PSYCHOMETRIKA— VOL. 70, NO. 3, 1–23
SEPTEMBER 2005
DOI: 10.1007/s11336-003-1141-x

HIGH-DIMENSIONAL MAXIMUM MARGINAL LIKELIHOOD ITEM FACTOR ANALYSIS BY ADAPTIVE QUADRATURE

STEPHEN SCHILLING

SCHOOL OF EDUCATION, UNIVERSITY OF MICHIGAN

R. DARRELL BOCK

CENTER FOR HEALTH STATISTICS, UNIVERSITY OF ILLINOIS AT CHICAGO

Although the Bock–Aitkin likelihood-based estimation method for factor analysis of dichotomous item response data has important advantages over classical analysis of item tetrachoric correlations, a serious limitation of the method is its reliance on fixed-point Gauss-Hermite (G-H) quadrature in the solution of the likelihood equations and likelihood-ratio tests. When the number of latent dimensions is large, computational considerations require that the number of quadrature points per dimension be few. But with large numbers of items, the dispersion of the likelihood, given the response pattern, becomes so small that the likelihood cannot be accurately evaluated with the sparse fixed points in the latent space. In this paper, we demonstrate that substantial improvement in accuracy can be obtained by adapting the quadrature points to the location and dispersion of the likelihood surfaces corresponding to each distinct pattern in the data. In particular, we show that adaptive G-H quadrature, combined with mean and covariance adjustments at each iteration of an EM algorithm, produces an accurate fast-converging solution with as few as two points per dimension. Evaluations of this method with simulated data are shown to yield accurate recovery of the generating factor loadings for models of up to eight dimensions. Unlike an earlier application of adaptive Gibbs sampling to this problem by Meng and Schilling, the simulations also confirm the validity of the present method in calculating likelihood-ratio chi-square statistics for determining the number of factors required in the model. Finally, we apply the method to a sample of real data from a test of teacher qualifications.

Key words: factor analysis, item response theory, latent variables, EM algorithm, marginal likelihood estimation, GLS estimation, adaptive quadrature, monte carlo integration.

1. Introduction

Not long after Thurstone (1947) perfected the centroid method of multiple factor analysis, applications of the method to dichotomous item responses rather than test scores appeared in the literature (Guilford, 1941). It was soon demonstrated that the use of phi coefficients was unsatisfactory because of so-called "difficulty" factors encountered when the p-values deviated appreciably from one-half (Ferguson, 1941). Karl Pearson's tetrachoric correlation coefficients performed much better in this respect; in fact, Thurstone (1947) prepared nomographs for obtaining the coefficients from the proportions of the 2×2 pairwise joint occurrence tables, using them as a labor saving device when computing correlations between test scores (Thurstone and Thurstone, 1941). In that application the cut points can be chosen sufficiently near the median to keep all the marginal proportions near one-half, avoiding zeros in the frequency tables. That is not possible in item factor analysis and the presence of large variation in item difficulty often leads to zero joint frequencies, leaving the value of the coefficient undefined. This impeded applications of item factor analysis even after an efficient algorithm for computation of tetrachoric correlations was devised (Divgi, 1979).

Requests for reprints should be sent to Stephen Schilling, Assistant Professor, University of Michigan, School of Education, Ann Arbor, MI 48109, USA. E-mail: schillsg@umich.edu

© 2005 The Psychometric Society

The development of item factor analysis based on item response theory (IRT) and maximum marginal likelihood (MML) estimation of factor loadings was a marked improvement in this respect (Bock and Aitkin, 1981; Bock, Gibbons, and Muraki, 1987; Bartholomew and Knott, 1999). Direct evaluation of the marginal likelihood of the model parameters, given the observed item response patterns, yields robust estimation with item difficulties near one or zero, without computation of any pair-wise measure of association. It also makes full use of information in all orders of association among responses, not just joint occurrence, and provides a likelihood-ratio criterion of the statistical significance of additional factors in the model.

But the Bock and Aiken method of solving the marginal likelihood equations has limitations. Its reliance on numerical integration (quadrature) and a slowly converging first-order iterative solution limits its practical implementation to five factors and two or three quadrature points per dimension for more than three dimensions. Three-point solutions are satisfactory for relatively short tests, say less than 20 items, where each subject's individual contributions to the likelihood are sufficiently diffuse to have appreciable values for all points in the quadrature space. As Meng and Schilling (1996) have shown, Gaussian quadrature formulas on a small number of points per dimension become inaccurate with longer tests, where the likelihood of many patterns become so concentrated as to fall largely between the points. This is a serious limitation for item factor analysis because the most interesting applications are often those involving many items (see, for example, Lionelli, Chang, Bock, and Schilling, 2000).

Meng and Schilling (1996) showed that this problem can be avoided by locally adaptive Monte Carlo integration via a Gibbs sampler, but at the expense of another difficulty – random variation in successive evaluations of the marginal log-likelihood resulting from the necessity to draw a new Monte Carlo sample for each response pattern at each step of the iterative solution. This variation makes it impossible to obtain any definite values for a likelihood-ratio criterion for when to stop adding factors to the model. Variation in numerical estimates of the marginal log-likelihood also hampered earlier efforts at applying adaptive Gauss-Hermite (G-H) quadrature as Bock and Schilling (1997) often found that the marginal log-likelihood actually decreased during EM iterations.

This paper presents a solution to the problem of MML estimation using fixed-point quadrature with small numbers of points per dimension by applying Naylor-Smith (1982) adaptive quadrature; we avoid the problem with the Gibbs sampler by adopting a form of Monte Carlo integration based on linear transformations of a fixed set of points chosen by normal-distribution importance sampling (NIS). We then discuss restrictions that must be imposed on the solution to identify parameters and resolve indeterminacies of location, scale, and rotation. In contrast to Bock and Schilling (1997), we show that a simple adjustment based on the mean and covariance of the latent distribution computed at each EM iteration yields convergent solutions to the log-likelihood equations for models of any order. In an empirical section, we compare the accuracy of several methods using both likelihood-based and second-order approaches for recovering the generating parameters of the factor models and determining the number of factors required in the model. Finally, we present an application of the methods to a sample of real data.

2. Methods

All of the methods assume a conventionally scaled multiple factor model in d-dimensions for item j, $j = 1, 2, \ldots, n$:

$$y_j = \sum_k^d \alpha_{jk}\theta_k + \epsilon_j. \tag{1}$$

The latent variable, θ_k, is assumed independent normal with mean 0 and variance 1, and ϵ_j, independent normal with mean 0 and variance

$$\sigma_j^2 = 1 - \sum_k^d \alpha_{jk}^2.$$

Then y_j is normal with mean 0 and variance 1, and the correlation matrix of y is sufficient for estimating the factor loadings.

For classical analysis of data under this model, we employed a beta version of the TESTFACT 3.0 program. It computes tetrachoric correlations using Divgi's method and performs principal factor analysis by MINRES communality iteration (Harman, 1987). See Bock, Gibbons, Muraki, Schilling, Wilson, and Wood (1999) for details.

For IRT analysis we take the same approach as Bock and Aitkin (1981), letting $u = [u_j]$, $j = 1, 2, \ldots, n$ be a pattern of item scores, $u_j = 1$ if correct, and 0 otherwise. We assume a normal ogive response function, $P(u_j = 1|\theta, v_j) = P_j(\theta) = \Phi_j(\theta)$ and $P(u_j = 0|\theta, v_j) = 1 - P_j(\theta)$, where $\Phi_j(\theta)$ is the standard normal distribution function, the argument of which, for item j, is a function of θ, and an $(d+1)$-vector parameter v_j, $d < n$. We then set $\Phi_j(\theta) = \Phi[z_j(\theta)]$, where

$$z_j(\theta) = (\gamma_j - \sum_k^d \alpha_{jk}\theta_k)/\sigma_j \tag{2}$$

$$= c_j - \sum_k^d a_{jk}\theta_k; \tag{3}$$

that is, the underlying variable is rescaled so that $\sigma_j = 1$. In this scale $P_j(\theta)$ is a multiple latent variable probit model. The parameter γ_j is the item difficulty in standardized form, i.e., the normal deviate corresponding to the item percent-correct in the population. The parameter c_j is referred to as an item *intercept* and the a_{jk}, as item *slopes*. Computations of the MML analysis are carried out with intercepts and slopes, but for purposes of comparing the several methods the resulting slopes are rescaled to factor loadings,

$$\alpha_{jk} = a_{jk}/\sqrt{1 + \Sigma_h^d a_{jh}^2}.$$

In the analysis, so-called "Heywood" cases, in which $\sigma_j = 0$, are avoided by imposing the constraint $\sum_k^d a_{jk}^2 < 1$ on the estimator. For this purpose, Bock, Gibbons, and Muraki (1987) employ a stochastic constraint based on the beta distribution. This constraint was not needed for any of the examples in this paper.

In what follows, we begin by describing the Bock–Aitken and Meng–Schilling MML estimation approaches. Then we describe our approach using both adaptive G-H and Monte Carlo normal importance sampling. Finally, we describe a critical component of the adaptive approach: adjusting the parameter estimates at each EM iteration for provision estimates of the latent distribution mean and covariance.

2.1. MML Estimation using Non-adaptive G-H Quadrature

If the number of respondents, N, is large relative to the number of items, it is advantageous to sort the response patterns and count the number of occurrences, r_ℓ, of distinct patterns, u_ℓ,

$\ell = 1, \ldots, s$, where $N = \sum_\ell^s r_\ell$. The r_ℓ are then the sufficient statistics for estimating the total parameter set, say ν. The kernel of the likelihood is

$$L(\nu) = \prod_{\ell=1}^{s} \bar{P}_\ell^{r_\ell}, \tag{4}$$

where for $-\infty < \theta < \infty$,

$$\bar{P}_\ell = \int L_\ell(\theta) g(\theta) \mathrm{d}\theta, \tag{5}$$

and

$$L_\ell(\theta) = \prod_{j}^{n} [P_j(\theta)]^{u_{\ell j}} [1 - P_j(\theta)]^{1-u_{\ell j}}. \tag{6}$$

is the likelihood of θ, given u_ℓ. The vector latent variable θ is assumed distributed in the population of respondents with d-variate standard normal probability density $g(\theta)$. Note that for fixed item parameters, $L_\ell(\theta)g(\theta)$ in (5) is the unnormalized posterior density of θ for response pattern u_ℓ. Conditional independence of the item responses, given θ, is necessarily assumed in these results.

