PSYCHOMETRIKA-VOL. 46, NO. 4.
DECEMBER, 1981.

MARGINAL MAXIMUM LIKELIHOOD ESTIMATION OF ITEM PARAMETERS: APPLICATION OF AN EM ALGORITHM

R. DARRELL BOCK

UNIVERSITY OF CHICAGO

MURRAY AITKIN

UNIVERSITY OF LANCASTER

Maximum likelihood estimation of item parameters in the marginal distribution, integrating over the distribution of ability, becomes practical when computing procedures based on an EM algorithm are used. By characterizing the ability distribution empirically, arbitrary assumptions about its form are avoided. The EM procedure is shown to apply to general item-response models lacking simple sufficient statistics for ability. This includes models with more than one latent dimension.

Key words: estimation of item parameters, EM algorithm, item analysis, latent trait, dichotomous factor analysis, Law School Aptitude Test (LSAT).

Despite recent advances in item-response theory [Lord, 1977, 1980], practical application of the results has been hampered by the lack of a proven and economical method of estimating the parameters in general item-response models. Although conditional maximum likelihood estimation is available for the one-parameter logistic binary response model and, with certain qualifications, for the corresponding rating scale model [see Andersen, 1973, 1980], this method cannot be applied to the two- and three-parameter logistic or normal ogive models or to the multiple category models of Samejima [1969] or Bock [1972]. Broad progress in this area of psychometrics clearly requires a less restrictive approach to item parameter estimation.

The difficulty with the general models is that the subject's ability, which appears as a nuisance parameter, cannot be eliminated from the likelihood by conditioning on a sufficient statistic in the manner proposed for the one-parameter logistic by Rasch [see Andersen, 1980, chap. 6]. Moreover, joint maximum likelihood estimation of the subject and item parameters is not generally possible because the number of parameters increases with the number of subjects and standard limit theorems do not apply. Haberman [1977] has shown that consistent estimates of the Rasch difficulty parameters are attained by the joint maximum likelihood method as both the number of items and the number of subjects increase without limit, but this condition is not realistic in practical settings where the number of items is finite and often small. Various workers, including Wright and Panchapakesan [1969], Wood, Wingersky, and Lord (1976), and Kolakowski and Bock [1973 a & b], have avoided this problem by assuming that subjects who have the same number-right score, or the same item-score pattern, or who have been assigned provisonally to homogeneous ability groups, have the same ability. On this assumption, the number of parameters is finite and standard asymptotics apply. Nevertheless, these so-called "fixed-effect" solutions,

Supported in part by NSF grant BNS 7912417 to the University of Chicago and by SSRC (UK) grant HR6132 to the University of Lancaster.

We are indebted to Mark Reiser and Robert Gibbons for computer programming. David Thissen clarified a number of points in an earlier draft.

Requests for reprints should be addressed to R. Darrell Bock, Department of Behavioral Sciences, The University of Chicago, 5848 South University Avenue, Chicago, Illinois, 60637.

0033-3123/81/1200-3031$00.75/0

although generally giving reasonable estimates of the item parameters, may be subject to bias or even become unstable when the number of items is small [Thissen, Note 1]. In any event, the assumption that abilities are fixed parameters and are finite in number, when in fact they are not identifiable and have a distribution in the population of subjects, is difficult to justify from a statistical point of view.

A better approach to estimation in the presence of a random nuisance parameter is that of integrating over the parameter distribution and estimating the structural parameters by maximum likelihood in the marginal distribution (marginal maximum likelihood, MML). Working with the two-parameter normal ogive model, Bock and Lieberman [1970] have already taken this approach and have estimated item thresholds and factor loadings on the assumption that subjects are a random sample from an $N(0, 1)$ distribution of ability. Using Gauss-Hermite quadrature to perform the necessary integration, they obtained stable estimates of these parameters for as few as five items.

Although applicable to virtually any type of item response model (Bock, 1972, has also applied it to the logistic, multiple nominal categories model), the Bock-Lieberman method for a number of reasons does not seem practical for general use. Computationally, the method is unattractive because the Newton-Raphson method used to solve the MML equations requires, for n items, the generation and inversion of a $2n \times 2n$ information matrix, each element of which is a sum of 2^n terms. Because this matrix must be generated and inverted four or five times in the Newton iterations, this method is not practical at present, even on the fastest computers, when n exceeds 12. From a statistical point of view, the method is also objectionable because it assumes that the form of the distribution of ability effectively sampled is known in advance. Since item calibration studies are typically carried out on arbitrarily selected samples, it is difficult to specify *a priori* the distribution of ability in the population effectively sampled.

In the present paper we show that, by a simple reformulation of the Bock-Lieberman likelihood equations, a computationally feasible solution is possible for both small and large numbers of items. The method obtained by this reformulation is closely related to the EM algorithm for maximum likelihood estimation as discussed by Dempster, Laird and Rubin [1977].

This formulation of the likelihood equations also makes it clear that the form of the distribution of ability does not necessarily need to be known beforehand. Instead, it can be estimated as a discrete distribution on a finite number of points (i.e., a histogram). The item parameters can then be estimated by integrating over the empirical distribution, thus freeing the method from arbitrary assumptions about the distribution of ability in the population effectively sampled.

Two other generalizations of the procedure also appear promising. The method can be extended to include estimation of structural models for facet designs on item operations, content, format, etc., as discussed by Micko [1969], Scheiblechner [1972], and Fischer [1973, 1980]. Perhaps most remarkable, the method applies straightforwardly to more than one latent dimension of ability. It can be used to estimate factor loadings on each dimension, thus providing full-information factor analysis of dichotomous and polytomous data [Bartholomew, 1980].

