PSYCHOMETRIKA—VOL. 51, NO. 2, 177–195
JUNE 1986

BAYES MODAL ESTIMATION IN ITEM RESPONSE MODELS

ROBERT J. MISLEVY

EDUCATIONAL TESTING SERVICE

This article describes a Bayesian framework for estimation in item response models, with two-stage prior distributions on both item and examinee populations. Strategies for point and interval estimation are discussed, and a general procedure based on the EM algorithm is presented. Details are given for implementation under one-, two-, and three-parameter binary logistic IRT models. Novel features include minimally restrictive assumptions about examinee distributions and the exploitation of dependence among item parameters in a population of interest. Improved estimation in a moderately small sample is demonstrated with simulated data.

Key words: Bayesian estimation, EM algorithm, hierarchical prior distributions, item response models, marginal maximum likelihood.

Introduction

Simultaneous estimation of many parameters can often be improved, sometimes dramatically so, if it is reasonable to consider one or more subsets of parameters as exchangeable members of corresponding populations (Efron & Morris, 1975; James & Stein, 1961; Kelley, 1927; Lindley & Smith, 1972). The idea is that while each observation may provide limited information about the parameters it is modeled directly in terms of, it also contributes information about the populations to which they belong. Knowledge about the populations, generally superior to knowledge about individual parameters, can in turn be brought to bear in the estimation of any individual parameter. Novick, Jackson, Thayer, and Cole (1972) and Rubin (1980), for example, provide Bayes and empirical Bayes solutions respectively to the problem of predicting student performance in a given law school when data are available from several law schools. Both studies obtained more stable estimates in small schools and improved cross-validation results when compared to independent estimation within schools.

This approach holds promise for parameter estimation in item response theory (IRT) models for mental tests—a thorny problem even in large samples, since additional parameters are introduced with each new item and each new examinee in the sample. Many individual parameters can be poorly determined despite an apparently large data base. Wainer and Thissen (1982), for example, show that several thousand responses are sometimes needed to obtain stable maximum likelihood estimates of an item's parameters under the three-parameter logistic model.

Such procedures have recently begun to appear in the psychometric literature. Bock and Aitkin (1981), Rigdon and Tsutakawa (1983), and Thissen (1982) address the problem of incidental examinee parameters by integrating over a population density to produce marginal likelihood functions for item parameters. Reiser (1981) and Mislevy and Bock

Requests for reprints should be sent to Robert J. Mislevy, Educational Testing Service, Princeton, NJ 08541.

This research was supported by a grant from the Spencer Foundation, Chicago, IL. Comments and suggestions on earlier drafts by Charles Lewis, Frederic Lord, Paul Rosenbaum, James Ramsey, Hiroshi Watanabe, the editor, and two anonymous referees are gratefully acknowledged.

0033-3123/86/0600-7092$00.75/0

(1985) extended this model by positing prior distributions for item parameters. Swaminathan and Gifford (1981, 1982, 1985) employ two-stage priors for examinee parameters and selected item parameters, then obtain the joint posterior mode for all individual parameters. Andersen and Madsen (1977), Mislevy (1984), and Sanathanan and Blumenthal (1978) provide maximum likelihood solutions for the parameters of examinee population distributions, conditional on item parameters. Finally, Bock and Aitkin (1981) and Bock and Mislevy (1982) derive posterior means and standard deviations of the parameters of individual examinees, conditional on item and examinee population parameters.

The aforementioned procedures can all be expressed as special cases of a more comprehensive Bayesian framework for estimation in item response models. Working along lines first suggested by Lindley and Smith (1972), we begin by introducing a model for item responses that employs two levels of prior distributions on both item and examinee parameters. Attention is restricted to those cases in which it is reasonable to assume exchangeability over all items and over all examinees; that is, no information other than item responses is available to lead us to different expectations about particular items or examinees. A general discussion of theoretical and practical considerations in estimating the parameters of such a model, including an EM computing algorithm (Dempster, Laird, & Rubin, 1977), follows. Procedures specific to some binary logistic item response models (Birnbaum, 1968; Lord, 1980; Rasch, 1960) are then detailed. We illustrate the techniques with simulated data and conclude by discussing possible extensions of the procedures to take into account auxiliary information about items or examinees.

The General Form of the Model

Let θ denote examinee ability and $p(\theta \mid \tau)$ its density, conditional on examinee population parameters τ. If θ follows a normal distribution, for example, $\tau = (\mu_\theta, \sigma_\theta^2)$, the mean and variance. τ is assumed in turn to follow a density $p(\tau)$. In the same manner, let ξ denote the parameter(s) of a test item and $p(\xi \mid \eta)$ denote its density, conditional on item population parameters η; η in turn follows density $p(\eta)$. Independence over examinees and items is assumed, given τ and η. The joint prior for all unknowns is thus given by

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \tau, \eta) = \prod_i p(\theta_i \mid \tau) \prod_j (\xi_j \mid \eta) p(\tau) p(\eta).$$

Let d_{ij} take the value 1 if examinee i is administered item j and 0 if not. For n items of interest, let $\mathbf{d}_i = (d_{i1}, \ldots, d_{in})'$, and for N examinees, let $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_N)'$. Let u_{ij} denote the response of examinee i to item j, taking the value 1 if the item was adminstered and answered correctly, and 0 otherwise; define $\mathbf{u}_i$ and $\mathbf{U}$ in analogy to $\mathbf{d}_i$ and $\mathbf{D}$. We shall assume that the mechanism by which items are administered is ignorable with respect to likelihood inference (Rubin, 1976); that is, $p(\mathbf{U}, \mathbf{D} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \tau, \eta) = Kp(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\xi}, \tau, \eta)$. Random item administration, targeted testing, and adaptive testing are examples of ignorable item administration mechanisms (Mislevy, in preparation). Denote by $L(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\xi})$ the likelihood of the possibly incomplete matrix of responses of subjects with abilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ to items with parameters $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$. By Bayes theorem, the posterior density of $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, τ, and η, given realized observations $\mathbf{U}$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}, \tau, \eta \mid \mathbf{D}, \mathbf{U}) \propto L(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\xi}) \cdot p(\boldsymbol{\theta} \mid \tau) \cdot p(\tau) \cdot p(\boldsymbol{\xi} \mid \eta) \cdot p(\eta). \tag{1}$$

After the forms of the likelihood function L and the prior densities $p(\boldsymbol{\theta} \mid \tau)$ and $p(\boldsymbol{\xi} \mid \eta)$ have been chosen, the highest level prior densities $p(\tau)$ and $p(\eta)$ have been specified, and the data $\mathbf{U}$ have been observed, (1) contains all information available about the parameters in the model. The intractability and the sheer incomprehensibility of a joint distribution of possibly thousands of variables, however, demand summarization in terms of

salient attributes, to be used in constructing point and interval estimates or for subsequent calculations.

The mean of the posterior has the desirable property that its value for any subset of parameters is invariant with respect to marginalization of (1) over any subset of remaining variables. In complex problems like the one at hand, however, where closed form solutions are not readily forthcoming, posterior modes are more frequently seen. They are not invariant with respect to marginalization, but they are often easier to approximate numerically.