Differentiating under the integral sign, reversing the order of integration and summation, and replacing the multiple integral with multiple quadrature, we have as the marginal likelihood equation for the parameter vector, ν_j, of item j

$$\frac{\partial \log L(\nu)}{\partial \nu_j} \simeq \sum_{q_d}^{Q} \cdots \sum_{q_1}^{Q} \frac{\bar{r}_{j,q_1 \ldots q_d} - \bar{N}_{q_1 \ldots q_d} \Phi_j(\mathbf{X}_{q_1 \ldots q_d})}{\Phi_j(\mathbf{X}_{q_1 \ldots q_d})[1 - \Phi_j(\mathbf{X}_{q_1 \ldots q_d})]} \left[\frac{\partial \Phi_j(\mathbf{X}_{q_1 \ldots q_d})}{\partial \nu_j} W(X_{q_1}) \cdots W(X_{q_d}) \right], \tag{7}$$

where

$$\bar{r}_{j,q_1 \ldots q_d} = \sum_{\ell}^{s} r_\ell u_{\ell j} L_\ell(X_{q_1 \ldots q_d}) / \bar{P}_\ell \tag{8}$$

estimates the expected number of respondents answering item j correctly at each quadrature point,

$$\bar{N}_{q_1 \ldots q_d} = \sum_{\ell}^{s} r_\ell L_\ell(X_{q_1 \ldots q_d}) / \bar{P}_\ell \tag{9}$$

estimates the expected number of respondents at each point (that is, an unnormalized density of the d-dimensional latent distribution), and $W(X_{q_1}) \cdots W(X_{q_d})$ are the weights corresponding to the points of the multiple G-H quadrature.

Because Bock and Aitkin (1981) deemed second-order solution of these equations using either Newton-Raphson or Fisher scoring impractical when the number of items is large, they resorted to an EM solution in which (8) and (9) constitute each E-step and a d-variable Fisher-scoring probit analysis constitutes each M-step. This is the fixed-point non-adaptive method of the present study as implemented in TESTFACT 3.0 and earlier versions.

2.2. MML Estimation using Monte Carlo Integration: Gibbs Sampling

Meng and Schilling (1996) approached the estimation of item factor loadings from a different perspective. Considering θ as missing data, they applied the EM algorithm directly, giving for E-step t:

$$E(\log L_\ell(v_j \mid \nu^{(t)}) = \sum_{\ell}^{s} r_\ell \int \log L_\ell(v_j \mid \theta) f_\ell(\theta \mid \nu^{(t)}) \, \mathrm{d}\theta \ , \tag{10}$$

where

$$L_\ell(v_j \mid \theta) = [P_j(\theta)]^{u_{\ell j}} [1 - P_j(\theta)]^{1-u_{\ell j}}$$

and $f_\ell(\theta \mid \nu^{(t)})$ is the normalized posterior distribution of θ for response pattern u_ℓ at the current value of the total parameter set $\nu^{(t)}$. The M-step consists of maximizing the n equations of (10).

Meng and Schilling (1996) addressed the integration problem by drawing K points from the posterior distributions in (10). Although making independent draws from these distributions is difficult, they showed that a Gibbs sampler could be used to draw correlated samples from the posterior distribution of θ, given $u_{j\ell}$ and the provisional parameters $\nu^{(t)}$. Because the Gibbs sampler mixed very fast in this application, the autocorrelations decayed quickly and nearly independent samples could be achieved by retaining every fifth draw. The E-step is given by

$$E(\log L(v_j \mid \nu^{(t)})) \approx \sum_{\ell=1}^{s} r_\ell \frac{1}{K} \sum_{k=1}^{K} \log L_\ell(v_j \mid \theta_{\ell,k}) \ , \tag{11}$$

where the $\theta_{\ell,k}$'s are the K-retained draws. These draws act as the observed data for the M-step probit analysis. This result is an application of what Wei and Tanner (1990) refer to as a Monte Carlo EM (MCEM) algorithm.

2.3. MML Estimation using Adaptive Quadrature

The key to implementing MML estimation using adaptive quadrature is recognizing that the dominant terms in the integrals in both Equations (5) and (10) are the individual posterior densities of θ for the response patterns u_ℓ. Then if the posteriors are approximately normal the adaptive quadrature approach detailed in Naylor and Smith (1982) can be effectively applied to approximate each of these individual integrals.

Specifically, Naylor and Smith (1982) proposed a numerical approach to Bayes estimation along the following lines. Consider the expectation of a function with respect to a probability distribution,

$$\mathcal{E}[h(\boldsymbol{x})] = \int h(\boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{12}$$

This may be written as

$$\mathcal{E}(h(\boldsymbol{x})) = \int \frac{h(\boldsymbol{x}) p(\boldsymbol{x})}{g(\boldsymbol{x}, \mu, \Sigma)} g(\boldsymbol{x}, \mu, \Sigma) \mathrm{d}\boldsymbol{x} = \int f(\boldsymbol{x}) g(\boldsymbol{x}, \mu, \Sigma) \, \mathrm{d}\boldsymbol{x}, \tag{13}$$

where $g(\boldsymbol{x}, \mu, \Sigma)$ is a multivariate normal density. Then applying the multivariable change of variable formula by setting

$$\boldsymbol{x}^* = \boldsymbol{T}^{-1}(\boldsymbol{x} - \mu),$$

where $\boldsymbol{T}$ is the (lower triangular) Cholesky factor of Σ, we have

$$\mathcal{E}(h(\boldsymbol{x})) = |\boldsymbol{T}| \int f(\boldsymbol{T}\boldsymbol{x}^* + \mu) g(\boldsymbol{x}^*, 0, I) \, \mathrm{d}\boldsymbol{x}^*, \tag{14}$$

a form amenable to product G-H quadrature rules. The integral in (14) can be approximated by

$$|\boldsymbol{T}| \int f(\boldsymbol{T}\boldsymbol{x}^* + \mu) g(x_1^*), \ldots, g(x_d^*)\, \mathrm{d}x_1^*, \ldots, \mathrm{d}x_d^* \simeq |\boldsymbol{T}| \sum_{i_d=1}^{q} W_{i_d} \ldots \sum_{i_1=1}^{q} W_{i_1} f(\boldsymbol{T}\boldsymbol{X}_{i_1,\ldots,i_d} + \mu), \tag{15}$$

where $\boldsymbol{X}_{1,\ldots,d} = [X_{i_1}, \ldots, X_{i_d}]$ is a point in quadrature space, X_{i_k} is a G-H quadrature point, and W_{i_k} is the corresponding weight. If $f(\boldsymbol{T}\boldsymbol{x}^* + \mu)$ can be well approximated by a polynomial of order $2q - 1$ in each orthogonal direction, the quadrature in (11) will be highly accurate because it is exact for an order $2q - 1$ polynomial. Note that $f(\boldsymbol{T}\boldsymbol{x}^* + \mu)$ will be closely approximated by a $2q - 1$ order polynomial if and only if the original integrand is closely approximated by the product of a multivariate normal and a $2q - 1$ polynomial.

The key issue for implementation is finding a multivariate normal distribution closely matching the original density. The change of variable formula adapts the product G-H quadrature rules to $h(\boldsymbol{x})p(\boldsymbol{x})$ divided by the matching multivariate normal density at each point. Naylor and Smith (1982) presented an iterative orthogonalization procedure for matching in a full Bayesian context. Specifically, they used (15) to compute the first and second moments of the distribution to be matched and used as the matching distribution the multivariate normal with the same moments. They then estimated improved estimates of the first and of the second moments and repeated the process until the moments converged. The drawback of this approach is the large number of function evaluations needed for convergence. A less costly approach is to estimate the mode $\tilde{\mu}$ and the inverse information at the mode $I^{-1}(\tilde{\mu})$ and use these as the estimates of μ and Σ (see Liu and Pierce, 1994); this is the approach we implement throughout this paper.

Because of the asymptotic normality (in n) of IRT posterior distributions of θ, we assume the definite integral in (5) to be well approximated by the product of a multivariate normal and a $2q - 1$ degree polynomial. In that case, the estimates of $\bar{P}_\ell$ are given by

$$\bar{P}_\ell \approx |\boldsymbol{T}| \sum_{i=1}^{q^d} \frac{L_\ell(\boldsymbol{T}X_i + \tilde{\mu}_\ell) g(\boldsymbol{T}X_i + \tilde{\mu}_\ell, 0, I)}{g(X_i, 0, I)} W(X_i), \tag{16}$$

where $\boldsymbol{T}$ is the Cholesky factor of the inverse information matrix at $\tilde{\mu}_\ell$, $W(X_i)$ is the product of the W_{i_k}'s over d dimensions, and the sum is over the entire set of q^d quadrature points. Similarly the estimates of the integrals in (10) are given by

$$\int \log L_\ell(\nu_j \mid \theta) f_\ell(\theta \mid \nu^{(t)})\, \mathrm{d}\theta \approx |\boldsymbol{T}| \sum_{i=1}^{q^d} \frac{\log L_\ell(\nu_j \mid \boldsymbol{T}X_i + \tilde{\mu}_\ell) f_\ell(\boldsymbol{T}X_i + \tilde{\mu}_\ell \mid \nu^{(t)})}{g(X_i, 0, I)} W(X_i), \tag{17}$$

where $L_\ell(\theta \mid \nu^{(t)})$ is $L_\ell(\theta)$ at the current value of the total parameter set $\nu^{(t)}$.