EM Solution of the Marginal Likelihood Equations

We follow Bock and Lieberman in assuming a normal cdf for the item response function. Let $x_{ij} = 1$ if subject i responds correctly to item j, and $x_{ij} = 0$ otherwise. Then

$$P(x_{ij} = 1 \mid \theta_i) = \Phi_j(\theta_i) = \frac{1}{(2\pi)^{1/2}} \int_{-z_j(\theta_i)}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt, \tag{1}$$

and

$$P(x_{ij} = 0 \mid \theta_i) = 1 - \Phi_j(\theta_i),$$

where

$$z_j(\theta_i) = a_j(\theta_i - b_j). \tag{2}$$

The "slope", a_j, and "threshold", b_j, in (2) are the so-called "invariant" item parameters (Lord & Novick, 1968, p. 353). For computational purposes, it is convenient to employ the item-intercept parameter, $c_j = -a_j b_j$, and write

$$z_j(\theta_i) = c_j + a_j \theta_i. \tag{3}$$

On the usual conditional independence assumption, the probability of subject i responding in pattern $\mathbf{x}_i = [x_{i1}\ x_{i2} \cdots x_{in}]$, conditional on ability θ_i, is

$$P(\mathbf{x} = \mathbf{x}_i \mid \theta_i) = \prod_j^n [\Phi_j(\theta_i)]^{x_{ij}}[1 - \Phi_j(\theta_i)]^{1 - x_{ij}}. \tag{4}$$

For a random subject sampled from a population with a continuous ability distribution $g(\theta)$, the unconditional probability is

$$P(\mathbf{x} = \mathbf{x}_i) = \int_{-\infty}^{\infty} P(\mathbf{x} = \mathbf{x}_i \mid \theta) g(\theta)\, d\theta \tag{5}$$

This probability can be approximated to any practical degree of accuracy by Gauss-Hermite quadrature, i.e., by the sum

$$\sum_k^q P(\mathbf{x} = \mathbf{x}_i \mid X_k) A(X_k),$$

where X_k is a tabled quadrature point (node) and $A(X_k)$ is the corresponding weight (see Stroud & Sechrest, 1966).

Now let the item score patterns observed in a random sample of N subjects be indexed by $l = 1, 2, \ldots, s$, where $s \leq \min(N, 2^n)$. The number of subjects who respond in pattern l is then denoted r_l, where

$$\sum_l^s r_l = N.$$

Since the counts of observed patterns effectively assign each subject to one and only one of 2^n categories, the frequencies r_l are multinomially distributed with parameters N and $P_l = P(\mathbf{x} = \mathbf{x}_l)$. The log likelihood is, therefore,

$$\log L = C + \sum_l^s r_l \log P_l, \tag{6}$$

where C does not depend upon the item parameters.

Bock and Lieberman take the standard normal pdf for $g(\Theta)$ and employ Gauss-Hermite quadrature to approximate the likelihood equation for an item parameter u_j by the sum

$$\frac{\partial \log L}{\partial u_j} \cong \sum_l^s (-1)^{x_{ij}+1} \frac{r_l}{\tilde{P}_l} \sum_k^q \frac{\partial z_j(X_k)}{\partial u_j} \left[\prod_{h \neq j}^n [\Phi_h(X_k)]^{x_{ih}} [1 - \Phi_h(X_k)]^{1 - x_{ih}} \phi_h(X_k) \right] A(X_k) = 0, \tag{7}$$

where

$$\tilde{P}_l = \sum_k^q \left[\prod_j^n [\Phi(X_k)]^{x_{lj}} [1 - \Phi(X_k)]^{1 - x_{lj}} \right] A(X_k)$$

$$= \sum_k^q L_l(X_k) A(X_k). \tag{8}$$

In (8), $L_l(x_k)$ is the conditional probability of $\mathbf{x}_l$ given $\theta = X_k$. In the standard tables and subroutines for Gauss-Hermite quadrature, Gauss's form of the error function is used rather than the normal pdf. For present purposes, we therefore multiply the tabled points by $(2)^{1/2}$ and the weights by $1/(\pi)^{1/2}$ (see Bock & Lieberman, 1970).

For small n, (7) can be solved to any degree of accuracy by Newton-Raphson iterations using the information matrix in place of the matrix of second derivatives:

$$I\begin{bmatrix} c_j \\ a_j \end{bmatrix}_{2m \times 2m} = N \sum_h^{2^n} \frac{1}{\tilde{P}_h} \left[\frac{\partial P_h}{\partial \begin{bmatrix} c_j \\ a_j \end{bmatrix}} \cdot \frac{\partial P_h}{\partial \begin{bmatrix} c_{j'} \\ a_{j'} \end{bmatrix}} \right], \qquad \begin{matrix} j = 1, 2, \ldots, n \\ j' = 1, 2, \ldots, n \end{matrix} \tag{9}$$

(see Rao, 1973, p. 370). Unfortunately, the sum in (9) runs over all 2^n possible patterns and not just the s patterns realized in the sample. This limits a direct Newton-Raphson solution to a small number of items.

In preparation for the alternative formulation of these likelihood equations, we recall that, if the subjects are assumed to be grouped into homogeneous groups, each with a distinct ability value X_k, the intercept and slope parameters can be estimated separately for each item by conventional probit analysis. The likelihood equations for c_j and a_j in this case are, respectively,

$$\sum_k^q \frac{r_{jk} - N_k \Phi_j(X_k)}{\Phi_j(X_k)[1 - \Phi_j(X_k)]} \phi_j(X_k) \frac{\partial z_j(X_k)}{\partial c_j} = 0 \tag{10}$$

and

$$\sum_k^q \frac{r_{jk} - N_k \Phi_j(X_k)}{\Phi_j(X_k)[1 - \Phi_j(X_k)]} \phi_j(X_k) \frac{\partial z_j(X_k)}{\partial a_j} X_k = 0. \tag{11}$$

The information matrix is

$$I\begin{bmatrix} c_j \\ a_j \end{bmatrix} = N \sum_k^q W_j(X_k) \begin{bmatrix} 1 & X_k \\ X_k & X_k^2 \end{bmatrix},$$

where

$$W_j(X_k) = \frac{\phi_j(X_k)}{\Phi_j(X_k)[1 - \Phi_j(X_k)]}.$$

The conditions under which (10) and (11) have a solution, which are quite general, have been discussed by Bock and Jones (1980 reprint).

