PSYCHOMETRIKA—VOL. 50, NO. 3, 349–364.
SEPTEMBER 1985

BAYESIAN ESTIMATION IN THE TWO-PARAMETER LOGISTIC MODEL

HARIHARAN SWAMINATHAN

UNIVERSITY OF MASSACHUSETTS, AMHERST

JANICE A. GIFFORD

MOUNT HOLYOKE COLLEGE

A Bayesian procedure is developed for the estimation of parameters in the two-parameter logistic item response model. Joint modal estimates of the parameters are obtained and procedures for the specification of prior information are described. Through simulation studies it is shown that Bayesian estimates of the parameters are superior to maximum likelihood estimates in the sense that they are (a) more meaningful since they do not drift out of range, and (b) more accurate in that they result in smaller mean squared differences between estimates and true values.

Key words: Bayesian estimates, item response model, two-parameter logistic model, modal estimates, maximum likelihood estimates.

Introduction

An issue that often faces users of item-response models is the choice of an appropriate model for the analysis of item response data. While the one-parameter or the Rasch model is the most convenient and the easiest to work with, recent studies (Slinde & Linn, 1979; Loyd & Hoover, 1980) have shown that the Rasch model may not be completely adequate for the solution of certain measurement problems. The alternative, using the more elaborate, and possibly more appropriate, two- and three-parameter models, presents problems with respect to parameter estimation.

In item response models, the simultaneous estimation of parameters that characterize the items and the abilities of the respondents is often required. This is the well-known problem of estimating "structural" parameters in the presence of "incidental" parameters (Neyman & Scott, 1948; Andersen, 1972). In general, the incidental parameters have a correspondence with the observations that are increased to yield stable estimates of the structural parameters. It is meaningful, therefore, to consider the ability parameters the incidental parameters and the item parameters the structural parameters.

The basic problem, pointed out by Neyman and Scott (1948) is that the maximum likelihood estimators of the structural parameters are not consistent in the presence of incidental parameters. Andersen (1972) demonstrated this in the case of the Rasch model when he showed that with a fixed number of items, the small sample bias of the maximum likelihood estimators of the item parameters persists when the number of examinees approaches infinity. He further showed that consistent maximum likelihood estimators can be obtained by conditioning the likelihood function on the number correct score, the minimal sufficient statistic for the ability parameters in the Rasch model. Unfortunately, sufficient statistics for the ability parameters are not available in the two-parameter (or

The research reported here was performed pursuant to Grant No. N0014-79-C-0039 with the Office of Naval Research.

Requests for reprints should be sent to H. Swaminathan, School of Education, University of Massachusetts, Amherst, MA 01003.

0033-3123/85/0900-4044$00.75/0

the three-parameter) logistic model. Hence, conditional maximum likelihood estimators of the item parameters cannot be obtained for the two- and three-parameter models.

The consistency of the maximum likelihood estimators of the structural parameters in the presence of incidental parameters becomes an issue when the number of structural parameters is fixed. Haberman (1975) has shown that the bias in the maximum likelihood estimators of the structural parameters vanishes as the number of incidental and structural parameters approaches infinity. Although Haberman's result is valid only for the Rasch model, empirical evidence (Swaminathan & Gifford, 1983) shows that the result may be valid for the two- and three-parameter models (in fact, this empirical result is in agreement with that obtained by Kiefer and Wolfowitz, 1956, who showed that maximum likelihood estimators of structural parameters are consistent if the incidental parameters are independently and identically distributed).

Despite these encouraging results, joint maximum likelihood estimation of item and ability parameters in the two-parameter (and the three-parameter) models presents problems. Lord (1968), in estimating item parameters and ability parameters jointly, found it necessary to impose an upper bound on the range of values taken by the "discrimination" parameter, a_j, in the two-parameter logistic model. This limit on the range of values taken by a_j was necessary to prevent the estimates from drifting out of bounds. This has been criticized severely by Wright (1977) who argued that this was sufficient evidence that the discrimination parameter was inestimable. Although the outward drift of the discrimination parameter is less marked in large samples, this fact raises doubts regarding the appropriateness of joint maximum likelihood estimation of parameters in the two-parameter case.

Bock and Lieberman (1970) derived "marginal maximum likelihood" estimators of the item parameters in the normal ogive model by integrating out the ability parameter. This procedure requires numerical integration and the evaluation of the likelihood function over 2^n response patterns, where n is the number of items, a tedious procedure indeed where n is even moderately large. More recently, Bock and Aitkin (1981), employing the E-M algorithm, obtained marginal maximum likelihood estimates of the item parameters more efficiently. While marginal maximum likelihood estimators are superior to joint maximum likelihood estimators of item parameters, at least in small samples, they do not offer protection from Heywood type cases where inadmissable estimates of the discrimination parameters may be obtained.

When several parameters have to be estimated simultaneously, and when, as in item response models, structural as well as incidental parameters have to be estimated, a Bayesian approach may be appropriate (Zellner, 1971, pp. 114–112). This is particularly true when prior information or belief about the parameters is available, since the incorporation of such information will certainly increase the meaningfulness and the "accuracy" of the estimates. In one sense, the practice of imposing limits on the range of values taken by the discrimination parameter is tantamount to the specification of prior belief, albeit without a Bayesian justification.