The question remains, "How many quadrature points are needed for effective implementation within the context of maximizing the marginal likelihood?". In this context, precise estimates of each integrand are not really necessary. Accuracy to within two or three significant digits is sufficient because of the large number of integrals summed in the E-step. There is, however, clearly a lower limit to the number of points needed. One-point G-H quadrature corresponds to joint maximum likelihood estimation with a $N(0, I)$ prior placed on the θ's. The well-known inconsistency of this method of estimation leads to failure of EM iterations to converge, especially in high dimensions. Similar problems can occur with 2, and 3 point G-H quadrature, because, unlike nonadaptive quadrature, the points no longer form a stable basis. In the Meng and Schilling (1996) approach, the random draws from the posterior obscures this fact; the EM iterates basically form a stochastic process about the converged values. An earlier implementation of adaptive

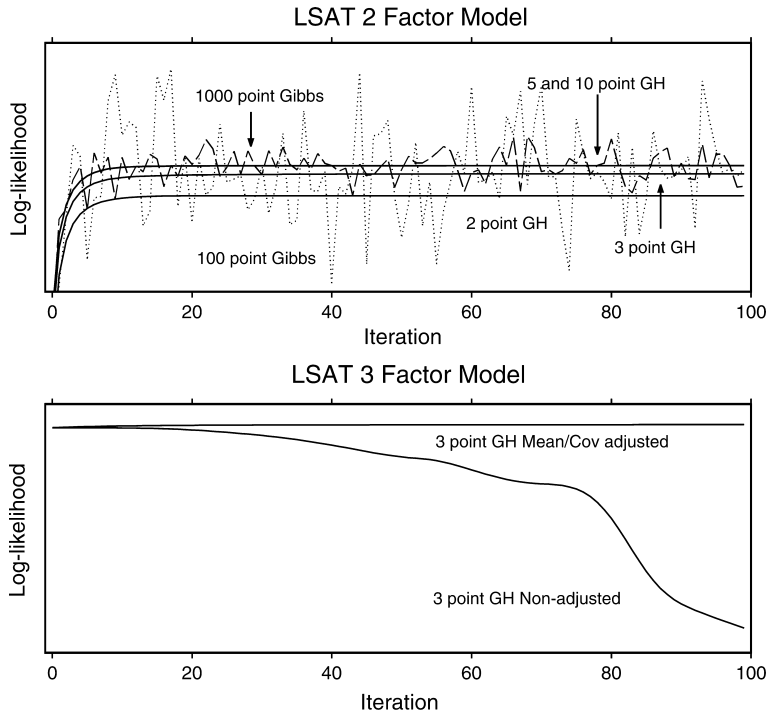

FIGURE 1.
The effect of integration methods on convergence – LSAT section 7 data.

quadrature in the context of the full-information item factor models (Bock and Schilling, 1997), fails to adequately address this issue. Therefore, Bock and Schilling (1997) were unable to obtain likelihood-ratio statistics for tests of models of higher order and often found a puzzling decrease in the log-likelihoods even when the EM iterates appeared to converge (for an example, see Figure 1).

In the next section we shall show that effective implementation of adaptive quadrature in the context of EM requires a stable basis at each iteration. This can be effectively accomplished by setting the mean and variance of the latent distribution to 0 and I respectively. In the context of EM, this can be viewed as an example of what Liu, Rubin, and Wu (1998) termed parameter extended EM (PX-EM).

2.4. Resolving Indeterminacy of Location, Scale and Direction in Multidimensional IRT Models

In unidimensional IRT, indeterminacy of location and scale during item parameter estimation is typically resolved by setting the mean of the latent distribution to 0 and the standard deviation to 1. This will happen automatically in principle if the quadrature points sum to 0 and the weights sum to 1. However, to allow for small deviations from these values due to computing approximations during the iterative solution of the likelihood equations, rescaling of the distribution and correcting the item parameters to yield exact values is advisable. These adjustments are implemented, for example, in the BILOG-MG program of Zimowski, Muraki, Mislevy, and Bock (1995) to fix the location and scale of the latent distribution and to speed convergence of the solution.

In *multidimensional* IRT, further restrictions are necessary to fix the covariance structure of the latent distribution. The usual choice is to set all covariances to zero, yielding a latent distribution with vector mean 0 and correlation matrix I. This should again be automatic if the multidimensional quadrature points are generated by a d-fold Kronecker product of the one-dimensional

points. But again it is advisable to impose this exact restriction on the distribution and parameter estimates by a linear transformation at each iteration. In fixed-point quadrature, provisional estimates of the mean and covariance matrix of the latent distribution are available for making these adjustments.

In adaptive quadrature, however, the required representation of the latent distribution as masses, computed at each E-step, on a specified set of points does not exist. In its place, we compute the mean and covariance matrix of provisional Bayes estimates of the *factor scores* corresponding to each response pattern in the data, and use these means and covariances to compute the mean and covariance of the latent distribution. If the number of quadrature points per dimension is eight or more, the resulting adjustments are usually negligible. But with the small number of points required for practical high-dimensional analysis, the adjustments are essential for efficient estimation of the item parameters and accurate computation of the marginal likelihood.

Interestingly, Liu, Rubin, and Wu (1998) found that similar adjustments during the EM solutions for multilevel models speed convergence even when the model is linear and quadrature is not involved. They refer to this procedure as "parameter expansion" and provide a detailed rationale for the use of these adjustments within EM. Except for slightly different notation, we use their description of the computing steps as follows:

During the E-step, approximate by quadrature the moments,

$$S_\theta = \sum_{\ell=1}^{s} \frac{r_\ell}{\bar{P}_\ell} \int \theta L_\ell(\theta)\phi(\theta)\,\mathrm{d}\theta \tag{18}$$

$$S_{\theta\theta'} = \sum_{\ell=1}^{s} \frac{r_\ell}{\bar{P}_\ell} \int \theta\theta' L_\ell(\theta)\phi(\theta)\,\mathrm{d}\theta\ ; \tag{19}$$

then the provisional estimates of μ and Σ are given by

$$\hat{\mu}^{(t+1)} = S_\theta / N \tag{20}$$

$$\hat{\Sigma}^{(t+1)} = S_{\theta\theta'}/N - \hat{\mu}^{(t+1)}\hat{\mu}^{(t+1)\prime}, \tag{21}$$

Note, $\hat{\mu}^{(t+1)}$ is the average of the individual posterior means, i.e., the latent distribution mean – $\hat{\Sigma}^{(t+1)}$ is the average of the individual posterior covariance matrices plus the covariance matrix of the individual posterior means, i.e., the latent distribution covariance matrix. We can rescale θ so that $\theta \sim N(0, I)$. Then if $a_j^{*(t+1)}$ is the EM updated unadjusted row vector of item slopes and $c_j^{*(t+1)}$ the EM updated unadjusted item intercept, the adjusted parameter estimates can be obtained by applying the inverse of the rescaling transformation to the item slopes and intercept,

$$\begin{aligned} z^{(t+1)}(\theta^*) &= a_j^{*(t+1)}\theta^* + c_j^{*(t+1)} \\ &= a_j^{*(t+1)} L L^{-1}(\theta^* - \mu^{(t+1)} + \mu^{(t+1)}) + c_j^{*(t+1)} \\ &= a_j^{*(t+1)} L\theta + a_j^{*(t+1)}\mu^{(t+1)} + c_j^{*(t+1)}, \end{aligned}$$

yielding

$$a_j^{(t+1)} = a_j^{*(t+1)} L \tag{22}$$

$$c_j^{(t+1)} = c_j^{*(t+1)} + a_j^{*(t+1)}\mu^{(t+1)}, \tag{23}$$

where L is the (lower triangular) Cholesky factor of $\hat{\Sigma}^{(t+1)}$. A justification for this adjustment in general terms is given in Liu, Rubin, and Wu (1998); the justification in this specific case borrows from that discussion and proceeds as follows.

The full-information item factor model assumes that the true mean and covariance of the latent distribution of θ are 0 and I; their computed value for the expanded model will equal the nominal values only at the maximum likelihood estimate. The E-step of the EM algorithm is therefore effectively imputing missing data under the wrong model with $\nu^{(t)} \neq \hat{\nu}$. The M-step of EM ignores this, while the M-step of the parameter expanded EM uses the expanded parameters as covariates, adjusting ν for the difference between the nominal and computed values of μ and Σ by regressing ν on this difference. In essence, the estimated latent distribution mean and covariances are acting as covariates for the item parameters, providing extra information about the true MLE.

In Liu, Rubin, and Wu (1998) this adjustment was shown to have a dramatic effect on EM's speed of convergence. But here this adjustment not only speeds convergence, it also stabilizes the computation of the log-likelihood so that convergence proceeds smoothly during the EM cycles. Without this adjustment five or more points per dimension are sometimes needed to obtain a convergent solution – with the adjustment convergent solutions can be obtained with as few as 2 points per dimension. Moreover, as we show in Section 3.2, this adjustment is almost always necessary for obtaining a convergent solution if the data are over-fitted relative to the number of factors. Because it is impossible to determine beforehand the number of factors necessary in empirical applications, stability of the computed log-likelihood is critical in the statistical test of when to stop factoring.