Bearing in mind (10) and (11), we observe that, in terms of the "weights" $l(X_k)$ in (8), (7) may be expressed as

$$\sum_l^s \frac{r_l}{\tilde{P}_l} \sum_k^q \left[\frac{x_{lj}}{\Phi_j(X_k)} - \frac{(1 - x_{lj})}{1 - \Phi_j(X_k)} \right] \phi_j(X_k) \frac{\partial z_j(X_k)}{\partial u_j} L_l(X_k) A(X_k)$$

$$= \sum_k^q \frac{\sum_l^s r_l x_{lj} L_l(X_k) / \tilde{P}_l - \left[\sum_l^s r_l L_l(X_k) / \tilde{P}_l \right] \Phi_j(X_k)}{\Phi_j(X_k)[1 - \Phi_j(X_k)]} \phi_j(X_k) \frac{\partial z_j(X_k)}{\partial u_j} A(X_k)$$

$$= 0. \tag{12}$$

Comparing (12) with (10) and (11) reveals it to be the likelihood equation of a probit analysis for item j in which

i) X_k is the value of the independent variable at level k,

ii) $\sum_l^s r_l x_{lj} L_l(X_k)A(X_k)/\tilde{P}_l = \bar{r}_{jk}$ is the "expected frequency" of correct response to item j at level k given x_{lj} and $\Phi_j(X_k)$ (the sum of these quantities with respect to k is the number of correct responses for item j), and

iii) $\sum_l^s r_l L_l(X_k)A(X_k)/\tilde{P}_l = \bar{N}_k$ is the "expected sample size" at level k given $\Phi_j(X_k)$ (the sum of these quantities with respect to k is N).

Repeated applications of (12) over the set of items constitutes an EM solution of (7). The numerical procedure consists of two steps per item per cycle:

E-step. For provisional c_j and a_j, compute $L_l(X_k)$, $k = 1, 2, \ldots, q$, and $\tilde{P}_l = \sum_k^q L_l(X_k)A(X_k)$ for pattern l, $l = 1, 2, \ldots, s$. Accumulate $\bar{r}_{jk}$ and $\bar{N}_k$ by summing with respect to l.

M-step. Obtain improved estimates of c_j and a_j by performing a probit analysis on the $\bar{r}_{jk}$ and $\bar{N}_k$, using X_k as the independent variable and weighting the corresponding term by $A(X_k)$. Estimates of c_j and a_j are the intercept and slope of the fitted probit regression line. (See Bock & Jones, 1980 reprint, p. 51.)

The EM cycles are continued until the estimates become stable to the required number of places. Convergence is only geometric, however, and becomes very slow as the solution point is approached. One method of speeding up the calculations is to employ an acceleration factor as proposed by Ramsay [1975] for solution of implicit equations. Another is to sort the response vectors into score groups and calculate likelihoods, $L_l(X_k)$, for successive patterns by changing factors only where the ones or zeroes differ between patterns. Because many item scores are the same in patterns within the same score group, considerable savings in computation result.

It may also be advantageous to stop the EM steps short of convergence and, if n is large, to approximate the matrix of second derivatives by sampling response patterns and applying (9). Provided the number of patterns sampled exceeds $2n$, this approximate information matrix will in general be positive-definite and may be inverted and used in one or two Newton-Raphson iterations to improve the nearly converged EM solution. If the information matrix is generated and inverted only once, the calculations should be within the capacity of a large computer. As an added benefit, the elements of the inverse information matrix will supply large-sample variances and covariances of the MML estimators. Thus, standard errors of the estimated item parameters, which are not provided by the EM solution, would be available.

An Alternative Derivation of the EM Procedure

The steps of the EM algorithm in this application can also be derived as an extension of the missing information principle used by Dempster, Laird, and Rubin [1977] to obtain maximum likelihood estimates when the probability model belongs to the exponential family. In the present context, the ability variable θ carries the missing information which, if known, would permit the item parameters to be estimated by conventional probit analysis as in (10) and (11). If the model belonged to the exponential family, a sufficient statistic would exist for θ, and, according to the missing information principle, expected values of this statistic given the observed data would be substituted in the probit or maximization step of the EM algorithm.

In the present case, no simple sufficient statistics exist for θ, so we take a direct likelihood approach and replace each individual observation θ_i by its conditional expectation, given the observed $\mathbf{x}_i$. From Bayes' theorem, the conditional distribution of θ given

$\mathbf{x} = \mathbf{x}_i$ is

$$g(\theta \mid x_i) = \frac{P(\mathbf{x} = \mathbf{x}_i \mid \theta) g(\theta)}{P(\mathbf{x} = \mathbf{x}_i)},$$

and therefore the conditional expectation of θ given $\mathbf{x} = \mathbf{x}_i$ is, using (4) and (5),

$$E(\theta \mid \mathbf{x}_i) = \frac{\int_{-\infty}^{\infty} \theta g(\theta) \prod_j^n [\Phi_j(\theta)]^{x_{ij}} [1 - \Phi_j(\theta)]^{1 - x_{ij}} \, d\theta}{\int_{-\infty}^{\infty} g(\theta) \prod_j^n [\Phi_j(\theta)]^{x_{ij}} [1 - \Phi_j(\theta)]^{1 - x_{ij}} \, d\theta}. \tag{13}$$

Approximating the integrals by q-point sums indexed by k as before, and recoding the i-th subject to the l-th score pattern, we have, from (8),

$$E(\theta \mid \mathbf{x}_i) \cong \frac{\sum_k^q X_k L_l(X_k) A(X_k)}{\tilde{P}_l}, \tag{14}$$

a simple weighted mean of the X_k. There are s distinct score patterns, and hence s values of $E(\theta \mid \mathbf{x}_i)$. The number of responses (the sample size for the probit model) at $\mathbf{x}_i$ is r_l, and the number correct on the j-th item is $\mathbf{x}_{lj} r_l$. Thus, a probit model may be fitted to s points, using $E(\theta \mid \mathbf{x}_i)$ as the (expected) ability variable, with $\mathbf{x}_{lj} r_l$ as the number of items correct out of r_l at this ability value. This appears to involve s ability values, but since $E(\theta \mid \mathbf{x}_i)$ is a weighted sum of only q terms in X_k, the fitting of the probit model to $E(\theta \mid \mathbf{x}_i)$ involves only the X_k taken as the (q) ability values. The number correct on the j-th item at this ability value is then

$$\bar{r}_{jk} = \frac{\sum_l^s r_l \mathbf{x}_{lj} L_l(X_k) A(X_k)}{\tilde{P}_l},$$

and the corresponding sample size is

$$\bar{N}_k = \frac{\sum_l^s r_l L_l(X_k) A(X_k)}{\tilde{P}_l},$$

as given previously.