Generally speaking, the posterior mean of a subset of parameters in multiparameter problems is better approximated by the mode after marginalization over one or more other parameters (O'Hagan, 1976). Improvements can be striking when marginalization is over large numbers of poorly determined "nuisance" parameters that appear in the joint posterior along with parameters of interest. Examinee parameters θ follow this description in the present context, and we integrate over their distribution to obtain

$$p(\xi, \tau, \eta \mid \mathbf{D}, \mathbf{U}) = \int p(\boldsymbol{\theta}, \xi, \tau, \eta \mid \mathbf{D}, \mathbf{U})\, d\boldsymbol{\theta}$$

$$\propto L(\mathbf{U} \mid \mathbf{D}, \xi, \tau) p(\tau) p(\xi \mid \eta) p(\eta). \tag{2}$$

Modal estimates from (2), as opposed to those from (1), are also preferable on the grounds of asymptotic behavior. Bayes modal estimates tend to normality under regularity conditions similar to those required for asymptotic normality of the maximum likelihood estimate, with the additional requirement of a continuous and positive prior distribution in the neighborhood of the modal value (Hartigan, 1983); in the limit, the modal value is equal to the mean, which is equal in turn to the maximum likelihood estimate. The precision matrix, or the inverse of the posterior covariance matrix, is given by the negative matrix of second derivatives of the log posterior evaluated at that point.

The regularity conditions are not satisfied, however, in the case of infinitely many nuisance parameters (Neyman & Scott, 1948)—a situation obtained when the number of items is held constant and the number of examinees increases without limit (Andersen, 1973). Asymptotic normality of the modal estimator under (1), then, need not follow. It *will* follow for modal estimators of ξ, τ, and η under (2) with the well-behaved item response functions and prior distributions in current use, because the number of parameters remains constant with increases in the number of examinees.

Even when the θ parameters of individual examinees are of interest, as they often are in practice, the reasoning above implies that their modal values in (1) may not be a good way to estimate them. Bock and Aitkin (1981) suggest approximating the relevant marginal posterior $p(\theta_i \mid \mathbf{U}, \mathbf{D})$ by $p(\theta_i \mid \mathbf{u}_i, \mathbf{d}_i, \hat{\xi}, \hat{\tau}, \hat{\eta})$, a distribution conditional on modal estimates of item and population parameters; Bock and Mislevy (1982) offer a tractable numerical approximation of this distribution, from which posterior means and standard deviations are readily calculated. The effects of ignoring uncertainty about item and population parameters prove negligible for more than, say, 10 appropriate items, but Dunmore's (1976) adjustment using the posterior precision matrix can be applied if desired. The exact form of $p(\theta_i \mid \mathbf{U}, \mathbf{D})$ in the special case of the case of the two-parameter normal ogive IRT model has been given by Lewis (1985).

It would seem appealing at first blush to apply the same techniques to the estimation of item parameters as well, first marginalizing over $\boldsymbol{\theta}$ and ξ to obtain

$$p(\tau, \eta \mid \mathbf{D}, \mathbf{U}) \propto \iint p(\boldsymbol{\theta}, \xi, \tau, \eta \mid \mathbf{D}, \mathbf{U})\, d\boldsymbol{\theta}\, d\xi, \tag{3}$$

obtaining modal estimates of τ and η, then calculating $p(\xi_j \mid U_i, D_j, \hat{\tau}, \hat{\eta})$, where U_j and D_j

represent data for item j only. This approach will not be pursued in the current presentation. A first reason stems from the fact that in applied work, the number of examinees is typically much larger than the number of items. This means the posteriors of items parameters are usually much better determined and more nearly symmetric, so that modal values of ξ in $p(\xi \mid \mathbf{U}, \mathbf{D}, \hat{\tau}, \hat{\eta})$ differ little from those in (2). A second reason pertains to IRT models with more parameters per item than response categories. As the number of examinees increases without limit, the likelihood $L(U_j \mid D_j, \xi_j)$ is maximized by all values of ξ that trace a response curve though the observed proportion correct. For binary IRT models with one parameter per item, for example, there is just one such value of ξ; but for models with more than one parameter per item, there are infinitely many. In this latter case, the density $p(\xi_j \mid U_j, D_j, \hat{\tau}, \hat{\eta})$ does not lead to unique item parameter estimates as the number of examinees increases; a density that retains joint response frequencies across items, namely $p(\xi \mid \mathbf{U}, \mathbf{D}, \hat{\tau}, \hat{\eta})$, must instead be addressed and the attraction of one-item-at-a-time solutions is lost.

If item population parameters are not of interest, they can be integrated out to yield

$$p(\xi, \tau \mid \mathbf{D}, \mathbf{U}) \propto \int_{\eta} \int_{\theta} p(\theta, \xi, \tau, \eta \mid \mathbf{D}, \mathbf{U})\, d\theta\, d\eta. \tag{4}$$

The remaining item parameters and examinee population parameters are typically of primary interest in the educational setting, although for many examinees and all but very short tests, their marginal modes under (2) and (4) will differ little.

An EM Algorithm for Parameter Estimation

This section provides a framework for parameter estimation in the general model outlined above, based on the variation of Dempster, Laird, and Rubin's EM algorithm introduced by Bock and Aitkin (1981) in the context of marginal maximum likelihood (MML) estimation of item parameters. The posterior density function in our model, marginalized with respect to θ, can be written as

$$p(\xi, \tau, \eta \mid \mathbf{D}, \mathbf{U}) = \left\{ \int L(\mathbf{U} \mid \mathbf{D}, \theta, \xi) p(\theta \mid \tau)\, d\theta \right\} \cdot \{p(\tau) p(\xi \mid \eta) p(\eta)\}. \tag{5}$$

The first bracketed factor on the right is the marginal probability of observed responses from a random sample of examinees from a population with density $p(\theta \mid \tau)$, while the second can be thought of as the prior distribution for ξ and τ. We now focus our attention on the first factor.

By maximizing the first factor of (5) with respect to parameters of interest, Bock and Aitkin (1981) obtain MML estimates of ξ given $p(\theta \mid \tau)$ and Mislevy (1984) obtains MML estimates of τ given $L(\mathbf{U} \mid \mathbf{D}, \theta, \xi)$. Both presentations employed the expedient of approximating integration over θ by summation over a finite grid of points X_q, $q = 1, \ldots, Q$, with associated weights $A(X_q \mid \tau)$ as follows:

$$\log L(\mathbf{U} \mid \mathbf{D}, \xi, \tau) \approx \sum_i \log \sum_q L(\mathbf{u}_i \mid \mathbf{d}_i, X_q, \xi) A(X_q \mid \tau). \tag{6}$$

Three methods were suggested for specifying points and weights. First, when $p(\theta \mid \tau)$ takes the form of a normal density or a mixture of normal densities, optimal points and weights for a given Q may be found in Stroud and Sechrest (1966). Second, a Monte Carlo approach generates a random sample of equally weighted points from $p(\theta \mid \tau)$. Third, a grid of Q equally-spaced points can be specified a priori and assigned weights proportional to $p(X_q \mid \tau)$.

Bock and Aitkin (1981) show that with the discrete approximation of the likelihood function, partial derivatives of the marginal likelihood, in which θ's are not observed but must be inferred from item responses, can be written in forms quite similar to their counterparts in a related "complete data" problem in which individual θ's are known. Under the assumption of iid θ's, we may write the partial derivative of the *complete data* log likelihood, or

$$\log L(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \xi, \tau) = \log L(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \xi) + \log p(\boldsymbol{\theta} \mid \tau), \tag{7}$$

with respect to a typical parameter v from ξ or τ in the form

$$\frac{\partial \log L(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \xi, \tau)}{\partial v} = \sum_i f_v(\mathbf{u}_i, \mathbf{d}_i, \theta_i, \xi, \tau), \tag{8}$$

for an appropriately defined gradient function f_v. It can be shown (e.g., Bock & Aitkin, 1981) that the corresponding derivative of the *marginal* log likelihood (6) can then be approximated as

$$\frac{\partial \log L(\mathbf{U} \mid \mathbf{D}, \xi, \tau)}{\partial v} \approx \sum_q f_v(\hat{\mathbf{r}}_q, \hat{\mathbf{N}}_q, X_q, \xi, \tau), \tag{9}$$

where

$$\hat{N}_{qj} = \sum_i d_{ij} P(X_q \mid \mathbf{u}_i, \mathbf{d}_i, \xi, \tau), \tag{10}$$

and

$$\hat{r}_{qj} = \sum_i d_{ij} u_{ij} P(X_q \mid \mathbf{u}_i, \mathbf{d}_i, \xi, \tau) \tag{11}$$

with

$$P(X_q \mid \mathbf{u}_i, \mathbf{d}_i, \xi, \tau) = \frac{L(\mathbf{u}_i \mid \mathbf{d}_i, X_q, \xi) A(X_q \mid \tau)}{\sum_s L(\mathbf{u}_i \mid \mathbf{d}_i, X_s, \xi) A(X_s \mid \tau)}. \tag{12}$$

