Results are rather similar, albeit slightly better, than the proportional reduction in mean-squared error achieved by use of the observed section score sum and observed total test score to predict the true section score (Haberman, 2005b). For these data, the estimated proportional reductions are 0.85757 for language arts, 0.87370 for mathematics, 0.81210 for social studies, and 0.86742 for science.

Estimated model parameters for the multivariate normal and polytomous cases are quite closely linked, although some care must be taken to treat differences in scaling of variables. This issue is especially significant when estimated item discriminations are studied. Conditional on the test component, the sample correlation of estimated item discriminations for any pair of

Table 5
Estimated Correlations of Ability Coordinates

Latent var.	Classes per var.	LA by M	LA by SS	LA by S	M by SS	M by S	SS by S
Normal		0.84538	0.83086	0.88139	0.72643	0.82708	0.89262
Polytomous	4	0.86155	0.79247	0.84659	0.69808	0.80617	0.81064
Polytomous	5	0.82869	0.81657	0.86311	0.71272	0.81040	0.87563
Polytomous	6	0.86808	0.86282	0.88860	0.81620	0.85890	0.90792
Polytomous	7	0.84035	0.82818	0.87485	0.72482	0.82361	0.88729

Note. LA = language arts, M = mathematics, SS = social studies, S science.

models is never less than 0.99880 and for the normal and 7-point polytomous cases, the sample correlation is at least 0.99994; however, sample means of item discriminations for the different models are quite different. The polytomous models with 4 or 6 points per dimension have item discriminations roughly half of those in the normal case, while the polytomous models with 5 or 7 points per dimension have item discriminations somewhat larger than for the other polytomous cases but appreciably smaller than in the normal case. In the case of the item intercepts, the sample correlations for a specific test are all at least 0.99958, and at least 0.99999 for the normal and 7-point polytomous pair. Here effects of scaling are much smaller, especially if the 4-point polytomous case is excluded.

Comparison of estimated distributions of $\boldsymbol{\theta}_i$ is more complex given the problem of scaling; however, it is worth note that estimated correlations of coordinates of $\boldsymbol{\theta}_i$ are somewhat similar for the various models, but they do not agree very precisely. Consider Table 5. The 7-class case exhibits particularly good agreement with normal results. The correlations are quite high, although mathematics and social studies are less highly correlated than are other pairs of disciplines.

5 Conclusions

The example suggests that either a multivariate normal or a polytomous ability distribution can be used to achieve rather similar results for 2PL models for multidimensional item response analysis. Either the stabilized Newton-Raphson methods or the EM algorithm may be employed in computations. In this example, the multivariate normal ability distribution generally had a slight advantage; however, the difference was remarkably modest. Client preferences could influence any decision concerning which model to use. It is possible that other examples will arise in which differences between approaches have more substantial consequences.

Computational burden for analysis appears acceptable, although many details of calculation would best be modified for much larger samples. It would probably be advisable to begin calculations with a few hundred or few thousand observations to establish good approximations to maximum-likelihood estimates. The approximations would then be used to complete computations with the full sample. When computational labor is a major issue, then it is likely that the use of adaptive quadrature in the multivariate normal case with only 2 or 3 points per dimension will be the most attractive option.

The use of multidimensional item response models to generate subscores is quite feasible, as evident from the example. Given the similarity in results to those based on classical test theory, client preferences may again be a significant consideration.

The example used in the analysis was selected because the skills assessed were not closely linked. It should be emphasized that multidimensional item response analysis is not likely to reveal anything useful when skills are very tightly linked. Indeed, the estimation of the ability distribution will become increasingly challenging as correlations of ability coordinates approach 1.

The techniques used in this report can be applied quite readily to multidimensional versions of generalized partial credit models and to cases in which covariates are present or in which not all items are presented to each examinees (Xu & von Davier, 2006); however, these generalizations have not yet been fully implemented for all cases considered in this report.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

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

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187.

Haberman, S. J. (1974a). *The analysis of frequency data*. Chicago: University of Chicago Press.

Haberman, S. J. (1974b). Log-linear models for frequency tables with ordered classifications. *Biometrics*, *30*, 589–600.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, *18*, 193–211.

Haberman, S. J. (2005a). *Latent-class item response models* (ETS Research Rep. No. RR-05-28). Princeton, NJ: ETS.

Haberman, S. J. (2005b). *When can subscores have value?* (ETS Research Rep. No. RR-05-08). Princeton, NJ: ETS.

Haberman, S. J. (2006). *Adaptive quadrature for item response models* (ETS Research Rep. No. RR-06-29). Princeton, NJ: ETS.

Heinen, T. (1996). *Latent class and discrete latent trait models, similarities and differences*. Thousand Oaks, CA: Sage Publications.

Rao, C. R. (1973). *Linear statistical inference and its applications* (second ed.). New York: John Wiley.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–642). Amsterdam: North-Holland.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BGROUP program to higher dimensions* (ETS Research Rep. No. RR-05-27). Princeton, NJ: ETS.

Stewart, G. W. (1973). *Introduction to matrix computations.* New York: Academic Press.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Rep. No. RR-08-06). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2007, April). *Fitting structured general diagnostic models.* Paper presented at the annual meeting of the National Council on Education in Measurement, Chicago, IL.

Zhang, J. (2004). *Comparison of unidimensional and multidimensional approaches to IRT parameter estimation* (ETS Research Rep. No. RR-04-44). Princeton, NJ: ETS.