It should be clear that a comparable computational procedure may be used if the logit, rather than the probit, model is assumed for the item response curve. The item by item logit analyses depend upon the θ_i, so the E-step again involves the replacement of θ_i by its conditional expectation.

We are indebted to an anonymous reviewer for pointing out that this extension of the EM principle is not the same as the general EM algorithm of Dempster, Laird, and Rubin [1977] (which, due to an error in their proof, is not rigorously established). For models outside the exponential family, they take the expected value of the log likelihood of the missing data given the observed data and the provisional values of the parameters. We, on the other hand, substitute the expected values of the missing data (given the observed data and the provisional values of the parameters) into the log likelihood. As we have seen, our treatment satisfies the marginal likelihood equations and, like theirs, reduces to the special EM algorithm when the model is within the exponential family and the log likelihood is linear in the sufficient statistics.

As is well known, the only item response function within the exponential family is the

one-parameter logistic model (Rasch). For that model, the marginal likelihood equation for the difficulty parameter γ_j is

$$\sum_{k}^{q} \bar{r}_{jk} A(X_k) - \sum_{k}^{q} \bar{N}_{jk} \Psi_j(X_k) A(X_k) = 0,$$

where $\Psi_j = 1/(1 + e^{-(\gamma_j - X_k)})$. Because the likelihood for the one parameter model depends only on the number-right score and not the response pattern [see Thissen, Note 2], this result is in the form of Dempster, Laird and Rubin's equation (2.13)—the so-called "striking representation" of likelihood equations for the exponential family as the difference of conditional and unconditional expectations of sufficient statistics.

Empirical Characterization of the Ability Distribution

The quadrature weight, $A(X_k)$, in (12) is approximately the normalized (i.e., $\sum_k^q A(X_k) = 1$) probability density of a $N(0, 1)$ random variable at the point X_k. (It approaches this quantity exactly as q becomes large). In Gaussian quadrature formulas, the points are chosen to best approximate the integral of certain classes of functions as a sum of a specified number of terms [see Stroud & Sechrest, 1966]. If some other prior distribution of ability is assumed, other points may be chosen and a normalized density point k substituted for $A(X_k)$ in (12). If a rectangular prior is assumed, for example, q points may be set at equal intervals over an appropriate range and the quadrature weights set equal to $1/q$.

Since probit analysis is essentially iterated straight-line least squares regression analysis, and since the quadrature weights serve only to weight the square and cross-products accumulated over the q points, we would expect the estimated item parameters to be little affected by the choice of prior distribution. This is borne out by the numerical examples below, in which the parameter estimates obtained with rectangular and empirical priors are all but indistinguishable from those obtained with the normal prior.

These results suggest that the marginal maximum likelihood approach can be freed of assumptions about the population of subjects by roughly estimating the distribution in the sample for some provisional estimates of the item parameters. Then the densities of this "empirical" prior at selected points can be used as weights in the final parameter estimation.

In principle, it is possible to estimate parameters of the prior distribution from the sample by maximum likelihood. Andersen and Madsen [1977] and Sanathanan and Blumenthal [1978] have successfully estimated the mean and variance of an assumed normal prior by this method. Because the likelihood is so insensitive to the shape of the prior, however, this approach could not be depended upon to estimate accurately the finer features of the ability distribution (e.g., coefficients of skewness and kurtosis) in practical size samples. We will therefore employ the simple, but effective, expedient of characterizing the continuous distribution of ability (which we assume to have finite mean and variance) by means of a discrete distribution on a finite number of equally spaced points (i.e., a histogram). As the estimate of the density of the ability distribution at the point X_k, we use the posterior density given the data. Again expressing the definite integral as a quadrature, we have

$$g(X_k) \cong \frac{\sum_{l}^{s} r_l L_l(X_k) A(X_k)}{\sum_{h}^{q} \sum_{l}^{s} r_l L_l(X_h) A(X_h)}. \tag{15}$$

The values from (15) can then be used in place of $A(X_k)$ in (12) to obtain improved estimates of the item parameters. This process could be repeated in the same manner that

Sanathanan and Blumenthal [1978] estimate the mean and variance of a normal prior, but the improvement in estimates of the item parameters would be negligible. This process in fact gives the nonparametric maximum likelihood estimator of the prior distribution discussed by Laird [1978].

We prefer to assume a normal prior in (15) as the starting point for the empirical prior on the grounds that it represents maximal uncertainty in a distribution with finite mean and variance [Rao, 1973, p. 162]. If a flat prior is assumed, the discrete posterior distribution may be excessively heavy in the tails, especially if a large proportion of all zero or all one response vectors occurs in the data.