An application of Bayes theorem will be recognized in (12), yielding a value approximately proportional to the posterior density of θ at the value X_q given $\mathbf{u}_i$, $\mathbf{d}_i$, ξ, and τ. The upshot is that the first derivatives (9) of the marginal likelihood are identical in form to the first derivatives (8) of the complete data likelihood, with expressions for subjects evaluated at θ_i with *observed data* u_{ij} and d_{ij} replaced by similar expressions evaluated at quadrature points X_q with *pseudo-data* $\hat{r}_{qj}$ and $\hat{N}_{qj}$. Likelihood equations are obtained by setting the partial derivatives (9) to zero.

It will be noted that $\hat{\mathbf{r}}_q$ and $\hat{\mathbf{N}}_q$ depend on ξ and τ. Solution must proceed iteratively taking the form described by Dempster et al. (1977, sec. 4.1.1) for EM cycles with missing values under multinomial sampling. In the E-step, (10) and (11) are evaluated with provisional estimates $\hat{\xi}^t$ and $\hat{\tau}^t$. This gives the expectations of $\hat{\mathbf{r}}_q$ and $\hat{\mathbf{N}}_q$ conditional on the data and the provisional parameter estimates. In the M-step, $\hat{\xi}^{t+1}$ and $\hat{\tau}^{t+1}$ are obtained by solving (9) (equated to zero) with $\hat{\mathbf{r}}_q$ and $\hat{\mathbf{N}}_q$ treated as known. Cycles continue in this manner until changes become negligible. An indication of the precision of estimation is given by the following approximation of the Fisher information matrix:

$$\mathbf{A} = \sum_i \left(\frac{\partial \log L(\mathbf{u}_i \mid \mathbf{d}_i, \xi, \tau)}{\partial(\xi, \tau)} \right) \left(\frac{\partial \log L(\mathbf{u}_i \mid \mathbf{d}_i, \xi, \tau)}{\partial(\xi, \tau)'} \right), \tag{13}$$

evaluated at $(\hat{\xi}, \hat{\tau})$.

The EM algorithm is readily extended to Bayes modal estimation (Dempster et al., 1977, p. 6). All of the foregoing procedures are applied as before, except that the marginal likelihood equations are replaced by modal estimation equations of similar form; for a typical element v of ξ or τ,

$$0 = \frac{\partial \log p(\xi, \tau \mid \mathbf{D}, \mathbf{U})}{\partial v}$$

$$= \frac{\partial \log p(\mathbf{U} \mid \mathbf{D}, \xi, \tau)}{\partial v} + \frac{\partial \log p(\xi \mid \eta)}{\partial v} + \frac{\partial \log(\tau)}{\partial v}. \tag{14}$$

The treatment of item population parameters η, which do not appear in (9), depends on whether they are to be integrated out or jointly estimated. Integrating them out modifies the form of the prior for ξ from $p(\xi \mid \eta)$ to $\int p(\xi \mid \eta)p(\eta)\, d\eta$. Estimating them requires the solution of additional equations

$$0 = \frac{\partial \log p(\xi \mid \eta)p(\eta)}{\partial \eta}.$$

When η has been integrated out in the problem at hand, the posterior precision matrix for ξ and τ takes the form

$$\mathbf{B} = -\frac{\partial^2 \log L(\mathbf{U} \mid \mathbf{D}, \xi, \tau)}{\partial(\xi, \tau)\partial(\xi, \tau)'} \frac{\partial^2 \log p(\xi)p(\tau)}{\partial(\xi, \tau)\partial(\xi, \tau)'}, \tag{15}$$

where

$$p(\xi) = \int p(\xi \mid \eta)p(\eta)\, d\eta.$$

Employing the well-known result on Fisher's information matrix and substituting observed values for expectations, we avoid calculating second derivatives of the log likelihood via the approximation

$$\mathbf{B} \approx \mathbf{A} - \frac{\partial^2 \log p(\xi)p(\tau)}{\partial(\xi, \tau)\partial(\xi, \tau)'}, \tag{16}$$

where $\mathbf{A}$ is given in (13). When η is estimated jointly with ξ and τ, the precision matrix is similarly approximated as

$$\mathbf{C} = \begin{cases} \mathbf{A} - \dfrac{\partial^2 \log p(\tau)p(\xi \mid \eta)}{\partial(\xi, \tau)\partial(\xi, \tau)'} & \text{(symmetric)} \\ -\dfrac{\partial^2 \log p(\tau)p(\xi \mid \eta)p(\eta)}{\partial(\xi, \tau)\partial\eta'} & -\dfrac{\partial^2 \log p(\tau)p(\xi \mid \eta)p(\eta)}{\partial\eta\, \partial\eta'} \end{cases} \tag{17}$$

It should be pointed out that solutions of the estimation equations are local extrema or saddle points of the posterior. Whether they are local maxima can be determined by examining the shape of the posterior in the neighborhoods of solutions, either empirically or through the matrix of second derivatives, which will be negative definite at local maxima. Whether a local maximum is a global maximum follows in certain cases from the form of the posterior (e.g., a member of the exponential family), but must be determined empirically in most cases by starting the iterative solution from a number of different initial values.

Procedures for Some Logistic Models

The balance of the article implements the procedures in the context of some logistic item response models, with special focus on Birnbaum's three-parameter model. The following sections provide details on functional forms for the likelihood and prior distributions, and on the corresponding forms of the fitting equations. For the first stage of priors, a multivariate normal density will be posited for item thresholds, log slopes, and logit asymptotes; both a mixture of normal components and a nonparametric approximation in the form of a histogram will be provided for examinee abilities. For the second stage, both diffuse and natural conjugate priors will be provided in all cases.

The Likelihood Function

The three-parameter logistic model for dichotomous items (Birnbaum, 1968) gives the probability of a correct response to item j from examinee i as

$$\begin{aligned} P_j(\theta_i) &= P(u_{ij} = 1 \mid \theta_i, a_j, b_j, c_j) \\ &= c_j + (1 - c_j)\Psi[Da_j(\theta_i - b_j)], \end{aligned} \tag{18}$$

where $\Psi(x)$ is the logistic function $1/(1 + \exp(-x))$. D is a scaling constant, taken as 1 by some writers for convenience and as 1.7 by others (e.g., Birnbaum, 1968) so that the units of the model will approximate those of normal ogive IRT models (Lord, 1952). One may obtain the two-parameter logistic model from (18) by fixing $c_j = 0$, and the one-parameter model (Rasch, 1960) by additionally fixing $a_j = 1$.

Indeterminacies of scale and origin are apparent in (18). If for any scalars m and x we define $\theta^* = m\theta + x$, $b^* = mb + x$, and $a^* = a/m$, then $P(u = 1 \mid \theta^*, a^*, b^*, c) = P(u = 1 \mid \theta, a, b, c)$. In this article we will specify higher-level prior distributions that resolve these indeterminacies.