2.5. Adaptive Monte Carlo Integration using a Normal Importance Sampling Distribution

As an alternative to adaptive G-H quadrature, Monte Carlo integration based on draws from a $N(0, I)$ distribution for each posterior distribution in (10) can be used to evaluate the E-step of the EM algorithm. Specifically the estimates for $\bar{P}_\ell$ are given by

$$\bar{P}_\ell \approx |\boldsymbol{T}| \sum_{i=1}^{N_p} \frac{L_\ell(\boldsymbol{T}X_i + \tilde{\mu}_\ell) g(\boldsymbol{T}X_i + \tilde{\mu}_\ell, 0, I)}{g(X_i, 0, I)}, \tag{24}$$

where the sum is over the N_p drawn from a $N(0, I)$ distribution. Similarly the estimates of the integrals in (10) are given by

$$\int \log L_\ell(\nu_j \mid \theta) f_\ell(\theta \mid \nu^{(t)}) \, \mathrm{d}\theta \approx \frac{1}{\bar{P}_\ell} |\boldsymbol{T}| \sum_{i=1}^{N_p} \frac{\log L_\ell(\nu_j \mid \boldsymbol{T}X_i + \tilde{\mu}_\ell) f_\ell(\boldsymbol{T}X_i + \tilde{\mu}_\ell \mid \nu^{(t)})}{g(X_i, 0, I)}, \tag{25}$$

where $L_\ell(\theta \mid \nu^{(t)})$ is $L_\ell(\theta)$ at the current value of the total parameter set $\nu^{(t)}$.

Adapting the draws from a $N(0, I)$ to the mode $\tilde{\mu}_\ell$ and corresponding inverse information $I^{-1}(\tilde{\mu}_\ell)$ is identical to using a $N(\tilde{\mu}_\ell, I^{-1}(\tilde{\mu}_\ell))$ importance sampling distribution (see Ripley, 1987) and is similar to adapting the G-H points. The only difference is the absence of the $W(X_i)$ in the above equations. A single set of draws from a $N(0, I)$ is used for each posterior distribution throughout the iterations of the EM algorithm – only the modes $\tilde{\mu}_\ell$ and $I^{-1}(\tilde{\mu}_\ell)$'s are updated at each iteration. This avoids the random variation associated with the Meng-Schilling approach. Here again, the mean and covariance adjustment described in the previous section is critical in effective application of normal importance sampling within the context of MML. Our experience is that the adjustment has the same effect, speeding convergence, and increasing the accuracy of the calculated log-likelihoods.

2.6. Fixing Rotational Indeterminacy

In addition to scaling indeterminacies, there is also rotational indeterminacy of the multiple factor model. Because of this indeterminacy, any rank d basis of the matrix factor loadings may be chosen arbitrarily. We construct a basis by performing a MINRES principal factor solution on the tetrachoric correlation matrix (conditioned to be positive-definite) and transforming the loadings so that $\alpha_{jk} = 0, k > j$. The transformation for this purpose is computed by a modified Gram-Schmidt or Householder (1964) triangular orthogonalization of the leading d rows of the principle factor loadings, say $\boldsymbol{A}_{11}$. The result is the so-called **QL** decomposition

$$\boldsymbol{A}'_{\mathbf{11}} = \boldsymbol{QL}', \ |\boldsymbol{A}_{\mathbf{11}}| \neq 0, \tag{26}$$

where **L** is lower triangular and **Q** is orthogonal. Then

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{\mathbf{11}} \\ \boldsymbol{A}_{\mathbf{21}} \end{bmatrix}, \ \boldsymbol{AQ} = \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{A}_{\mathbf{21}}\boldsymbol{Q} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{B} \end{bmatrix}, \tag{27}$$

as required. We estimate B and the nonzero elements of L but later transform the restricted solution to a preferred basis, either the principle-factor or Kaiser's (1958) varimax pattern.

3. Empirical Investigation of Numerical Integration Procedures in Maximum Marginal Likelihood Estimation

In this section we evaluate the accuracy of competing methods in estimating the parameters of item factor models and in testing the statistical significance of additional factors included in the model. Apart from the Gibbs and GLS solutions, all calculations were performed with a beta version of the TESTFACT 3.0 program of Bock, Gibbons, Muraki, Schilling, Wilson, and Wood (1999). The adaptive quadrature routines are currently implemented in TESTFACF 4.0 (Wood, Wilson, Gibbons, Schilling, Muraki, and Bock, 2003).

3.1. Recovery of Generating Parameters from Simulated Data

Random values from uniform (0.0,1.0) distributions required in these simulations were generated by Schrage's (1979) algorithm. Random normal deviates were computed by the composition method of Ahrens and Dieter (1979). As described in Section 1, a normal ogive item response function with argument linear in the number of factors was assumed.

3.1.1. Five-factor Simulations

We simulated responses of 1250 subjects to three different five-factor tests with the following factor patterns:

Simulation 1. A 64-item test with a principal factor pattern in which the signs of the loadings were those of the general effect and four main effect contrasts in the basis matrix of a 2^6 factorial design (see Bock, 1975/1985). The absolute values in successive columns of loadings were 0.64, 0.48, 0.36, 0.28, and 0.24, with communality 0.9056. For computational purposes, these loadings were converted to factor slopes as in (3) with divisor $\sqrt{1 - 0.9056}$. The item intercepts in (3) were all set to zero, and the factor scores were drawn randomly from an $N(0, I)$ distribution.

Simulation 2. A 32-item test with a principal factor sign pattern corresponding the general and four main effect contrasts of a 2^5 factorial design. The absolute values of the factor loadings were drawn from uniform (0.5, 0.6), (0.4, 0.5), (0.3, 0.4), (0.2, 0.3), and (0.05, 0.15)

distributions and converted to slopes. The communalities ranged from 0.51 to 0.94 with a median value of 0.69. The intercepts were drawn randomly from a uniform $(-1.75, 1.75)$ distribution.

Simulation 3. A 30-item test with quasi-simple structure suitable for varimax rotation. Each of five successive blocks of items had partly overlapping positive factor loadings as shown in Table 1. All other loadings were zero. The communalities ranged from 0.53 to 0.91 with a median value of 0.67. The intercepts were drawn from the uniform $(-1.75, 1.75)$ distribution.

To assess the accuracy in recovering the generating factor loadings, we performed factor analyses of the simulated data by the following methods:

1. Meng and Schilling's (1996) Monte Carlo EM method using a 243-point Monte Carlo sample drawn from the posterior via Gibbs sampling (GS).
2. EM using 243-point Monte Carlo integration by importance sampling from an assumed multivariate normal posterior distribution with provisional mean and covariance matrix approximated by the Bayes mode and corresponding inverse information (Maximum A Posteriori estimation) during each EM cycle (NIS).
3. Adaptive G-H quadrature with five points per dimension (total of 3,125 points in total) (G-H5).
4. Adaptive G-H quadrature with three points per dimension (243 points) (G-H3).
5. Adaptive G-H quadrature with two points per dimension (32 points) (G-H2).
6. Nonadaptive G-H quadrature with three points per dimension (243 points).
7. Classical MINRES (i.e., unweighted least squares) item factor analysis of a conditioned positive-definite tetrachoric correlation matrix (see Bock, Gibbons, and Muraki, 1987).
8. Generalized least squares (GLS) (Muthén, 1984) as implemented in the Mplus program (Muthén and Muthén, 1998–2001). Here we used the diagonal weight option, which performed best of all GLS methods for the examples examined in this paper.

Accuracy was assessed by computing the root mean square error (RMSE) between the estimated and generated loadings for each factor. The item slopes were estimated in MML but

TABLE 1.
Five-factor simulation: generating factor loadings for simulation 3 (other loadings 0.0).

Items				
1 – 6 F1	5 – 12 F2	11 – 18 F3	17 – 24 29 – 30 F4	23 – 30 F5
0.8216	0.6340	0.5672	0.5715	0.6391
0.7728	0.6529	0.5318	0.5773	0.5886
0.8489	0.7279	0.7443	0.8434	0.8274
0.9005	0.8250	0.7878	0.8184	0.8110
0.6340	0.8914	0.7980	0.8203	0.9353
0.6529	0.7823	0.7308	0.8181	0.8799
	0.5672	0.5715	0.6391	0.6756
	0.5318	0.5773	0.5886	0.5685
			0.6756	
			0.5685	

TABLE 2.
Root-mean-square errors of recovering generating factor loadings: five-factor simulated Data N = 1,250.

Simulation	Factor	Gibbs	NIS	GH 5	GH 3	GH 2	NA 3	ULS	GLS
Principal	1	0.020	0.025	0.018	0.022	0.026	0.033	0.077	0.020
Factor	2	0.023	0.023	0.023	0.024	0.024	0.047	0.071	0.025
Rotation	3	0.031	0.033	0.031	0.031	0.032	0.063	0.060	0.034
Factorial	4	0.035	0.038	0.037	0.038	0.044	0.045	0.046	0.042
Design	5	0.031	0.032	0.033	0.034	0.038	0.052	0.045	0.040
	Mean	0.029	0.031	0.029	0.030	0.034	0.049	0.061	0.034
Principlal	1	0.042	0.042	0.042	0.042	0.045	0.051	0.058	0.057
Factor	2	0.045	0.046	0.045	0.045	0.045	0.051	0.096	0.072
Rotation	3	0.036	0.037	0.036	0.036	0.038	0.041	0.075	0.052
Random	4	0.040	0.041	0.039	0.040	0.043	0.043	0.051	0.049
Loadings	5	0.057	0.059	0.058	0.058	0.062	0.144	0.173	0.128
	Mean	0.045	0.046	0.045	0.045	0.047	0.077	0.101	0.077
Varimax	1	0.043	0.045	0.044	0.043	0.046	0.051	0.094	0.043
Factor	2	0.038	0.040	0.039	0.039	0.039	0.043	0.049	0.047
Rotation	3	0.045	0.045	0.045	0.046	0.047	0.044	0.051	0.046
Simple	4	0.050	0.052	0.050	0.049	0.047	0.043	0.067	0.047
Structure	5	0.031	0.031	0.030	0.031	0.032	0.037	0.081	0.061
	Mean	0.042	0.043	0.042	0.042	0.043	0.044	0.071	0.049

converted to loadings before computing the errors. However, the arbitrary basis of the estimated loadings had to be rotated to agree with that of the generating loadings to produce meaningful results. For Simulations 1 and 2 the MML basis was converted to the principal vectors of the correlation matrix constructed from the estimated loadings. In Simulation 3 the generating loadings shown in Table 1 were transformed orthogonally to varimax loadings for comparison with the similarly transformed estimated values.