Test of Fit

The likelihood ratio chi-square statistic for testing the assumed model against a general multinomial alternative is

$$G^2 = 2\left(\sum_l^s r_l \log_e \frac{r_l}{N\tilde{P}_l}\right) \tag{16}$$

on $s - 2n$ degrees of freedom (i.e., neglecting patterns with $r_l = 0$) for $2n < s < 2^n$ or $s - 2n - 1$ for $s = 2^n$. If 2^n is large relative to N, the frequency table will be sparse and (16) will tend to be unstable numerically because $N\tilde{P}_l$ will frequently be less than 5. In such cases, the frequencies of patterns with small expectations should be pooled until all expected frequencies equal or exceed 5. In principle, it does not matter which patterns are pooled, but for purposes of displaying the observed and expected frequencies, it is convenient to pool neighboring frequencies when the patterns are ordered according to an estimate of ability corresponding to each (see Section 6). Pooling should be held to a minimum so that the number of frequencies, say s_0, after pooling is greater than $2n$. The L.R. chi-square on $s_0 - 2n - 1$ degrees of freedom after pooling then provides a conservative test of fit since the likelihood has not actually been maximized in the pooled data.

Example 1. LSAT Sections 5 and 6

The EM procedures described in the second and third sections (above) were applied to the data for the Law School Aptitude Test (LSAT) presented in Bock and Lieberman [1970]. When the normal prior is assumed, the estimated item thresholds and slopes agree to about half an order of magnitude in the third decimal place with those obtained by Bock and Lieberman, even though fewer quadrature points were used in the present analysis. As shown in Table 1, and with the possible exception of the 2-point prior for Section 7, the EM solutions with $q = 10$ and $q = 2$ give essentially the same values as the Bock and Lieberman solution. The likelihood ratio chi-squares for goodness-of-fit of these solutions are also nearly identical, again with the exception of the 2-point solution for Section 7.

To illustrate the insensitivity of MML estimation to the form of the prior distribution, we also show in Table 1 the parameter estimates when the rectangular and the empirical prior distribution of ability are assumed. Because the location and scale of the estimates depend upon the arbitrary mean and variance of the prior, we have made all of the estimates comparable by applying the restrictions

$$\prod_j^n a_j = 1,$$

and

$$\sum_i^n b_j = 0.$$

TABLE 1

ITEMS FROM LSAT SECTIONS 6 AND 7: RESTRICTED ESTIMATES OF THRESHOLDS AND SLOPES ASSUMING VARIOUS DISTRIBUTIONS OF ABILITY (WITH TESTS OF FIT)

	Item	Bock-Lieberman (1970)	10-point normal	2-point normal	10-point rectangular	10-point empirical
Section 6						
	1	-.6785	-.6787	-.6998	-.6701	-.6819
	2	.3159	.3161	.3155	.3898	.3170
Threshold	3	.7863	.7878	.7702	.7754	.7871
	4	.0920	.0923	.0851	.0814	.0931
	5	-.5159	-.5174	-.4710	-.4964	-.5153
	1	.9798	.9788	.9660	.9890	.9768
	2	1.0160	1.0149	1.0367	1.0198	1.0170
Slope	3	1.2593	1.2652	1.1974	1.2228	1.2606
	4	.9482	.9476	.9510	.9440	.9490
	5	.8413	.8397	.8769	.8589	.8414
Fit (G^2) df = 21		21.28	21.29	23.70	22.40	21.28
Section 7						
	1	-.3097	-.3086	-.2821	-.2966	-.3107
	2	.3841	.3836	.3737	.3825	.3859
Threshold	3	.2017	.1998	.1630	.1862	.2050
	4	.4487	.4480	.4020	.4267	.4510
	5	-.7248	-.7229	-.6566	-.6989	-.7312
	1	.9585	.9606	.9979	.9747	.9563
	2	1.1084	1.1086	1.1298	1.1220	1.1101
Slope	3	1.6877	1.6797	1.6590	1.7001	1.6911
	4	.7922	.7927	.7182	.7481	.7954
	5	.7040	.7053	.7445	.7190	.7004
Fit (G^2) df = 21		31.59	31.67	42.25	34.83	31.51

The discrete distribution estimated for LSAT 6 and 7 by (15) are compared with the weights for a normal prior in Table 2. As would be expected, the ability distributions for this sample of subjects based on items from Section 6 and Section 7 are nearly identical. Relative to the normal, the empirical distributions are skewed somewhat toward higher abilities.

TABLE 2

NORMAL AND EMPIRICAL DISCRETE DISTRIBUTIONS OF ABILITY

Point	1	2	3	4	5	6	7	8	9	10
Deviate	-4.86	-3.58	-2.48	-1.47	-.48	.48	1.47	2.48	3.58	4.86
Normal	.431 x10^{-5}	.758 x10^{-3}	.191 x10^{-1}	.136	.345	.345	.136	.191 x10^{-1}	.258 x10^{-3}	.431 x10^{-5}
Empirical										
Section 6	.264 x10^{-6}	.944 x10^{-4}	.470 x10^{-2}	.690 x10^{-1}	.270	.411	.215	.357 x10^{-1}	.153 x10^{-2}	.892 x10^{-5}
Section 7	.410 x10^{-6}	.800 x10^{-4}	.245 x10^{-2}	.324 x10^{-1}	.221	.450	.252	.411 x10^{-1}	.172 x10^{-2}	.995 x10^{-5}

The parameter values and fits for solutions based on these nonnormal priors are so similar to the normal prior solutions that there is little to choose between them. Nevertheless, we would prefer to avoid arbitrary assumptions by employing the empirical prior in those psychometric applications where the calibration sample is selected arbitrarily.

Note that the two-point solution is almost identical to fitting a two-latent class model. The results suggest that in some cases (e.g., LSAT 6) the latent class and latent trait model cannot be distinguised, but in other cases (e.g., LSAT 7) they can. The fit of the two-latent class model to the LSAT data is examined by Aitkin in the discussion of Bartholomew [1980].