Rather than obtaining a posterior for **a**, **b**, and **c** directly, we work with the transformed item parameters

$$\alpha_j = \log a_j,$$

$$\beta_j = b_j,$$

and

$$\gamma_j = \log\left(\frac{c_j}{1 - c_j}\right).$$

It is readily inferred that $a_j = \exp \alpha_j$ and $c_j = \Psi(\gamma_j)$. While this formulation does not permit the boundary values of 0 and 1 for c_j, it serves our purposes adequately by allowing c's arbitrarily close to these values. If prior beliefs suggest positive probabilities for zero c values, neither this transformation nor the multivariate normal prior on item parameters will be appropriate; the beta prior on untransformed c's suggested by Swaminathan and Gifford (1981) may be preferable. Nonpositive a's are also disallowed; careful examination of fitted and empirical response curves will obviously be required in applications where faulty items and incorrect keys can occur.

Reparameterization achieves two ends. The first is a more rapid attainment of large-sample results. The impediment against normality represented by the finite range of c, for example, is removed by reexpression in terms of γ. The second is convenience in specifying higher level prior densities. With unrestricted ranges for all parameters, the imposition of multivariate normal priors on parameters within items but independent across items is

not unreasonable. This may be the simplest way to allow for the possibility of dependence among parameters a, b, and c in a population of items.

Letting ξ represent $(\alpha_1, \beta_1, \gamma_1, \ldots, \alpha_n, \beta_n, \gamma_n)$, the estimation equations for item parameters take the form

$$0 = \frac{\partial \log L(\xi, \tau)}{\partial \xi} + \frac{\partial \log p(\xi \mid \eta)}{\partial \xi}. \tag{19}$$

Formulas for the second term appear in the following section. Those for the first term are approximated as

$$\frac{\partial \log L}{\partial \alpha_j} \approx D (\exp \alpha_j)(1 - c_j) \sum_q e_{qj} W_{qj} (X_k - b_j), \tag{20}$$

$$\frac{\partial \log L}{\partial \beta_j} \approx -D(1 - c_j) \sum_q e_{qj} W_{qj} a_j, \tag{21}$$

$$\frac{\partial \log L}{\partial \gamma_j} \approx \frac{c_j \sum_q e_{qj}}{P_{qj}}, \tag{22}$$

where $e_{qj} = \hat{r}_{qj} - \hat{N}_{qj} P_{qj}$ with $\hat{N}_{qj}$ and $\hat{r}_{qj}$ given in (10) and (11) and

$$P_{qj} = c_j + (1 - c_j)\Psi[Da_j(X_k - b_j)],$$

with

$$W_{qj} = \left[\frac{c_j(1 - c_j)}{P_{qj}} + (1 - c_j)^2 \right]^{-1}.$$

Given $\hat{N}_{qj}$ and $\hat{r}_{qj}$, the estimation equations, (19), corresponding to parameters of item j do not involve item parameters or pseudo-data from any other item. This means that the M-step task of finding zeros of (19), along with additional estimation equations for examinee- and possibly item-population parameters, need not address all $3n$ equations for item parameters simultaneously. Zeros for the parameters of a given item within an M-step may be obtained rapidly by methods such as Newton-Raphson iterations, which require second derivatives of the log posterior, or Davidon-Fletcher-Powell iterations, which do not.

Prior Structures on Item Parameters

Let the prior distribution on the parameters for item j be given by $\xi_j = (\alpha_j, \beta_j, \gamma_j) \sim \text{MVN}(\mu_\xi, \Sigma_\xi)$, where $\mu_\xi = (\mu_\alpha, \mu_\beta, \mu_\gamma)$. Hence (μ_ξ, Σ_ξ) plays the role of the item population parameter η in the more general notation of the preceding section. Assuming independence over items, the joint prior log density of item parameters is then given by

$$\log p(\xi \mid \mu_\xi, \Sigma_\xi) \propto -\frac{n}{2} \log |\Sigma_\xi| - \frac{1}{2} \sum_j (\xi_j - \mu_\xi)' \Sigma_\xi^{-1} (\xi_j - \mu_\xi). \tag{23}$$

The partial derivatives of (23) with respect to the parameters for item j are obtained as

$$\frac{\partial \log p(\xi \mid \mu_\xi, \Sigma_\xi)}{\partial \xi_j} = -\Sigma_\xi^{-1} (\xi_j - \mu_\xi). \tag{24}$$

These terms are added to the partial derivatives of the log likelihood of (20) to (22), and the results set to zero to give the estimation equations for the parameters of item j.

In IRT models with independent unimodal prior distributions on item parameters, the contribution of prior information in the estimation equation for a given parameter depends upon its distance from the center of the distribution of parameters of its same type. That is, parameters of a given type "shrink" toward a single point, namely the mean of parameters of that type, by amounts inversely proportional to the information available for each individually. It will be seen in (24) that under the structure proposed here, the contribution of the prior also depends on the distance of the item's parameters of other types from the centers of their respective populations. Parameters of a given type now shrink toward a plane, namely their conditional expectations given the values of the items' parameters of other types.

Let us suppose further that $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ follows the natural conjugate prior distribution for the multivariate normal, namely multivariate normal for $\boldsymbol{\mu}_\xi$ given $\boldsymbol{\Sigma}_\xi$ and inverted Wishart for $\boldsymbol{\Sigma}_\xi$ (Ando & Kaufman, 1965); that is,

$$\log p(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi) \propto -\left|\frac{(m+1)}{2}\right| \log |\boldsymbol{\Sigma}_\xi| - \frac{1}{2}(\boldsymbol{\mu}_\xi - \mathbf{y}_\xi)\boldsymbol{\Sigma}_\xi^{-1}(\boldsymbol{\mu}_\xi - \mathbf{y}_\xi)b - \frac{1}{2}\operatorname{tr} \boldsymbol{\Sigma}_\xi^{-1}\mathbf{H}. \tag{25}$$

Here b and m are scalars ($m > 2p$ for a proper distribution under the p-parameter IRT model), $\mathbf{y}_\xi$ is a vector, and $\mathbf{H}$ is a 3-by-3 positive symmetric matrix—all to be specified in such a way that $\mathbf{H}$ corresponds to the covariance of $m - p$ values of ξ and $\mathbf{y}_\xi$ corresponds to the average of the b values of ξ. A practical example of how these parameters can be specified appears in the analysis of simulated data in a subsequent section.

The indeterminacies of scale and origin in the two- and three-parameter models can be conveniently resolved at this point by specifying their locations; for example, $p(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ is null everywhere except where $\mu_\alpha = 0$ and $\mu_\beta = 0$. Only the latter constraint would enter into the one-parameter model.

If $\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\xi$ are to be estimated jointly with ξ and τ, partial derivatives must first be obtained for all terms in the log posterior in which they appear, namely $\log p(\xi \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$, (23), and $\log p(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$, (25). Differentiating, equating to zero, and simplifying yields the following estimating equations:

$$\boldsymbol{\mu}_\xi = \frac{n\bar{\xi} + b\mathbf{y}_\xi}{n + b}, \tag{26}$$

and

$$\boldsymbol{\Sigma}_\xi = (n + m + 1)^{-1}\{\mathbf{S} + n(\bar{\xi} - \boldsymbol{\mu}_\xi)(\bar{\xi} - \boldsymbol{\mu}_\xi)' + b(\boldsymbol{\mu}_\xi - \mathbf{y}_\xi)(\boldsymbol{\mu}_\xi - \mathbf{y}_\xi)' + \mathbf{H}\},$$

where

$$\bar{\xi} = n^{-1}\sum_j \xi_j,$$

and

$$\mathbf{S} = \sum_j (\xi_j - \bar{\xi})(\xi_j - \bar{\xi})'.$$

A familiar theme in Bayesian estimation appears in (26), where a mean is estimated as a weighted average of a sample mean and a prior mean. It should be pointed out that $\bar{\xi}$ will generally not be equal to the simple mean of the item parameter estimates that would have been obtained under straight marginal maximum likelihood (MML) estimation. This is because the item parameters ξ_j are being estimated at the same time, and each is shrinking back from its MML value in *inverse* proportion to the amount of information

about it; items therefore contribute toward the estimation of the item population mean in *direct* proportion to the amount of information about them.