The RMSEs given in Table 2 reveal that the accuracies of the five adaptive MML methods are very similar. Any differences can be attributed to the relative accuracies of the numerical integrations involved. The Monte Carlo integrations were each based on 243 draws from the posterior distribution of θ for each response pattern in the sample. Of the two, GS was slightly more accurate because it draws directly from the posterior, rather than the normal approximation to the posteror in NIS. Among the G-H methods, the accuracies at five-points per dimension quadratures were essentially the same as GS but were of course based on many more points in the θ space. More interesting, however, is that the RMSEs of the three-point and two-point quadratures were so closely comparable to the computationally more demanding methods – practically indistinguishable.

All the methods examined tended to exhibit larger RMSEs for the smaller factor loadings, as indicated by the larger RMSEs for the fifth as compared to first factors in Simulations 1 and 2. As expected, all the adaptive methods were more accurate than the nonadaptive method, except in Simulation 3. In short tests where posteriors are diffuse, there is little to be gained by adaptive quadrature. Simulation 3 not only had the fewest items (only 30), but the near simple structure meant that there were effectively only six or eight items measuring each factor.

With the sample sizes in these simulations (1,250), the MML methods clearly were more accurate than MINRES analysis of estimated tetrachoric correlations. As mentioned previously, principal factor analysis with communality iteration performs reasonably well when samples are large and the items are predominantly near 50% difficulty. This is the case in Simulation 1, where responses were generated from items with zero intercepts. However, even in that favorable cir-

TABLE 3.
Eight-factor simulation: generating quasi-simple structure factor loadings (other loadings 0.0).

Items							
1 – 6 F1	5 – 12 F2	11 – 18 F3	17 – 24 F4	23 – 30 F5	29 – 36 F6	35 – 42 47 – 48 F7	41 – 48 F8
0.885	0.560	0.665	0.687	0.550	0.590	0.566	0.555
0.854	0.578	0.632	0.589	0.600	0.489	0.558	0.676
0.863	0.839	0.851	0.865	0.850	0.889	0.791	0.807
0.817	0.936	0.793	0.915	0.808	0.837	0.861	0.843
0.560	0.813	0.794	0.879	0.904	0.807	0.794	0.874
0.578	0.786	0.881	0.823	0.807	0.801	0.789	0.904
	0.665	0.687	0.550	0.590	0.566	0.555	0.548
	0.632	0.589	0.600	0.489	0.558	0.676	0.606
						0.548	
						0.606	

cumstance, tetrachoric MINRES shows substantially larger RMSEs than the other methods. GLS performs better, with RMSEs essentially identical to those of adaptive MML. However, Simulations 2 and 3, with their randomly generated intercepts, lead to many more poorly conditioned estimates of the tetrachoric correlations; the RMSEs of the GLS solution were correspondingly greater for these tests. This was especially true for the fifth factor of Simulation 2, and the fifth factor in Simulation 3, where the RMSEs were about twice as large as those for any adaptive MML method. There is also evidence in these results that the accuracy improves with increasing proportions of common variance accounted for by the factors. This is seen in the average errors for the five factors given in Table 2; Simulation 1, with the most items, and thus small measurement error variance, has the largest average factor loadings and the smallest RMSEs.

3.1.2. Eight-factor Simulations

We evaluate here the accuracy of four of the above methods in recovering parameters of an eight-factor model from 1,250 simulated responses to the following tests:

Simulation 4. A 128-item test with eight principal factors and sign pattern of the general effect and seven main effect contrasts of a 2^7 factorial design. The absolute values of the loadings in successive columns were 0.55, 0.45, 0.36, 0.28, 0.21, 0.15, 0.10, and 0.06, for a communality of 0.7932. The intercepts were drawn from a uniform $(-1.75, 1.75)$ distribution.

Simulation 5. A 48-item test with eight quasi-simple structure factors; the intercepts and nonzero loadings are shown in Table 3. The communalities ranged from 0.48 to 0.94 with a median value of 0.70. The intercepts were drawn from a uniform $(-1.75, 1.75)$ distribution.

The simulated data were analyzed by methods 1, 2, 5, 7, and 8; however, method 5 had a total of 256 quadrature points rather than 32. The RMSE results are shown in Table 4. For both tests the accuracies of MML methods were virtually identical. The adaptive methods were more accurate than MINRES and GLS, although the gains were greater for Test 4 than Test 5, where the six and eight effective items per factor resulted in more diffuse posterior densities for each response pattern. GLS and adaptive MML were nearly equivalent in the simple structure simulation. Where there were differences, they were greatest for the proportionally small loadings of the later factors.

TABLE 4.
Root-mean-square errors of recovering generating factor loadings: eight-factor simulated data N = 1,250.

	Factor								
Method	1	2	3	4	5	6	7	8	Mean
	Simulation 1 – 128 Items – Random intercepts								
GS 243-point MCEM	0.034	0.041	0.042	0.039	0.039	0.042	0.052	0.075	0.047
IS 243-point MCEM	0.034	0.040	0.042	0.039	0.039	0.042	0.055	0.076	0.048
2-point adaptive G-H	0.034	0.041	0.043	0.039	0.039	0.042	0.053	0.076	0.048
Tetrachoric MINRES	0.101	0.090	0.072	0.064	0.107	0.190	0.178	0.145	0.126
GLS	0.038	0.044	0.052	0.056	0.064	0.124	0.114	0.119	0.083
	Simulation 2 – 48 Items – Simple Structure								
GS 243-point MCEM	0.042	0.043	0.039	0.047	0.045	0.047	0.041	0.045	0.044
IS 243-point MCEM	0.045	0.042	0.041	0.048	0.046	0.047	0.040	0.046	0.044
2-point adaptive G-H	0.042	0.044	0.040	0.049	0.046	0.049	0.043	0.047	0.045
Tetrachoric MINRES	0.048	0.081	0.045	0.053	0.090	0.048	0.069	0.051	0.063
GLS	0.046	0.049	0.042	0.052	0.051	0.047	0.046	0.052	0.048

3.2. Testing the Statistical Significance of an Added Factor

Although the number of items in the simulations are too large to allow a chi-square test of the number of factors based on response pattern frequencies as in Bock and Aitkin (1981), Haberman (1977) has shown that the difference of twice the log-likelihood ML analysis with the addition of q free parameters to the model is distributed in large samples as chi-square with q degrees of freedom. In the present context the difference chi-square is,

$$X^2 = 2\sum_{\ell=1}^{s} r_\ell \log \hat{P}_\ell - 2\sum_{\ell=1}^{s} r_\ell \log \hat{P}'_\ell, \tag{28}$$

where $\hat{P}_\ell$ is the estimated marginal probability of pattern ℓ under the smaller model and $\hat{P}'_\ell$ is the corresponding probability under the larger model. The degrees of freedom when increasing the number of factors for the orthogonal model from d to $d+1$ is $n-d$.

Numerical integration of the likelihood equations in MML estimation is much easier than obtaining the correct absolute value of the likelihood. This is because the integral appears both in the numerator and denominator, canceling any systematic bias in its values, similar to Tierney and Kadane (1986). With small numbers of items we can verify the accuracy of integration by checking that the calculated probabilities of all 2^n patterns sum to one. This was done using the LSAT-7 data of Bock and Lieberman (1970); Table 5 gives results for the sum of the probabilities using adaptive G-H with 2, 3, and 8 quadrature points per dimension and models with 1 through 3 factors.

Table 5 reveals that the 2- and 3-point G-H quadratures underestimate the sum, while the 8-point quadratures are accurate to five places for 1- and 2-factor models, and to four places for 3-factor models. Considering that the posterior densities for a 5-item test can be strongly skewed and long-tailed, the degree of inaccuracy for 2- and 3-point quadratures is not surprising. These inaccuracies affect the overall likelihood in that the likelihoods for the addition of the third factor (which is certainly over-factoring in this case) do not increase. Three points are enough to yield a fairly accurate value for the difference chi-square, indicating marginal significance of a second factor ($p = 0.06$ on four degrees of freedom), but the 2-point value is too low to do so. However,

TABLE 5.
Sum of response pattern probabilities and corresponding likelihoods and differce chi-squares for 1-, 2-, and 3-factor models for the 5-item LSAT-7 data evaluated by 2-, 3-, and 8-point Gauss-Hermite quadrature.

			Number of factors		
			1	2	3
Sum	Points	2	0.99748	0.99614	0.99366
		3	0.99880	0.99826	0.99413
		8	1.00000	1.00000	1.00005
$2 \times$ log-likelihood	Points	2	−5323.80	−5318.90	−5321.35
		3	−5320.80	−5312.40	−5318.82
		8	−5317.57	−5308.85	−5306.46
Chi-square	Points	2	4.90	–	
		3	8.40	–	
		8	8.72	2.39	

with larger numbers of items the posteriors can be expected to become essentially normal and integration with 2-point quadratures should be reasonably accurate.

The 2-factor model for the LSAT-7 data is also useful for demonstrating the convergence properties of the adaptive procedures. Panel 1 of Figure 1 shows MML the log-likelihoods for 100 iterations of Gibbs MCEM based on 100 and 1000 points, and the adaptive G-H with 10, 5, 3, and 2 points per dimension. This panel illustrates that 2- and 3-point G-H quadrature tends to underestimate the log-likelihoods while the 5-point quadrature accurately approximates the true log-likelihood and its path of convergence. The equivalence paths of the 5- and 10-point log-likelihoods suggest that results for 5-point G-H are almost exactly the results that would be obtained if all integrals could be integrated exactly. Note that the Gibbs log-likelihoods exhibit substantial random variability about the converged 10- and 5-point G-H log-likelihoods, especially the 100-point Gibbs log-likelihoods. This effectively rules out this method for significance tests of additional factors. Panel 2 of Figure 1 shows the effect of the mean and covariance adjustment on log-likelihood convergence for the 3-factor model. The non-adjusted 3-point log-likelihood iterates decrease rapidly, a result of EM losing its monontone convergence property with a numerical E-step. The decreasing log-likelihood often occurs whenever the full-information item factor model is over-fit to the data without the mean and covariance adjusted M-step. In contrast, the log-likelihood for the mean and covariance adjusted 3-point iterates converges to a maximum after 90 iterations. The covariance adjustment is often necessary to obtain convergent results, particularly when the number of factors is large relative to the number of items.