Estimating Ability

Assuming the item parameters to be known from a previous calibration, the ability of a subject with pattern $\mathbf{x}_i = [x_{ij}]$, $j = 1, 2, \ldots, n$, may be estimated by one of the following methods.

i) Maximum likelihood estimator, $\hat{\theta}_i$:

Solve

$$\sum_{i}^{n} a_j B_{ij} = 0 \tag{17}$$

by Newton-Raphson with second derivative

$$\sum_j^n a_j^2 \{B_{ij} C_{ij} - W_j(\theta)\}_{\theta=\hat{\theta}_i} \tag{18}$$

starting from

$$\hat{\theta}_i^{(0)} = \Phi^{-1}\left(\frac{\sum_j^n r_{ij}}{n}\right),$$

where

$$B_{ij} = \frac{[x_{ij} - \Phi_j(\theta)]\phi_j(\theta)}{\Phi_j(\theta)[1 - \Phi_j(\theta)]},$$

$$C_{ij} = -z_j(\theta) - \phi_j(\theta)\left\{\frac{1}{\Phi_j(\theta)} - \frac{1}{1 - \Phi_j(\theta)}\right\}$$

and

$$W_j(\theta) = \frac{\phi_j^2(\theta)}{\Phi_j(\theta)[1 - \Phi_j(\theta)]}.$$

The variance of this estimator is, for large n,

$$V(\hat{\theta}_i) = \left[\sum_j^n a_j^2 W_j(\theta)\right]^{-1}_{\theta=\hat{\theta}_i} \tag{19}$$

ii) Bayes modal (or MAP, "maximum *a posteriori*") estimator, $\dot{\theta}_i$:

Solve

$$\left[\sum_j^n a_j B_{ij} + \frac{\partial g(\theta)}{\partial \theta}\right]_{\theta=\dot{\theta}_i} = 0, \tag{20}$$

for any $g(\theta)$ with finite mean and variance by Newton-Raphson with second derivative

$$\left[\sum_j^n a_j^2[B_{ij} C_{ij} - W_j(\theta)] + \frac{\partial^2 g(\theta)}{\partial \theta^2}\right]_{\theta=\dot{\theta}_i} \tag{21}$$

beginning from $\theta_i^{(0)}$ as above.

The variance of this estimator for large n is given by

$$V(\dot{\theta}_j) = \left[\sum_j^n a_j^2 W_j(\theta) + \frac{\partial^2 g(\theta)}{\partial \theta^2}\right]^{-1}_{\theta=\dot{\theta}_i} \tag{22}$$

iii) Bayes (or EAP, "expected *a posteriori*") estimator, $\bar{\theta}_i$:

Evaluate (13) by the quadrature (14) with $\mathbf{x}_l = \mathbf{x}_i$. The variance of this estimator may be obtained by the quadrature,

$$V(\bar{\theta}_i) = \sum_k^q \frac{(X_k - \bar{\theta}_i)^2 L_i(X_k) A(X_k)}{\tilde{P}_i}. \tag{23}$$

The maximum likelihood estimator has the disadvantage of not giving finite values for patterns with all correct or all incorrect. Plausible values can be assigned to these patterns, of course, but there is a risk of biasing sample statistics based on such scores [see Goldstein,

1980]. Certain "unlikely" patterns can also cause the Newton-Raphson iterations to fail due to what Andersen [1980] calls the "spider web" effect. In these cases, an interval-bisection method must be substituted for Newton-Raphson.

The two Bayes estimators are free of these problems, but the MAP estimator has the disadvantage of requiring a proper continuous density function for the prior distribution. The EAP estimator implemented by quadrature, on the other hand, can use a discrete prior such as the empirical prior proposed above. All of these estimates require computation of a latent ability estimate for each pattern of item scores. Again great economies of calculation can result if patterns are sorted into score groups before estimation.

We note in passing that objections to Bayes estimation of individual differences have sometimes been raised [Bock, 1972a]. These objections are without force, however, if the scores of two or more subjects are compared for purposes of selection only when their abilities are estimated with respect to the same prior. Local or conditional priors should be used only when for some inferential purpose it is desired to compare subjects to others like themselves in the sense of having the same demographic background, or whatever.

The One-Parameter Logistic Model (Rasch)

MML for the one-parameter logistic model is an alternative to conditional maximum likelihood (CML) estimation, and it is much easier to compute. In Table 3, we compare the MML Rasch difficulty parameter estimates for Sections 6 and 7 of the LSAT with the CML estimates reported by Andersen [1980, pp. 253–254]. In both cases, the estimates are constrained to sum to zero, but their scale is free and is determined by the logit function. Likelihood ratio chi-square for goodness-of-fit is shown for both estimators. The marginal estimates assume a normal prior.

The marginal and conditional estimates are almost identical for Section 6, where the one-parameter model fits the data well. This is to be expected since both estimators are consistent, and the sample size is large. Neither of these estimators is subject to the $n/(n-1)$ bias that is encountered in fixed-effect estimation for this model [Andersen, 1980].

In Section 7 of the LSAT, where the one-parameter model fits poorly, possibly because item 3 is somewhat more discriminating than the other items, the agreement of the conditional and marginal solutions is not as good. With such poor fit, neither the MML nor the CML solution can be relied upon. Threshold estimates based on the one-parameter model can be substantially in error when the assumption of homogeneous slopes is violated. Thissen (1981) has pointed out that, for the one parameter logistic model, the likelihoods, $L(X_k)$ are the same for all patterns in the same score group. Thus, at most $n+1$ distinct likelihoods need to be evaluated for the purposes of MML estimation with this model. The EM implementation of MML is therefore the most economical method of obtaining Rasch difficulty parameters; it is faster than the fixed-effect maximum likelihood solution and faster by far than the conditional solution.

Structural Item Parameters

The linear logistic one-parameter models discussed by Micko [1969], Scheiblechner [1972], and Fischer [1973, 1980] for the study of item-facet structures can be extended to one or more parameters of other response models. The proposed EM procedure can be used to estimate the structural parameters of such models, or indeed of any many-one function of the item parameters,

$$u_j = f_j(v_1, v_2, \ldots, v_m); \qquad j = 1, 2, \ldots, m; \qquad m < n.$$

In this case the derivatives, $\partial u_j/\partial v_h$, appear in (7), which then contains a sum over j.