To achieve maximum likelihood estimation of $\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\xi$, again jointly with $\boldsymbol{\xi}$ and $\boldsymbol{\tau}$, one may specify that $\mathbf{H} = 0$ and $m = 2$, and omit the quadratic term involving $\mathbf{y}_\xi$ in and after (25). This gives an improper diffuse prior, justifiable along the lines of invariance with respect to reparameterization (Jeffreys, 1961). The estimation equations simplify in obvious ways.

If modal values of $\boldsymbol{\xi}$ and $\boldsymbol{\tau}$ marginalized with respect to $\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\xi$ are desired, these latter parameters may be integrated out and then estimation equations for item parameters can be modified in the following manner. Focusing on the relevant terms of the posterior, we can write

$$p(\boldsymbol{\xi} \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi) \cdot p(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi) \propto |\boldsymbol{\Sigma}_\xi|^{-(n+m+1)/2} \times \exp\{-\tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}_\xi^{-1}[\mathbf{S} + \mathbf{H} + n(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_\xi)(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_\xi)' + b(\boldsymbol{\mu}_\xi - \mathbf{y}_\xi)(\boldsymbol{\mu}_\xi - \mathbf{y}_\xi)']\}.$$

Integration over $\boldsymbol{\Sigma}_\xi$ yields a multivariate-t distribution for $\boldsymbol{\mu}_\xi$ (Ando & Kaufman, 1965), and subsequent integration over $\boldsymbol{\mu}_\xi$ yields the marginal distribution of $\boldsymbol{\xi}$:

$$p(\boldsymbol{\xi}) \propto |\mathbf{C}|^{1/2},$$

where

$$\mathbf{C} = \frac{\mathbf{S} + \mathbf{H}}{n+1} + \frac{nb}{(n+b)^2}(\mathbf{y}_\xi - \bar{\boldsymbol{\xi}})(\mathbf{y}_\xi - \bar{\boldsymbol{\xi}})'.$$

The terms to be added to the partial derivative of the log marginal likelihood to obtain an estimation equation for ξ_j, now marginalized with respect to $\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\xi$, become

$$\frac{\partial \log p(\boldsymbol{\xi})}{\partial \xi_j} = -\frac{\mathbf{C}^{-1}}{n+1}\left[\xi_j - \bar{\xi} + \frac{\bar{\boldsymbol{\xi}} - \mathbf{y}_\xi}{n+1}\right].$$

This result is similar in form to (24), the contribution when $\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\xi$ are estimated jointly with $\boldsymbol{\xi}$.

Prior Structures on Examinee Parameters

This section presents details for two types of prior distributions on examinee parameters $\boldsymbol{\tau}$, namely a nonparametric prior in the form of a histogram and a mixture of homoscedastic normal distributions in unknown proportions. The latter choice includes the familiar standard normal prior as a special case (in which case no scale-setting restrictions would be imposed on the item parameter distributions).

Recalling the form of the posterior distribution for $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and $\boldsymbol{\tau}$, or

$$p(\boldsymbol{\xi}, \boldsymbol{\tau}, \boldsymbol{\eta} \mid \mathbf{D}, \mathbf{U}) = \left\{\int_{\boldsymbol{\theta}} L(\mathbf{U} \mid \mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta} \mid \boldsymbol{\tau})\, d\boldsymbol{\theta}\right\} \cdot \{p(\boldsymbol{\tau})\} \cdot \{p(\boldsymbol{\xi} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta})\},$$

we note that (a) contributions to the estimation equations for $\boldsymbol{\tau}$ come from its prior and the marginal likelihood and (b) these contributions are the same regardless of whether $\boldsymbol{\eta}$ is being estimated jointly or integrated out. In the interest of conserving space, only estimation equations and not partial derivatives for $\boldsymbol{\tau}$ are presented here. Detailed calculations of the contributions from the marginal likelihood may be found in Mislevy (1984).

A nonparametric solution. If $p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ is a smooth continuous density it may be approximated by a discrete distribution over a finite number of points X_q, $q = 1, \ldots, Q$. Letting p_q denote the density at point X_q, we approximate the log marginal likelihood as

$$\log L(\mathbf{U} \mid \mathbf{D}, \xi, \tau) \approx \sum_{i=1}^{N} \log h(\boldsymbol{\mu}_i)$$

where

$$h(\boldsymbol{\mu}_i) = \sum_{q=1}^{Q} L(\boldsymbol{\mu}_i \mid \mathbf{d}_i, X_q, \xi) p_q.$$

The continuous density $p(\boldsymbol{\theta} \mid \tau)$ is thus replaced by a discrete distribution with parameters $p_1, \ldots, p_Q$, with

$$\sum_{q=1}^{Q} p_q = 1.$$

The natural conjugate prior for this distribution is the Dirchlet, which takes the following form:

$$p(p_1, \ldots, p_Q \mid M_1, \ldots, M_Q) \propto \prod_k p_k^{M_k}.$$

Prior beliefs about $p_1, \ldots, p_Q$ are thus expressed as values of the proportions $(M_1)/M^+, \ldots, (M_Q)/M^+$, where $M^+ = \sum M_q$, is interpretable as a fictitious sample size to indicate the strength of prior beliefs. The following estimation equations, based on $\partial[\log L(\mathbf{M} \mid \mathbf{D}, \xi, \mathbf{p}) + \log p(\mathbf{p} \mid \mathbf{M})]/\partial \mathbf{p} = \mathbf{0}$, result:

$$p_q = \frac{\hat{N}_q + M_q}{N + M^+}, \qquad q = 1, \ldots, Q,$$

where

$$\hat{N}_q = \sum_i p(X_q \mid \mathbf{u}_i, \mathbf{d}_i, \xi, \mathbf{p})$$

$$= \frac{\sum_i L(\mathbf{u}_i \mid \mathbf{d}_i, \mathbf{X}_q, \xi, \mathbf{p}) p_q}{\sum_r \sum_i L(\mathbf{u}_i \mid \mathbf{d}_i, X_r, \xi, \mathbf{p}) p_r}.$$

The posterior density at point X_q, therefore, is a weighted average of its prior density and the expectation of its density at that point conditional on the data and the densities themselves.

To obtain maximum likelihood estimates, we may take a uniform diffuse prior with $M_q \equiv 0$. An alternative diffuse prior with $M_q \equiv -1$ may be preferred, however, on the grounds of robustness with respect to the choice of quadrature points.

It is possible to resolve the indeterminacies of the IRT model at this point, by specifying that the distribution $p(\mathbf{p} \mid \mathbf{M})$ can take nonzero values only when the following equality constraints are satisfied:

$$\sum_q X_q p_q = 0,$$

and

$$\sum_q X_q^2 p_q = 1.$$

Values of $\mathbf{M}$ specified in an informative prior should satisfy these constraints as well.