3.2.1. Likelihood Ratio Statistics for the Five-factor Simulations

Tests of the significance of an added fifth and sixth factor in the analysis of the 5-factor simulations by the NIS MCEM method, 5-, 3-, and 2-point adaptive G-H, and 3-point non-adaptive G-H are shown in Table 6. Although factor models are identified by their rotated solutions, the likelihoods are invariant with respect to the basis and are in fact computed from the fully-identified working solution of the TESTFACT program. The likelihood ratio chi-squares computed by TESTFACT are also compared to asymptotic chi-square tests of fit produced for generalized weighted least squares in GLS (see Muthén and Muthén, 1998–2001).

For the principal factor simulations, the difference chi-squares of all the integration methods accepted the presence of the fifth factor but differed with respect to the non-existent sixth factor.

TABLE 6.
Log-likelihoods and likelihood-ratio difference chi-squares five-factor simulated data N = 1,250.

	Simulation 1			Simulation 2			Simulation 3		
	Chi-square		Prob	Chi-square		Prob	Chi-square		Prob
Method	4 – 5	5 – 6	5 – 6	4 – 5	5 – 6	5 – 6	4 – 5	5 – 6	5 – 6
IS 243-pt	5512	20	1.0000	97	24	0.6210	746	41	0.0230
5-pt adaptive G-H	5595	71	0.1358	92	35	0.1406	761	43	0.0148
3-pt adaptive G-H	5536	101	0.0005	91	34	0.1710	760	51	0.0017
2-pt adaptive G-H	5434	90	0.0060	86	23	0.6594	762	39	0.0370
3-pt NAd G-H	5536	−104	–	146	−52	–	359	−108	–
GLS	27569	85	0.0146	191	55	0.0012	1556	56	0.0004

All the methods showed an order of magnitude difference in the chi-squares for the fifth and sixth factors across all the simulations. The 2- and 3-point G-H methods yielded nominally significant chi-squares for the sixth factor on 59 degrees of freedom for the first simulation, although with the conventional conservative rule requiring a chi-square twice the degrees of freedom for significance of all the adaptive methods would reject the sixth factor. For the second simulation with a principal factor sign pattern and random loadings, the chi-squares with 28 degrees of freedom for all methods accepted the fifth factor, and for the adaptive methods rejected the sixth factor. Results for the quasi-simple structure simulation showed generally correct acceptances and rejections using the conservative rule, with the exception of 3-point G-H, where the chi-square of 51 was only slightly larger than twice the degrees of freedom. Note that for all three simulations the 3-point non-adaptive method here was not accurate enough to produce a non-negative chi-square.

The asymptotic chi-squares produced by GLS likewise always accepted the presence of a fifth factor, but differed considerably from the MML methods with respect to the test of a sixth factor. GLS performed well in this respect for the first simulation, where responses were generated from items with zero intercepts. However, Simulations 2 and 3, with their randomly varying intercepts resulted in GLS chi-square values 57 and 10% greater than those produced by the adaptive MML methods.

3.2.2. Likelihood Ratio Statistics for Eight-factor Simulations

Likelihoods and likelihood-ratio chi-squares for the eight-factor simulations are shown in Table 7 only for the NIS and two-point G-H, along with the asymptotic chi-squares for GLS. The MML methods all showed a large decrease in the chi-squares from 8- to 9-factor models for both simulations – the GLS method also showed a large decrease but the chi-squares for the ninth factor remained very large, particularly for Simulation 4. Factor 8 was accepted in both simulations for all methods, but the non-existent ninth factor was not rejected for the 128-item principal factor simulation for the adaptive MML methods, although it would have been if the twice-the degrees-of-freedom rule for the difference chi-square had been used. In both simulations GLS produced chi-square values for the test of the ninth factor so large that a ninth factor would be accepted by any criterion. Moreover, the size of the differences in the chi-squares between GLS and the adaptive MML methods is substantial even for Simulation 5 where the difference in the RMSEs was relatively small.

TABLE 7.
Log-likelihoods and likelihood-ratio difference chi-squares eight-factor simulated data $N = 1{,}250$.

	Simulation 4			Simulation 5		
	Chi-square		Prob	Chi-square		Prob
Method	7 – 8	8 – 9	8 – 9	7 – 8	8 – 9	8 – 9
IS 243-pt	262	182	0.0002	902	48	0.2137
2-pt adaptive G-H	331	175	0.0007	887	33	0.8189
GLS	2373	1568	0.0000	1873	120	0.0000

TABLE 8.
Likelihood ratio difference chi-squares pedagogical content knowledge questionnaire $N = 640$.

	Chi-square				
Method	1 – 2 ($df = 32$)	2 – 3 ($df = 31$)	3 – 4 ($df = 30$)	4 – 5 ($df = 29$)	5 – 6 ($df = 28$)
NIS 243-point MCEM	411.24	100.16	91.10	65.44	50.12
5-point adaptive G-H	378.86	100.58	101.46	62.14	51.66
3-point adaptive G-H	351.00	101.60	99.20	65.26	49.86
2-point adaptive	302.86	109.08	65.14	64.90	51.64
GLS	489.00	71.47	186.90	73.50	62.36

3.3. Real data: Pedagogical Content Knowledge (PCK) Questionnaire

As a second example of the adaptive MML methods to real data, we analyzed a sample of 640 teachers from professional development institutes across California to a questionnaire examining pedagogical content knowledge (PCK) in mathematics (Hill, Schilling, and Ball, 2004). The test evaluated teacher knowledge in three content areas of the California grade school curriculum – (1) number concepts, (2) operations, and (3) patterns, functions, and algebra – and also examined understanding of typical student mistakes in number concepts and operations. In as much as patterns, functions, and algebra was a relatively recent addition to the curriculum, thereby not necessarily mastered by all teachers, we expected that a clear factor would emerge for these items. We also expected understanding of student thinking would identify a distinct factor.

The chi-square statistics for the sixth factor in Table 8 are significant at the 0.01 level for all the MML methods, even though conceptually there are only five types of items. This suggests the conservative rule which accepts a fifth factor but rejects a sixth factor for all the adaptive methods. Alternatively, most of the methods show a moderate decrease in the chi-squares from 4- to 5- factor models, suggesting a four-factor model. Comparison of the Promax-rotated factor loadings for the 4- and 5-factor models showed them to be about the same with the exception of two items that loaded on a separate factor for the five-factor model.

Overall the Promax rotated factor loadings for the MML and GLS methods for the five-factor model showed no substantive differences. Both showed a clear factor for the pattern functions and algebra items with a lesser tendency for the other items to load on separate factors relating to the conceptual division of the remaining items. But while the factor loadings failed to show substantive differences between MML and GLS, the chi-square statistics for GLS were consistently ten or more greater than those of MML. In our earlier simulations this occurred when the

recovery of factor loadings was less accurate for GLS, which in turn occurred when there was substantial variation in the item difficulties. The items of the Content Knowledge Questionnaire showed substantial variation in difficulty, ranging from 12 to 94% correct. While this is less variation than our designed examples, it is typical of the greater variation that is typically observed in educational examples. These items also differed from our simulations in their relatively small communalities, ranging from 0.06 to 0.93 with a median value of 0.30. These properties are likely to affect the relative performance of MML and GLS item factor analysis and are worthy of further examination.

In order to examine the effect of these factors on the relative performance of GLS and MML methods, we simulated responses of 1000 subjects in two conditions based on the estimated factor loadings from the 3-point G-H solution: (1)the original factor loadings and item intercepts with high variability obtained by randomly choosing the intercepts from a uniform $(-1.75, 1.75)$ distribution; and (2) the original factor loadings and item intercepts with low variability obtained by dividing the above item intercepts by 2. These simulations were repeated 25 times to allow us to compute RMSEs separately for each item and each factor loading. Chi-square statistics for the test of the sixth factor for 3-point and 2-point G-H EM, GLS, and NIS MCEM are presented in Figure 2. RMSE differences for GLS and 3-point G-H EM, Gibbs MCEM and 3-point G-H EM, 2-point G-H EM and 3-point G-H EM, and NIS MCEM and 3-point G-H EM for the high variability and low variability conditions are presented in Figures 3 and 4.

When the item difficulties exhibit low variability all the methods give essentially the same results, although the MML likelihood methods show a small but statistically significant advantage over GLS for both chi-square differences and RMSEs. But in the high variability condition the relative advantage of MML over GLS increases, with the MML methods showing both smaller chi-square differences for the test of the sixth factor and smaller RMSEs. There is little difference between the MML methods in this condition, with the exception of the 2-point G-H EM, where the RMSEs are significantly larger than the other MML methods and the chi-square differences are more variable. The larger RMSEs for the 2-point G-H EM estimates is largely due to the small communalities and smaller factor loadings for this example. When combined with highly variable item difficulties this leads to posterior distributions that are more strongly skewed and less normal in appearance. The effect is to essentially reduce the number of items when compared to applications with higher communalities and factor loadings. However, even here the 2-point G-H EM estimates still show significantly smaller RMSEs and chi-squares compared to GLS.