TABLE 3

COMPARISON OF CONDITIONAL AND MARGINAL MAXIMUM LIKELIHOOD ESTIMATES FOR THE ONE-PARAMETER LOGISTIC MODEL

	Item	CML (Andersen, 1980)	MML 10-pt. normal prior
Section 6			
	1	-1.256	-1.2552
	2	0.475	0.4763
	3	1.236	1.2350
	4	0.168	0.1684
	5	-0.623	-0.6245
		$G^2 = 3.1$	$G^2 = 21.80$
		df = 12	df = 25
Section 7			
	1	-0.641	-0.5413
	2	0.583	0.5359
	3	-0.134	-0.1340
	4	0.758	0.8054
	5	-0.566	-0.6660
		$G^2 = 35.9$	$G^2 = 43.90$
		df = 12	df = 25

Only m likelihood equations in m unknowns then need to be solved, but they are not in general separable in the manner of (12). Thus, one m-dimensional Newton-Raphson solution will be required in each M-step of the algorithm. In typical applications, m is much smaller than n, however, and EM estimation of parameters of the facet structure should be feasible.

More Than One Latent Dimension

Nothing in the proposed EM procedure limits it to one dimension of ability. Consider, for example, a two-factor model. Employing Thurstone's [1947] multiple factor model, we assume for item j the latent variable

$$y_j = \alpha_{j1}\theta_1 + \alpha_{j2}\theta_2 + \varepsilon_j, \tag{24}$$

with

$$y_j \sim N(0, 1) \quad \text{and} \quad \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right),$$

which implies

$$\varepsilon_j \sim N(0, 1 - \alpha_{j1}^2 - \alpha_{j2}^2).$$

Subject i is assumed to score $x_{ij} = 1$ on item j if $y_{ji} = \alpha_{ji}\theta_{i1} + \alpha_{j2}\theta_{i2} + \varepsilon_{ji}$ is greater than the threshold γ_j, and $x_{ij} = 0$ otherwise.

Then

$$P(x_{ij} = 1 \mid \theta_{i1}, \theta_{i2}) = \frac{1}{(2\pi)^{1/2}} \int_{-z_j(\theta_i)}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt, \tag{25}$$

where

$$-z_j(\theta_i) = \frac{\gamma_j - \alpha_{j1}\theta_{1i} - \alpha_{j2}\theta_{2i}}{\sigma_j},$$

and

$$\sigma_j = (1 - \alpha_{j1}^2 - \alpha_{j2}^2)^{1/2}$$

In terms of the computationally more convenient intercept and slope parameters,

$$z_j(\theta_i) = c_j + a_{j1}\theta_{i1} + a_{j2}\theta_{i2}, \tag{26}$$

where

$$c_j = -\frac{\gamma_j}{\sigma_j}, \qquad a_{j1} = \frac{\alpha_{j1}}{\sigma_j}, \quad \text{and} \quad a_{j2} = \frac{\alpha_{j2}}{\sigma_j}.$$

The unconditional probability of score pattern $\mathbf{x}_l$ can be computed by two-dimensional Gauss-Hermite quadrature:

$$\tilde{P}_l = \sum_{k_2}^{q} \sum_{k_1}^{q} L_l(X_{k_1}, X_{k_2}) A(X_{k_1}) A(X_{k_2}). \tag{27}$$

The only changes required to extend the EM procedure to two factors are (i) calculating the conditional probabilities L_l at all pairs of quadrature points corresponding to $k_1 = 1, 2, \ldots, q$ and $k_2 = 1, 2, \ldots, q$; (ii) accumulating for each item the $q \times q$ table of provisional expected frequencies of correct response and the expected sample sizes [cf. (12)]; (iii) performing a two-variable "multiple" probit analysis on the expected frequencies and expected sample sizes to obtain provisional maximum likelihood estimates of c_j, a_{j1}, and a_{j2}; (iv) performing steps 1, 2, and 3 to complete one EM cycle; (v) repeating EM cycles until the values of the estimates become stable. As in the one-dimensional case, it may be advantageous to complete the process with one Newton-Raphson step after a limited number of EM cycles. The MML estimates of the factor loadings and threshold are then calculated by substituting estimates for parameters in $\alpha_{j1} = a_{j1}/d_j$, $\alpha_{j2} = a_{j2}/d_j$, and $\gamma_j = -c_j/d_j$, where $d_j = (1 + a_{j1}^2 + a_{j2}^2)^{1/2}$. Notice that boundary solutions in which one or more slopes are infinite correspond to Heywood cases in estimating factor loadings.

As in the unidimensional case, the solution can also be freed of arbitrary assumptions about the prior distribution. For example, an empirical bivariate distribution on q^2 points could be substituted for the bivariate normal. In general, the prior could be correlated so that the weights are $A(X_{k_1}, X_{k_2})$, rather than independent with weights equal to the product of the marginal weights.

Because the multiple factor model is unique only up to a rotation of the factor space, an unrestricted solution by this method is nonunique and depends upon the starting values. The goodness-of-fit statistic is unique, however, and any particular solution can be carried into any other [e.g., the varimax solution (Kaiser, 1958)] by a rotation.

EAP estimates of a subject's ability on each dimension can be readily computed from his item score pattern $\mathbf{x}_i$. The estimators are the marginal means of the bivariate *a posteriori* distribution of θ_1 and θ_2, given $\mathbf{x}_i$. They can be evaluated by quadrature. For example,

$$\bar{\theta}_1 = \sum_{k_1}^{q} X_{k_1} \left[\sum_{k_2}^{q} L_i(X_{k_1}, X_{k_2}) A(X_{k_2}) \right] \frac{A(X_{k_1})}{\tilde{P}_i}$$

The variances of the estimators can be evaluated in similar marginal quadratures of $(X_{k_1} - \bar{\theta}_1)^2$ and $(X_{k_2} - \bar{\theta}_2)^2$, and their covariance by two-dimensional quadrature of $(X_{k_1} - \bar{\theta}_1)(X_{k_2} - \bar{\theta}_2)$.