A mixture of normal components. Suppose that the distribution is a mixture of K normal components, with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and common variance σ^2. Let $\mathbf{p} = (p_1, \ldots, p_K)$ be the unknown proportions of the mixture. Define the marginal probability of

response pattern $\mathbf{u}$ given ξ and $\underline{\tau} \equiv (\boldsymbol{\mu}, \mathbf{p}, \sigma^2)$ as

$$h(\mathbf{u}) = \sum_p p_k \int_\theta L(\mathbf{u} \mid \mathbf{d}, \boldsymbol{\theta}, \xi, \tau) f_k(\theta)\, d\theta,$$

where

$$f_k(\theta) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[\frac{-(\theta - \mu_k)^2}{2\sigma^2}\right].$$

Approximating integration by summation over a fixed grid of equally-spaced quadrature points $X_1, \ldots, X_Q$, we obtain for the log marginal likelihood for N examinees

$$\log L(\mathbf{U} \mid \mathbf{D}, \xi, \tau) \approx \sum_i \log \sum_k p_k \sum_p L(\mathbf{u}_i \mid X_q) f_k(X_q),$$

where

$$L(\mathbf{u}_i \mid X_q) = L(\mathbf{u}_i \mid \mathbf{d}_i, X_q, \xi, \tau).$$

A natural conjugate prior for τ is Dirichlet-normal-inverse gamma:

$$\log p(\mathbf{p}, \boldsymbol{\mu}, \sigma^2) = \sum_k M_k \log p_k + \sum_k \frac{-(\mu_k - y_k)^2}{2\sigma^2} - \left(\frac{v}{2} + 1\right) \log \sigma - \left(\frac{s}{2\sigma^2}\right).$$

Here $\mathbf{M}$, $\mathbf{y}$, v, and s are the parameters of the prior distribution, to be supplied by the user. $\mathbf{M}$ can be thought of as the number of examinees in each of the components from a sample of size $M^+ = \sum M_k$; $\mathbf{y}$ can be thought of as anticipated locations for the means of the components. v and s are the parameters of the inverted gamma distribution, possibly more easily specified after one has in mind a mean and variance of such a distribution that incorporates prior belief about σ^2:

$$v = \frac{2\ \text{mean}^2}{\text{variance}} + 4,$$

and

$$s = \frac{\text{mean} \cdot \text{variance}}{2(\text{mean} + \text{variance}^2)}.$$

The indeterminacies of the IRT model can also be resolved at this point, by specifying that the total mean and within-component variance take specified values, say 0 and 1. That is, $p(\tau)$ is zero except where

$$\sum_k p_k \mu_k = 0,$$

and

$$\sigma^2 = 1.$$

When $K = 1$, a standard normal density is effectively specified for θ by this procedure.

Estimation equations are now obtained as the sums of partial derivatives of the log marginal likelihood and the log prior. Again writing equations in terms of K p's constrained to a sum of one, we obtain

$$p_k = \frac{\sum_q \hat{N}_{kq} + M_k}{N + M^+}, \qquad k = 1, \ldots, Q,$$

$$\mu_k = \frac{\sum_q \hat{N}_{kq} X_q + y_k}{\sum_q \hat{N}_{kq} + 1},$$

and

$$\sigma^2 = \frac{\sum_k \sum_q \hat{N}_{kq}(X_q - \mu_k)^2 + \sum_k (\mu_k - y_k)^2 + s}{N + \left(\frac{v}{2} + 1\right)}.$$

where

$$\hat{N}_{kq} = \sum_i h^{-1}(\mathbf{u}_i) p_k L(\mathbf{u}_i \mid X_q) f_k(X_q).$$

A diffuse prior may be obtained by omitting terms involving $\boldsymbol{\mu}$ and setting $M_k \equiv 1$, $s = 0$, and $v = 0$.

A Numerical Example

Satisfactory procedures for item parameter estimation have been available for some time for both large and small samples under the one-parameter logistic (Rasch) IRT model and for large samples of both persons and items under the three-parameter logistic (Birnbaum) IRT models. The same cannot be said about moderate-to-small samples under the three-parameter model, and it is to this problem we apply the procedures of the preceeding sections.

A perusal of the recent literature on Bayesian item parameter estimation suggests that such efforts were motivated not so much by the pursuit of minimum mean squared error or by a conviction that all unknowns should be expressed in probabilistic terms, but rather by a more practical desire to obtain "reasonable" item parameter estimates—in particular, finite ones.

The essential difficulty with parameter estimation under the three-parameter model is that the parameters of a given item are often poorly determined by the data at hand; apparently discrepant triples (a, b, c) can trace similar response curves in the region of the ability scale where the sample of examinees is to be found. Such poor resolution is manifest as a likelihood surface nearly flat along one or more dimensions, yielding unstable maximum likelihood estimates (MLE's). A trivially higher likelihood may be produced, for example, by taking a particular item's values of a and c to be 200 and .6 rather than the more reasonable values of 2 and .25.

Extreme and infinite parameter estimates can be avoided by using a single-stage Bayesian prior, but not without introducing an additional hazard. A fully-specified prior will indeed have the desired effect of pulling extreme but ill-determined values toward the center of the prior distribution. If the prior has been poorly specified, however, this center may be far from the actual center of the parameter values of interest; estimates of all such parameters will be biased in the same direction. These "ensemble biases" have serious implications for subsequent estimation of examinee individual or population parameters, for while such estimation is resistent to *random* errors in item parameters, it reflects in direct measure *systematic* errors in a's and b's, and, through the systematic errors in a's and b's they cause, systematic errors in c's as well.

As a means of overcoming these difficulties, second stage priors may be introduced. Their forms and parameters will be influenced by a number of factors, including not only the analyst's prior beliefs about the values of the parameters in question but considerations of convenience and expediency as well. The rationale behind the "floating priors" option for item parameters in the BILOG computer program (Mislevy & Bock, 1982), serves to illustrate the process.

The author's experience with the full multivariate normal prior on transformed item parameters has indicated that even samples consisting of a few hundred examinees and ten items provide sufficient information to approximate the central tendencies of item parameters through $\boldsymbol{\mu}_\xi$, so that its prior may be diffuse. Much larger sample sizes (thousands of examinees, fifty or more items) are required, however, to estimate the associated covariance matrix $\boldsymbol{\Sigma}_\xi$; its prior must therefore be informative. Informative, and strong. Anomalies such as estimated zero variances for α and γ occur increasingly often with weaker priors. Fixed values of the diagonal elements of $\boldsymbol{\Sigma}_\xi$ are thus posited, at user-specified values or at the program defaults. Zero off-diagonals are posited for computational convenience. The consequence is that parameter estimates tend to shrink toward their centers at user-controlled rates, but those centers are estimated from the data.

The program default values for $\Sigma_{\alpha\alpha}$, $\Sigma_{\beta\beta}$, and $\Sigma_{\gamma\gamma}$ are .25, 4.00, and .25 respectively. These values are sufficiently mild to affect most item parameters minimally when the data supply information about them, but keep all parameters within a "reasonable" range. If the geometric average of the slopes were one, for example, then $\Sigma_{\alpha\alpha} = .25$ would mean we would expect most estimated a's to lie between 1/3 and 3 (about two standard deviations in the log metric). And if the logit average of lower asymptotes were .25, then $\Sigma_{\gamma\gamma} = .25$ would mean we would expect most c's to lie between .15 and .35. Information in the data can produce occasional individual estimates outside these ranges, but infinite values and extreme values for parameters for which the data provide little information will be eliminated. The prior variance of 4.00 on β's, which are better determined that α's or γ's, usually has little effect but to pull in a few extreme values that would have resulted from unusually low a's.

Some of the effects of these second-stage priors can be seen in two analyses of a simulated data set, with responses of 1000 simulated examinees, selected at random from a unit normal population, to 20 test items. The parameters of the items were also generated from independent normal distributions; for the $\alpha = \log a$, the mean and variance were 0.0 and .5; for $\beta = b$, .5 and 1.0; and for $\gamma = \text{logit } c$, -1.39 and .16. Item parameters were estimated in two ways.

1. Marginal maximum likelihood (MML): Using the BILOG computer program, the following likelihood equation was maximized with respect to item parameters ξ and weights p_q at ten equally spaced quadrature points X_q between -4 and $+4$:

$$L = \prod_i \sum_q p(\mathbf{u}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, X_q) p_q .$$

2. Bayes estimation: To obtain Bayes modal estimates of item parameters, a posterior of similar form was maximized:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{p}, \boldsymbol{\mu}_\xi \mid \mathbf{U}, \boldsymbol{\Sigma}_\xi) = \prod_i \sum_q p(\mathbf{u}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, X_q) \cdot p_q \cdot p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi).$$

BILOG's "floating priors" option (described above) was employed to set the values of $\boldsymbol{\Sigma}_\xi$.