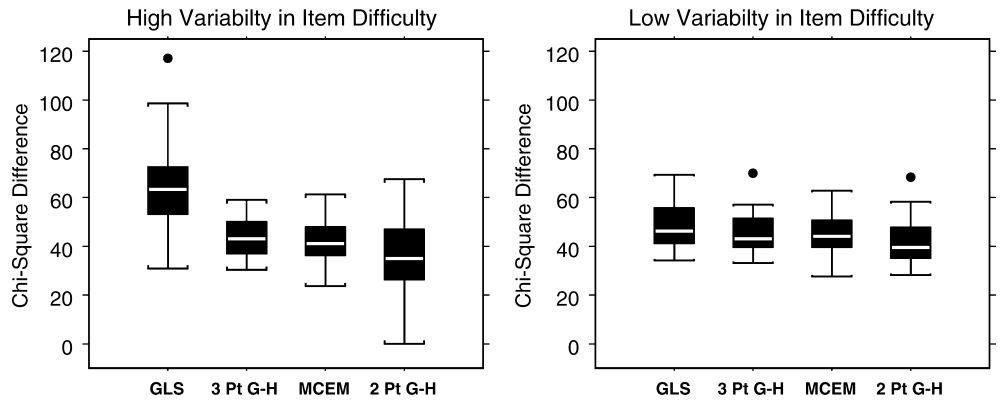

FIGURE 2.
Comparing chi-square differences for the test of a sixth factor.

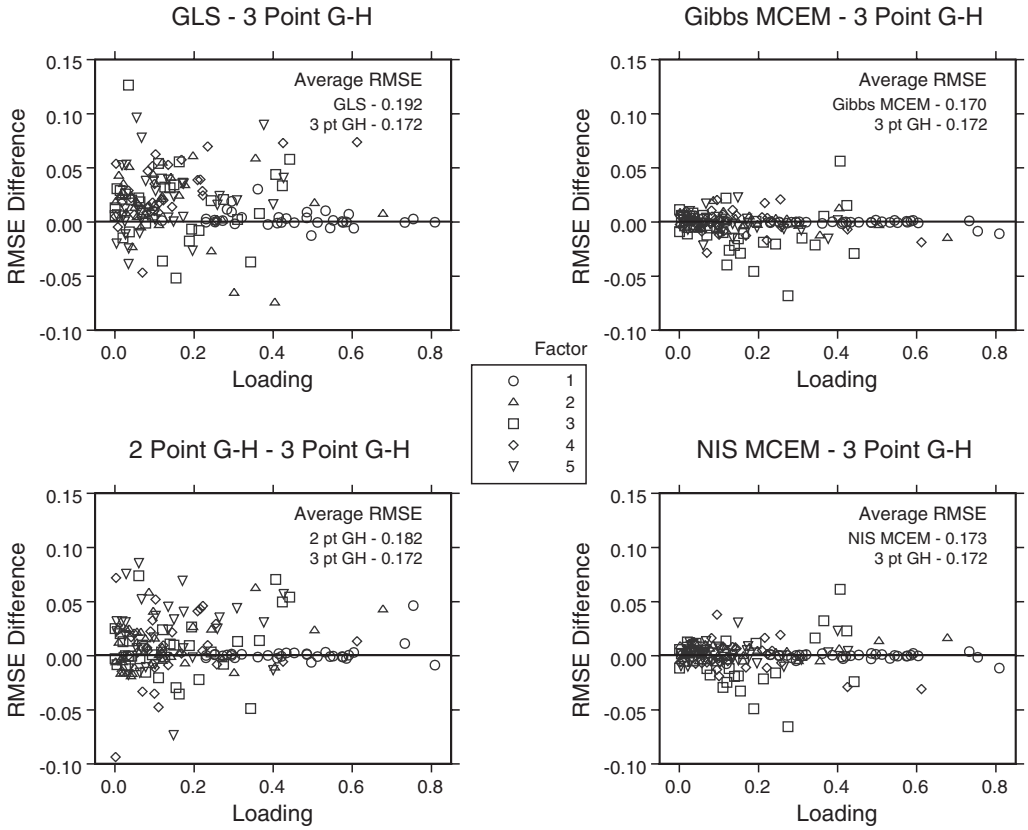

FIGURE 3.
Comparing RMSE differences: high variability in item difficulties.

4. Discussion and Conclusions

Multidimensional IRT models present unique challenges for estimation. For smaller numbers of items and factors adaptive quadrature with, say, five points per dimension and straightforward optimization algorithms such as Newton-Raphson can be employed. For example, the GLLAMM program developed by Rabe-Hesketh and her colleagues while this paper was under review (Rabe-Hesketh, Pickles, and Skrondal 2001; Rabe-Hesketh, Skrondal, and Pickles 2005a) uses adaptive quadrature and Newton-Raphson with a numerically estimated Hessian. Using the authors' recommended five points per dimension, this approach works well for fitting our model with smaller numbers of items, response patterns, and factors as the estimated log-likelihood accurately approximates the true log-likelihood. However, many applications in educational and psychological testing involve large numbers of items, response patterns, and possibly high dimensional factor structures. The number of parameters entails a high dimensional optimization problem, but the large number of response patterns and the necessity to perform integration over the latent space makes evaluation of the objective function, i.e., the marginal log-likelihood, very costly in absolute terms. Newton-Raphson is typically impractical in this situation, due to the large size of the Hessian and the tendency of the Hessian to become indefinite far from the solution. Conjugate gradient methods (Polak, 1971) or direction set methods (Powell, 1964) provide possible alternatives, but in our experience they typically require on the order of hundreds or thousands of function evaluations.

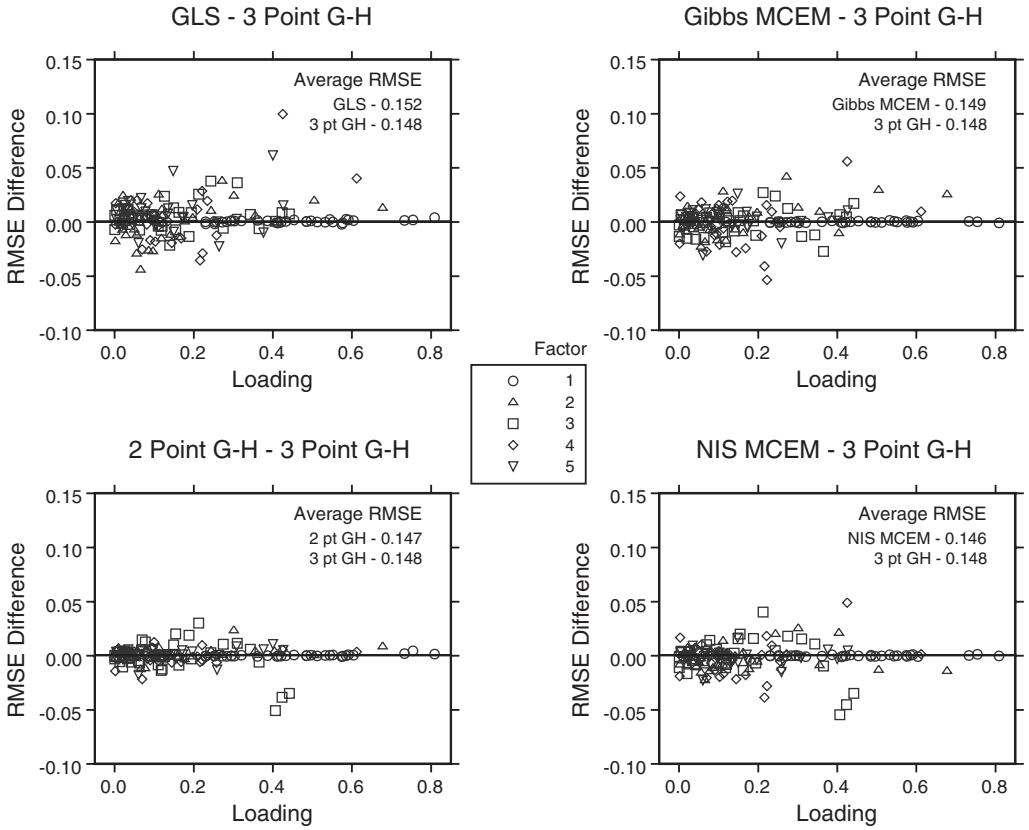

FIGURE 4.
Comparing RMSE differences: low variability in item difficulties.

We found that the EM algorithm with mean and covariance adjustments gives exactly the same results as these two methods in a fraction of time. For example, a two-factor model applied to the PCK questionnaire using Powell's direction set algorithm with GH-3 took over 7 h to converge on a 1.6 GZ laptop compared to only 32 s for EM; the five-factor model took less than 11 min for EM to converge. The time required for GH EM runs of other examples was linear in the number of integration points, parameters, and subjects. Time required for NIS EM was nearly identical, while Gibbs EM method of Meng and Schilling typically required 40% more computing time, due to the time needed to generate the MCMC draws.

Alternatively, a full Bayesian analysis using Markov chain Monte Carlo (Schilling, 1993; Ansari and Jedidi, 2000; Fox and Glass, 2001) could be employed, but this raises a host of issues, such as appropriate specification of priors to avoid non-recurrent chains and monitoring convergence of MCMC in high-dimensional space. Even if one chooses a Bayesian approach using MCMC, the methods we describe can easily be adapted to find the mode of the posterior distribution of the item parameters, which can provide an efficient check on the convergence of the MCMC procedure.

Because the full information item factor model is an example of two-level hierarchical data analysis in which the model is non-linear and the latent variables possibly of high dimensionality, the likelihood equations must be evaluated numerically. Although various linear approximation methods have been proposed in this situation (Lindstrom and Bates, 1990; Raudenbush, Yang, and Yosef, 2000), the only methods that have been applied to any extent are those based on numerical integration of functions of posterior densities (see Bartholomew and Knott, 1999). In

this study, we have shown that Gibbs, NIS, and G-H quadrature perform well in recovering the generating parameters of simulated data across conditions spanning the possible combinations of communality, item difficulty, and structure of the factor loadings. As might be expected, Gibbs performs slightly better than NIS or G-H because the normal approximation is not involved. In computing likelihoods, Gibbs is accurate on an average over the iterations of the solution, but it varies randomly from trial to trial to such an extent that the likelihoods of two different analyses at arbitrary stopping points cannot be compared. Because we consider the likelihood ratio test of the statistical significance of an additional factor in the model an indispensable part of item factor analysis, we do not recommend the Gibbs method for practical work. Our version of NIS behaves much better in this respect: if the number of random points drawn is in the order of several hundreds, this method attains virtually the same likelihood at convergence as the average Gibbs method. Because it performs well both in estimation and in determining the number of factors, and has no definite limit on the number of items or factors, our results broadly support the use of this method of MML item factor analysis in high dimensions.