Bayesian procedures for estimating parameters have been successfully applied in numerous situations. However, Bayesian methods have found only a limited application in the area of item response theory. Birnbaum (1969) and Owen (1975) have obtained Bayes estimates of ability parameters in item response models under the assumption that the item parameters are known. Swaminathan and Gifford (1982) have successfully applied a hierarchical Bayesian model suggested by Lindley (1971), Lindley and Smith (1972) and exemplified in the works of Novick, Lewis, and Jackson (1973) to the simultaneous estimation of parameters in the Rasch model. The hierarchical procedure advocated by Lindley (1971) provides a powerful method for the analysis of item response data and it is the purpose of this paper to demonstrate the applicability of this procedure for the estimation of parameters in the two-parameter logistic item response model. The Bayesian procedure described in this paper is appealing in that the nonconvergence of the estimates of the discrimination parameters can be effectively controlled by specifying an appropriate prior distribution. An added bonus is that the increased accuracy in the estimation of discrimination parameters results in increased accuracy in the estimates of the ability and difficulty parameters.

The Model

Let U_{ij} denote a random variable that represents the binary response of an examinee i ($i = 1, \ldots, N$) on item j ($j = 1, \ldots, n$). For a correct response $U_{ij} = 1$ while for an incorrect response, $U_{ij} = 0$. Further assume that the latent space is unidimensional and that the probability of correct response, $P[U_{ij} = 1]$ is given by

$$P_j(\theta_i) = P[U_{ij} = 1 \mid \theta_i, a_j, b_j] = [1 + \exp\{-a_j(\theta_i - b_j)\}]^{-1}, \tag{1}$$

while,

$$Q_j(\theta_i) = P[U_{ij} = 0 \mid \theta_i, a_j, b_j] = [1 + \exp\{a_j(\theta_i - b_j)\}]^{-1}. \tag{2}$$

From the assumption of local independence (Lord, 1980, p. 19) it follows that the joint probability of responses is

$$P[\mathbf{U} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}] = \prod_{i=1}^{N} \prod_{j=1}^{n} [P_j(\theta_i)]^{U_{ij}}[Q_j(\theta_i)]^{1-U_{ij}}, \tag{3}$$

where

$$\mathbf{U}' = [U_{11} \quad U_{12} \quad \cdots \quad U_{1n} \quad \cdots \quad U_{Nn}]$$

$$\boldsymbol{\theta}' = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_N]$$

$$\mathbf{a}' = [a_1 \quad a_2 \quad \cdots \quad a_n]$$

and

$$\mathbf{b}' = [b_1 \quad b_2 \quad \cdots \quad b_n].$$

The above statement is a statement of probability. However, once the observations are made, this becomes the likelihood function of the parameters $\boldsymbol{\theta}$, $\mathbf{a}$, $\mathbf{b}$, given by

$$L(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) = \prod_{i=1}^{N} \prod_{j=1}^{n} [P_j(\theta_i)]^{u_{ij}}[Q_j(\theta_i)]^{1-u_{ij}}, \tag{4}$$

where $\mathbf{u}' = [u_{11} \quad u_{12} \quad \cdots \quad u_{1n} \quad \cdots \quad u_{Nn}]$ is the vector of observed responses.

The relevant information concerning the parameters of interest, $\boldsymbol{\theta}$, $\mathbf{a}$, and $\mathbf{b}$ is contained in the posterior distribution of these parameters. It is well known that the joint posterior density of $\boldsymbol{\theta}$, $\mathbf{a}$, and $\mathbf{b}$ given the observation $\mathbf{u}$ can be expressed as

$$p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b} \mid \mathbf{u}) \propto L(\mathbf{u}, \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b})p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}). \tag{5}$$

The joint density $p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b})$ is the prior distribution of the vectors of parameters and is an expression of the prior belief or information the investigator has regarding these parameters.

Prior Distributions

In the first stage of the hierarchical model we assume that, a priori, the parameters $\boldsymbol{\theta}$, $\mathbf{a}$, and $\mathbf{b}$ are independent, i.e.,

$$p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) = p(\boldsymbol{\theta})p(\mathbf{a})p(\mathbf{b}). \tag{6}$$

We further assume that the information on the ability parameters θ_i is exchangeable (Novick, et al., 1973), and that a priori, the parameters θ_i are independently and identically normally distributed, i.e.,

$$\theta_i \mid \mu_\theta, \phi_\theta \sim N(\mu_\theta, \phi_\theta). \tag{7}$$

The notion of exchangeable prior information is reasonable since this merely asserts that prior belief about any θ_i is no different from that about any other. The assumption of normality also appears reasonable and has been made by numerous authors, e.g., Lord and Novick (1968).

The second stage of the Bayesian hierarchical model requires the specification of prior information on μ_θ and ϕ_θ. However, since the parameters are not identified, an identification condition may be imposed at this stage. For convenience μ_θ may be taken as zero and ϕ_θ as one to effect identification.