The value of $-2 \log L$ under the MML solution was found to be 22,295, while the value obtained by substituting the Bayes estimates into the likelihood function was 22,300. This trivial difference implies that the Bayes estimates explain the observed data nearly as well as the MML estimates. Indeed, with a few exceptions (more on these

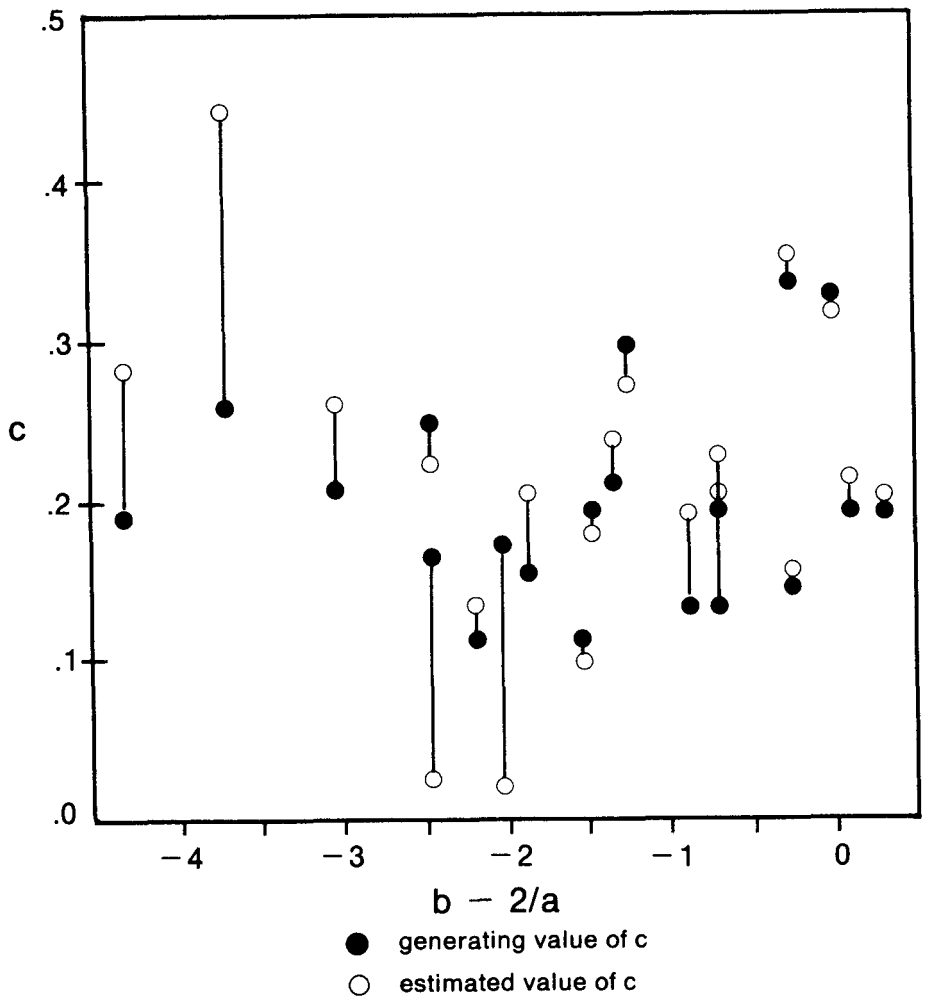

FIGURE 1
Generating and MML estimated values of c, against generating $b - 2/a$.

below), MML and Bayes estimates of α and β were quite similar, with α's tending to shrink slightly toward their estimated mean of .21.

Estimates of asymptotes were more significantly affected, as can be seen in Figures 1 and 2. These figures plot generating and estimated values of c, MML and Bayes solutions respectively, against generating values of the quantity $b - 2/a$, a heuristic index based on the observation that less information is obtained about c as items become easier or less reliable (Lord, 1975). Items with high values of this index are seen to have estimated c's near their generating values under both estimation procedures, but certain items with low values are regressed strongly toward the estimated mean of about .21. To anthropomorphize, we might say that the Bayes solution felt true c's for these items were probably more similar to the c's that it could estimate well than to the atypical and unstable MML values based on sparse information.

It is instructive to consider the estimated a's and b's of these items, to see how item parameters can "trade off" against one another. Values for the six items showing the largest differences between MML and Bayes estimated c's are shown in Table 1.

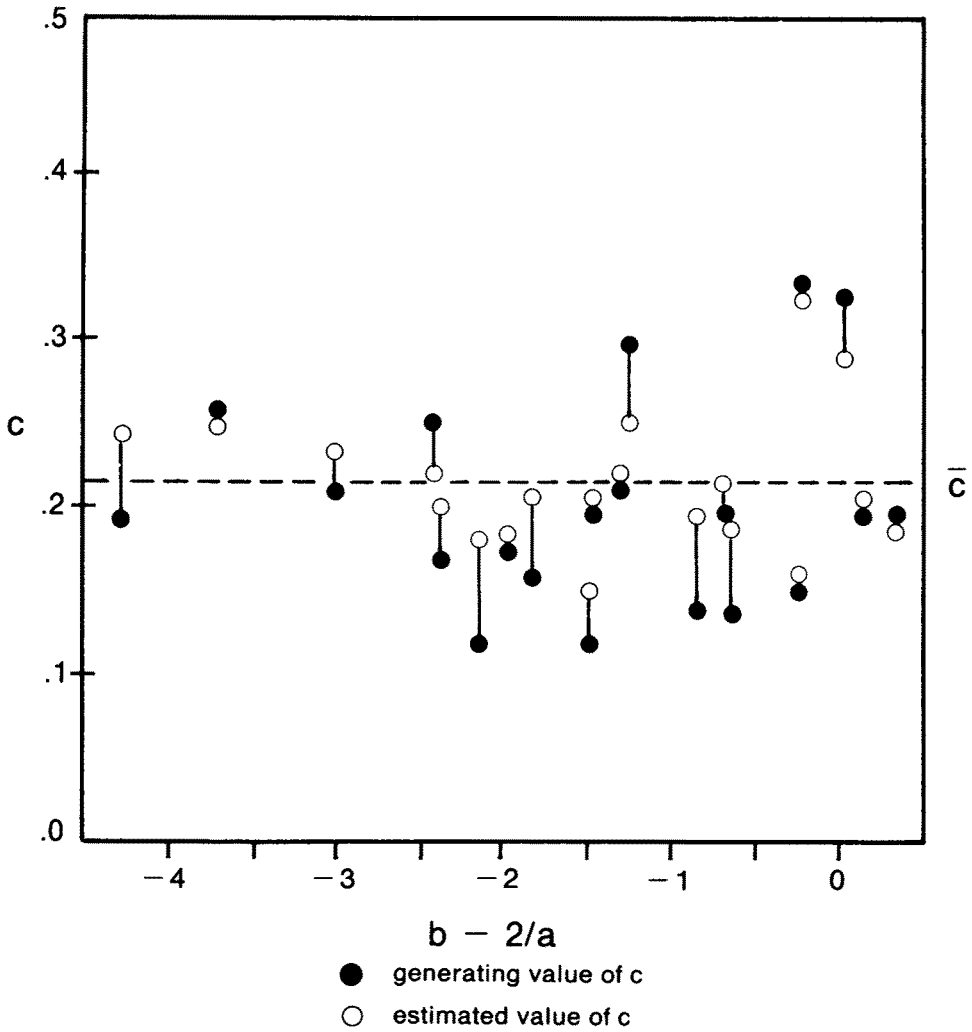

FIGURE 2
Generating and Bayes estimated values of c, against generating $b - 2/a$.