In our study of G-H quadrature, a key result is the improved accuracy gained by adapting the quadrature points to the individual posterior distributions rather than using a fixed grid over the whole latent space. Except with short tests where the posteriors are sufficiently diffuse to support by a fixed grid, the gain in accuracy of parameter estimation and likelihood evaluation from adaptive quadrature is readily apparent in our simulations. The new and surprising result is adjusting the EM iterates for the estimated mean and covariance of the latent distribution, an adjustment designed to speed convergence of the EM algorithm, actually facilitates integration to the extent that fewer quadrature points can be used while still obtaining a convergent solution with respect to the log-likelihood. Our experience is that this adjustment is critical in stabilizing the log-likelihood computations to enable convergence, particularly when a model is over-fit to the data. As noted earlier, this stability of the computed log-likelihood is essential in the statistical test of when to stop factoring.

In tests with small numbers of items, where the posteriors may be strongly skewed and far from normal, the two-point integrations tend to underestimate the marginal probabilities of the response patterns and thus the overall likelihood of the parameter set. But with the large numbers of items typically encountered in item factor analysis, this bias becomes negligible. In applications with up to eight factors, the faster computing time favors adaptive G-H quadrature for routine use. When the number of factors is relatively small, less than six, it remains advantageous for the most accurate integration to use, say, 10 points per dimension for 1 or 2 factors, 5 for 3, 4 for 4, and 3 for 5. For 6–10 factors, however, only 2 points per dimension is feasible for routine work in present computers. Beyond that, one must resort to the NIS method with up to perhaps 500 draws per integration. Until now we have had no experience with applications of this scope, but they might arise in factor studies of the item pools of large-scale testing programs. In respect to the limited information methods, we had expected from previous experience with tetrachoric factor analysis that MML would be superior to MINRES and GLS, when item difficulties vary widely, as is often the case in real data. Our simulations bore this out – the RMSEs and chi-squares showed a substantial advantage for MML when the item difficulties varied widely. Even when item difficulties were less extreme, MML had relative advantage in testing the statistical significance of an added factor.

However, our results indicate that using a strict criterion of statistical significance at the 0.05 level for the likelihood ratio chi-square is ill-advised, particularly for the 2- and 3-point G-H methods; such a criterion leads to incorrect acceptance of an additional factor for more than 50% of the tests we considered. A better alternative is the conservative rule of requiring a chi-square of twice the degrees of freedom such as Akaike's information criterion (AIC). This always produced a correct decision in our simulated examples. Alternatively, one can make use of the tendency of the likelihood ratio chi-square to decrease precipitously when a model is overfit to the data, a con-

stant across our simulations. Our real data examples suggest that this might be the most effective approach, since real data often have the likelihood ratio chi-squares inflated by some degree of overdispersion. Note that the tendency for inflated likelihood-ratio statistics in this context differs from the tendency of deflated chi-square statistics in the mixture of chi-square problem identified by Raudenbush and Bryk (2002, page 284), since the null hypothesis here is not on the boundary of the parameter space and encompasses many parameters in the form of factor loadings. Further research on this topic is clearly needed.

In the course of the present study we have identified and implemented all of the necessary conditions that must be met when applying the EM algorithm to maximum marginal likelihood item factor analysis, either by Monte Carlo integration or adaptive quadrature. These developments will also be important in any attempt to generalize binary-item factor analysis to polytomously scored items. There, the much greater information carried by graded scoring leads to concentrated posteriors in tests with many fewer items than a comparable dichotomously scored test. Some form of adaptive quadrature is essential in that type of application. Similar considerations apply in other multilevel treatments of binary and polytomous data (see Hedeker and Gibbons, 1994), a fact confirmed by a number of recent papers on generalized linear mixed models (Lesaffre and Spiessens, 2001; Rabe-Hesketh, Skrondal, and Pickles 2002, 2005b). One interesting finding specific to this application is that the iterative estimation of the first and second moments advocated by Naylor and Smith (1982) facilitates estimation in multilevel models with more than two levels of nesting (see Rabe-Hesketh, Skrondal, and Pickles 2005b).

In the context of item response theory, the main limitation to our approach in the estimation using adaptive numerical integration is the absence of a fixed grid of points to support a finite approximation of the latent distribution. That type of representation has many uses, for example resolving the latent distribution into normal components (see Mislevy, 1984). But in very large samples, this obstacle could perhaps be overcome by accumulating the pattern posterior densities into the cells of a frequency table defined by a limited number of intervals on each dimension of the latent distribution. Alternatively, one might represent the corresponding cumulative latent distributions as multidimensional spline functions (see Ramsay, 1998).

References

Ahrens, J.H. & Dieter, U. (1979). Computer methods for sampling from the exponential and normal distributions. *Communications of the Association for Computing Machinery*, *15*, 873–882.

Ansari, A. & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65, (4)*, 475–496.

Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*. New York: Oxford.

Bock, R.D. (1975/1985). *Multivariate Statistical Methods in Behavioral Research*. New York: McGraw-Hill; 1985 reprint, Chicago: Scientific Software International.

Bock, R.D. & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika*, *35*, 179–197

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

Bock, R.D., Gibbons, R.D., & Muraki, E. (1987). Full information item factor analysis. *Applied Psychological Measurement*, *12, (3)*, 261–280

Bock, R.D. & Schilling, S.G. (1997). High-dimensional full-information item factor analysis. In M. Birkane (Ed.), *Latent Variable Modeling and Applications to Causality* (pp. 163-176). New-York: Springer.

Bock, R.D., Gibbons, R.D., Muraki, E., Schilling, S.G., Wilson, D.T. & Wood, R. (1999). *TESTFACT 3: Test Scoring, Item Statistics, and Full-information Item Factor Analysis*. Chicago: Scientific Software International.

Divgi, D.R. (1979). Calculation of the tetrachoric correlation coefficient. *Psychometrika*, *44*, 169–172.

Ferguson, G.A. (1941). The factorial interpretation of test difficulty. *Psychometrika*, *6*, 323–329.

Fox, J.P. & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66, (2)*, 271–288.

Guilford, J.P. (1941). The difficulty of a test and its factor composition. *Psychometrika*, *6*, 66–77.

Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, *5*, 1148–1169.

Harman, H.H. (1987). *Modern Factor Analysis*. Chicago: University of Chicago Press.
Hedeker, D., & Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.
Hill, H.C., Schilling, S.G. & Ball, D.L. (2004). Developing measures of teachers mathematics knowledge for teaching. *Elementary School Journal*, in press.
Householder, A.S. (1964). *The Theory of Matrices in Numerical Analysis*. New York: Blaisdell.
Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.
Leonelli, B.T., Chang, C.H., Bock, R.D., & Schilling, S.G. (2000). A full-information item factor analysis interpretation of the MMPI-2: Normative Sampling with Non-pathonomic Descriptors. *Journal of Personality Assessment*, *74, (3)*, 400–422.
Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, *50*, 325–335.
Lindstrom, M.J. & Bates, D.M. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics*, *46*, 673–687.
Liu, C., Rubin, D.B., & Wu, Y.N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, *85, (4)*, 755–770.
Liu, Q. & Pierce, D.A. (1994). A note on G-H quadrature. *Biometrika 81, (3)*, 624–629.
Meng, X.L. & Schilling, S.G. (1996). Fitting full-information factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*, 1254–1267.
Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika*, *49, (3)*, 359–381.
Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika 49*, 115–132
Muthén, L.K., & Muthén, B.O. (1998–2001). *Mplus User's Guide (Second edition).* Los Angeles, CA: Muthén & Muthén.
Naylor, J.C. & Smith. A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, *31*, 214–225.
Polak, E. (1971). *Computational Methods in Optimization.* New York: Academic Press.
Powell, M.J.D. (1964). An efficient method for several variables without calculating derivatives. *Computer Journal*, *7*, 155–162.
Rabe-Hesketh, S., Pickles, & A., Skrondal, A., (2001). *GLLAMM Manual*. Tech. rept. 2001/01. Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London. Downloadable from http://www.gllamm.org.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*, 1–21.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005a). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.
Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2005b). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, in press.
Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society, Series B.*,*60*, 365–375.
Raudenbush, S.W., Yang, M., & Yosef (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9, (1)*, 141–157.
Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks: Sage.
Ripley, B.D. (1987). *Stochastic Simulation.* New York: Wiley
Schilling, S.G. (1993). *Advances in Full Information Item Factor Analysis using the Gibbs Sampler*. (Unpublished doctoral dissertation, University of Chicago).
Schrage, L. (1979). A more portable fortran random number generator. *Association for Computing Machinery: Transactions on Mathematical Software*, *5*, 132–138.
Thurstone, L.L. (1947). *Multiple Factor Analysis*. Chicago: The University of Chicago Press.
Thurstone, L.L. & Thurstone, T.G. (1941). Factorial studies of intelligence. *Psychometric Monographs No. 2*. Chicago: University of Chicago Press.
Tierney, L. & Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.
Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, *85*, 699–704.
Wood, R., Wilson, D.T., Gibbons, R.D., Schilling, S.G., Muraki, E., & Bock, R.D. (2003). *TESTFACF 4: Test Scoring, Item Statistics, and Full-information Item Factor Analysis*. Chicago: Scientific Software International.
Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1995) *BILOG-MG: multiple-group item analysis and test scoring*. Chicago: Scientific Software International.

Manuscript received 28 OCT 2003
Final version received 29 MAY 2004