TABLE 4

GLS (VARIMAX ROTATED) AND MML TWO-FACTOR SOLUTIONS
FOR LSAT SECTION 7
(GLS SOLUTION USED AS STARTING VALUES FOR MML)

	Item	GLS Christoffersson (1975)	MML 5-pt. normal	MML 3-pt. normal
Intercept				
	1	1.6172	1.6177	1.4936
	2	.4734	.4722	.4719
	3	1.4303	1.3977	1.2492
	4	.2968	.2938	.2933
	5	1.0902	1.0896	1.0847
Slope 1				
	1	1.3551	1.3539	1.2498
	2	.2297	.2312	.2368
	3	.4059	.3884	.3966
	4	.3572	.3448	.3421
	5	.3356	.3300	.3252
Slope 2				
	1	.2485	.2646	.2171
	2	.5538	.5384	.5386
	3	1.5815	1.5505	1.4397
	4	.2836	.2835	.2820
	5	.2436	.2498	.2344
		$G^2 = 0.63$	$G^2 = 21.23$	$G^2 = 21.63$
		df = 1	df = 17	df = 17

These ability estimates are, of course, nonunique and depend upon the particular choice of rotation.

All of these results extend straightforwardly to more than two factors. If, as Example 2 suggests, three quadrature points ($q = 3$) are sufficient in these calculations, extensions to at least 5 factors should be possible on fast computers. MML estimation would then provide a full-information alternative to the procedures for factor analysis of dichotomous data proposed by Christoffersson [1975], Muthén [1978], and Bartholomew [1980].

Example 2. LSAT Section 7

Seeing that the items of LSAT Section 7 did not fit the one-dimensional normal ogive model very well, we examined the fit of the two-factor model. We could have obtained suitable starting values for the slopes form an orthogonal unweighted least squares factor analysis of the matrix of inter-item tetrachoric correlations and, for the intercept, from the inverse normal transform of the item percent correct. But for the present data, better values exist in Christofferson's [1975] generalized least squares (GLS) two-factor estimates. Converting Christofferson's threshold estimates and varimax rotated factor loadings to intercept and slope parameters defined in (26), we obtained the values shown in Table 4. Both three- and five-point Gauss-Hermite quadrature show that an MML solution exists very near the GLS solution. Results of the two-point quadrature were less satisfactory and are not shown.

That the goodness-of-fit chi-square for the MML solution is not significant ($p = .216$) supports a two-factor interpretation of the Section 7 LSAT items. (The number of degrees of freedom is the number of independent frequencies, 31, minus the number of intercept estimates, 5, minus the number of *independent* slope estimates, namely, $5 + 4 = 9$; thus, $df = 31 - 5 - 9 = 17$.) That the component chi-square, $31.67 - 21.13 = 10.54$, *is* significant on 4 degrees of freedom ($p = .032$) is further evidence that the second-factor loadings are not all zero. The varimax solution indicates that the first factor is best defined by item 1 and the second by item 3.

REFERENCE NOTES

1. Thissen, D. Personal communication, 1979.
2. Thissen, D. *Marginal maximum likelihood estimation for the one-parameter logistic model.* Manuscript submitted for publication, 1981.

REFERENCES

Andersen, E. B. Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 31–44.

Andersen, E. B. *Discrete statistical models with social science applications.* Amsterdam: North-Holland, 1980.

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

Bartholomew, D. J. Factor analysis for categorical data (with Discussion). *Journal of the Royal Statistical Society, Series B*, 1980, *42*, 293–321.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29–51.

Bock, R. D. Review of *The dependability of behavioral measurements* by Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *Science*, 1972a, *178*, 1275–1275A.

Bock, R. D., & Jones, L. V. *The measurement and prediction of judgment and choice.* Chicago: International Educational Services, 1980 (reprint).

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

Christoffersson, A. Factor analysis of dichotomized variables. *Psychometrika*, 1975, *40*, 5–32.

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

Fischer, G. Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 1973, *37*, 359–374.

Fischer, G. H. Individual testing on the basis of the dichotomous Rasch model. In L. J. T. Van der Kamp, W. F. Langerak, & D. M. N. de Gruijter (Eds). *Psychometrics for Educational Debates.* Chichester: Wiley, 171–188, 1980.

Goldstein, H. Dimensionality, bias, independence, and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, 1980, *33*, 234–246.

Haberman, S. J. Maximum likelihood estimates in exponential response models. *Annals of Statistics*, 1977, *5*, 815–841.

Kaiser, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, *23*, 187–200.

Kolakowski, D. & Bock, R. D. *NORMOG : Maximum likelihood item analysis and test scoring; normal ogive model.* Chicago: International Educational Services, 1973a.

Kolakowski, D. & Bock, R. D. *LOGOG: Maximum likelihood item analysis and test scoring; logistic model for multiple item responses.* Chicago: International Educational Services, 1973b.

Laird, N. M. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 1978, *73*, 805–811.

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117–138.

Lord, F. M. *Applications of item response theory to practical testing problems.* New York: Erlbaum Associates, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading (Mass.): Addison-Wesley, 1968.

Micko, H. C. A psychological scale for reaction time measurement. *Acta Psychologica*, 1969, *30*, 324–335.

Muthén, B. Contributions to factor analysis of dichotomized variables. *Psychometrika*, 1978, *43*, 551–560.

Ramsay, J. O. Solving implicit equations in psychometric data analysis. *Psychometrika*, 1975, *40*, 337–360.

Rao, C. R. *Linear statistical inference and its applications (2nd ed.).* New York: Wiley, 1973.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, No. 17*, 1969.

Sanathanan, L. & Blumenthal, S. The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, 1978, *73*, 794–799.

Scheiblechner, H. Das Lernen und Losen komplexer Denkaufgaben. *Zeitschrift fur Experimentelle und Angewandte Psychologie*, 1972, *19*, 476–506.

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

Thurstone, L. L. *Multiple factor analysis.* Chicago: University of Chicago Press, 1947.

Wood, R. L. Wingersky & Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters.* Princeton (N.J.): Educational Testing Service, 1976.

Wright, B. D. & Panchapakesan, N. A. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23–48.

Manuscript received 10/8/80
Final version received 8/3/81