Item 1 in table 1 is relatively easy, so that the increased c value obtained by the Bayes solution has little effect on the estimated a and b. As it turns out, the generating c for this item was lower and more atypical than either MML or Bayes obtained, but since most of the examinees were well above the chance level, it did not really matter. Item 4 is similar, in that a large degree of shrinkage of the estimated c on an easy items has little effect on the other parameters. This time (and, the model assumes, more often than not) the Bayes estimate is closer to the true value.

Item 2 shows an extremely high c under MML shrunken back by Bayes procedures to a lower, more nearly correct, value. While the estimated a's are similar, the estimated b under Bayes is correspondingly reduced somewhat, again closer to its true value. The point here is that spuriously over-estimated c's induce spuriously over-estimated b's, a result guarded against in two ways when priors are enforced on both parameters.

Items 3 and 6 show items with high MML a estimates being shrunken back toward their mean under Bayes, and extreme c's correspondingly regressed. Both items are rela-

TABLE 1

Generating and Estimated Parameters of Selected Items

	a			b			c		
Item	True	MML	Bayes	True	MML	Bayes	True	MML	Bayes
1	1.1	1.2	1.3	-.4	-.4	-.3	.11	.14	.17
2	.5	.4	.4	.2	.8	.6	.19	.28	.24
3	.9	1.5	1.1	-1.3	-.6	-1.0	.26	.44	.27
4	1.4	1.2	1.4	-1.0	-1.2	-1.0	.17	.03	.19
5	1.5	2.2	2.4	-.3	-.2	-.2	.13	.12	.14
6	2.5	4.5	3.4	-1.1	-1.2	-1.1	.18	.03	.18

tively easy, but it is seen that pulling down a spuriously high c (item 3) affects b whereas increasing a spuriously low c (item 6) does not.

Finally, Item 5 shows an atypically low c regressing toward its mean, causing a corresponding shift in *a away* from its mean. The estimated b's are similar under both models.

Discussion

Maximum likelihood (ML) estimation is justified by its asymptotic properties alone. Taking the data for each parameter at face value no matter how sparse, ML will often yield infinite or implausible parameter estimates in small samples. At least for certain parameters, a sample size of 10,000 examinees can be a small sample in the context of the three-parameter logistic IRT model (Wainer & Thissen, 1982); estimation procedures therefore stand to profit from the incorporation of additional information. The hierarchical Bayesian framework given in the previous sections supplies such information in a very modest way. In effect, it quantifies beliefs such as the following.

1. If the items for which we can reasonably estimate c's yield values between .1 and .3, then the items for which less information is available probably have c's in this range as well.

2. If most of the items have a's between 1/3 and 3, then the a for this particular item is probably not 957.

3. If all of the other examinees seem to have θ's between -3 and $+3$, the θ for this examinee is probably not $+\infty$, even though he did correctly answer both items he was presented.

Such strictures are implied by the assumption that parameters belong to respective well-behaved populations, the higher-level parameters of which little or nothing needs to be assumed. The effect of this so-called assumption of exchangeability is to "shrink" estimates from where they would have been under ML toward the centers of the respective populations—with those centers estimated from the data.

When it is not reasonable to assume a common population, however, exchangeability is violated. Graphic examples of the absurdities that can result are suggested by proponents as well as critics of "shrunken" estimators. Should one expect to obtain better estimates of the true batting averages of baseball players, for instance, by including data on

the price of wheat? The point is that shrinking estimates toward a common center is justified only when a common population best represents the extent of our prior knowledge. The imposition of exchangeability across all units, and estimation procedures that require it, are not strictly appropriate when additional information differentiating the units is at hand. Assuming exchangeability only within distinct subpopulations of batting averages and wheat prices would be a more reasonable way to proceed in the example given above.

It is in fact this latter case that typically prevails in educational and psychological measurement. Already known, or available more economically than responses from examinees, is information from several sources:

1. Cognitive processing requirements of items can be specified, at least to some degree. Mental rotation items, for example, can be characterized in terms of the number of degrees the target object has been rotated; differential calculus items, an example discussed by Fischer (1973), can be characterized in terms of the derivation rules they demand for solution.

2. Surface features of items can be identified which can suggest a need for distinguishing subpopulations of items. Free-response and multiple-choice items in the same test may be distinguished, for example, as may be analogy items from vocabulary items in the SAT.

3. Item content can be often be identified. In a test of reading comprehension, one might wish to differentiate items associated with narrative passages, poetry, and documents.

4. Quantitative information, such as percents-correct from pretesting, may be available.

5. Examinees may be differentiated with respect to qualitative features such as sex, educational program, or racial/ethnic background; or with respect to quantitative variables such as scores on previously administered tests.

More comprehensive Bayesian procedures currently under investigations by the present writer and others (e.g., Leonard & Novick, 1985; Morgan, 1985), will provide for the utilization of such information. They will also provide for means of determining when such information makes material differences in item and population parameter estimates.

References

Andersen, E. B. (1973). *Conditional inference and models for measuring.* Copenhagen: Danish Institute for Mental Health.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. *Psychometrika, 42,* 357–374.

Ando, A., & Kaufman, O. M. (1965). Bayesian analysis of the independent normal process—neither mean nor precision known. *Journal of the American Statistical Association, 60,* 347–358.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

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

Dunsmore, I. R. (1976). Asymptotic predictor analysis. *Biometrika, 63,* 627–630.

Efron, B., & Norris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70,* 311–319.

Fisher, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Hartigan, J. A. (1983). *Bayes theory.* New York: Springer-Verlang.

James, W., & Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Probability and Statistics* (Vol. 1). Berkeley: University of California Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.) Oxford: Clarendon Press.

Kelley, T. L. (1927). *The interpretation of educational measurements.* New York: World Press.

Leonard, T., & Novick, M. R. (1985). *Bayesian inference and diagnostics for the three-parameter logistic model* (ONR Technical Report 85-5). Iowa City, IA: University of Iowa.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known item response functions.* Paper presented at the meeting of the Psychometric Society, Nashville, TN.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society* (Series B), *34*, 1–41.

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph No. 7, 17*(4, Pt. 2).

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (RB-75-33). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

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

Mislevy, R. J. (in preparation). *Stochastic test designs* (Research Report). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1985). Implementation of an *EM* algorithm in the estimation of item parameters. In D. J. Weiss (Ed.), *Proceedings of the IRT/CAT Conference.* Minneapolis, MN: Computerized Adaptive Testing Laboratory, University of Minnesota.

Morgan, G. (1985). *Computing approximate marginal distributions in the Rasch logistic test model using prior information.* Paper presented at the annual meeting of the American Educational Research Association in Chicago, IL.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1–32.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. (1972). Estimating multiple regressions in *m*-groups: A cross-validational study. *British Journal of Mathematical and Statistical Psychology, 5*, 33–50.

O'Hagan, A. (1976). On posterior joint and marginal modes. *Biometrika, 63*, 329–333.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reiser, M. R. (1981, June). *Bayesian estimation of item parameters in the two-parameter logistic model.* Paper presented at the annual meeting of the Psychometric Society in Chapel Hill, NC.

Rigdon, S., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48*, 567–574.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. *Journal of the American Statistical Society, 75*, 801–827.

Sanathanan, L., & Blumenthal, N. (1978). The logistic model and latent structure. *Journal of the American Statistical Association. 73*, 794–798.

Stroud, A. H., & Sechrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice-Hall.

Swaminathan, H., & Gifford, J. A. (1981). *Bayesian estimation in the three-parameter logistic model.* Paper presented at the annual meeting of the Psychometric Society, Chapel Hill, NC.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175–192.

Swaminathan, H., & Gifford, H. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349–364.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Wainer, H., & Thissen, D. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397–412.

Manuscript received 5/6/85
Final version received 12/9/85