Specification of prior information about the difficulty parameters, b_j, is accomplished somewhat differently. As with θ_i, we assume that the prior information about b_j is exchangeable and that

$$b_j \mid \mu_b, \phi_b \sim N(\mu_b, \phi_b), \tag{8}$$

in the first stage. In the second stage, prior information about μ_b and ϕ_b needs to be specified. We shall assume that (i) μ_b and ϕ_b are independent, (ii) μ_b is uniform, and (iii) ϕ_b has the inverse chi-square distribution (Novick & Jackson, 1974, p. 109), i.e.,

$$p(\mu_b, \phi_b)\, d\mu_b\, d\phi_b \propto b^{-(\nu_b/2+1)} \exp\left(\frac{-\lambda_b}{2\phi_b}\right) d\mu_b\, d\phi_b. \tag{9}$$

The assumption of normality appears to be reasonable (Lord & Novick, 1968) especially if an item bank is available. Furthermore, as a result of the hierarchical Bayes model, departures from this assumption appear to have a negligible effect on the estimates of b_j.

Specification of prior distributions about the discrimination parameters presents a problem since there does not appear to be any precedent in this case. The parameter a_j is the slope of the item characteristic curve at the point of inflection, and hence, is almost certainly positive (in a typical test, items are chosen so that the "classical" discrimination parameter is positive) and bounded away from zero. Since only positive values of a_j are meaningful, this suggests a prior density that belongs the the Gamma family. In addition, it is well known that (Lord & Novick, 1968, p. 399),

$$|\Psi\{1.7a_j(\theta_i - b_j)\} - \Phi\{a_j(\theta_i - b_j)\}| < .01$$

where Ψ, and Φ are the logistic and the normal cumulative distribution functions respectively. This correspondence suggests that a_j has the form of the reciprocal of the standard deviation.

In Bayesian analysis, the prior distribution for the variance is usually taken as the inverse chi-square, described earlier. The reciprocal of the variance therefore has the chi-square distribution. It follows then that the reciprocal of the standard deviation has a chi distribution. The form of the chi distribution can be shown to be (Appendix A)

$$p(\chi \mid \nu, \omega)\, d\chi \propto \chi^{\nu-1} \exp\left(\frac{-\chi^2}{2\omega}\right) d\chi \tag{10}$$

where ν is the degrees of freedom and ω is a scale parameter. We therefore assume that, a priori,

$$p(a_1, a_2, \ldots, a_n) = p(a_1)p(a_2), \ldots, p(a_n),$$

and that

$$p(a_j)\, da_j \propto a_j^{\nu_j - 1} \exp\left(\frac{-a_j^2}{2\omega_j}\right) da_j. \tag{11}$$

As mentioned earlier, estimation of the discrimination parameter often poses problems. The specification of prior information for this parameter appears to solve such problems as nonconvergence. The difficulty and ability parameters do not pose problems in general and hence specifying prior information for these parameters may not be necessary. Therefore, vague or non-informative priors may be specified for these parameters. In this case, we take the prior distributions of these parameters as follows:

$$p(b_j)\, db_j \propto db_j, \qquad j = 1, \ldots, n \tag{12}$$

$$p(\theta_i)\, d\theta_i \propto d\theta_i, \qquad i = 1, \ldots, N. \tag{13}$$

This approach, that of specifying non-informative priors for b_j and θ_i but an "informative" prior for a_j is appealing for several reasons. The estimation procedure is clearly less cumbersome. In addition, this results in less regressed estimates of θ_i and b_j, a feature that may be more acceptable to non-Bayesians. However, it should be noted that in this case the advantage offered by informative priors, that of finite ability parameter estimates corresponding to perfect or zero scores, is lost.

The Posterior Distribution

As noted in (5), the joint posterior distribution of the item and ability parameters is given by

$$\begin{aligned} &p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mu_b, \phi_b \mid \mathbf{u}, \mu_\theta, \phi_\theta, \nu_b, \lambda_b, \nu_1, \nu_2, \ldots, \nu_n, \omega_1, \omega_2, \ldots, \omega_n) \\ &\qquad \propto L(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) \left\{ \prod_{i=1}^{N} p(\theta_i) \right\} \\ &\qquad \cdot \left\{ \prod_{j=1}^{n} p(b_j \mid \mu_b, \phi_b) \cdot p(\mu_b, \phi_b \mid \nu_b, \lambda_b) \right\} \left\{ \prod_{j=1}^{n} p(a_j \mid \nu_j, \omega_j) \right\}. \end{aligned} \tag{14}$$

The likelihood function $L(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b})$ is given by (4) while

$$\prod_{i=1}^{N} p(\theta_i) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \theta_i^2\right)$$

and

$$\left\{ \prod_{j=1}^{n} p(b_j \mid \mu_b, \phi_b) \right\} p(\mu_b, \phi_b \mid \nu_b, \lambda_b) \propto \phi_b^{-(n + \nu_b + 2)/2} \exp - \left\{ \frac{\lambda_b + \sum_{j=1}^{n} (b_j - \mu_b)^2}{2\phi_b} \right\}.$$

Integrating out the nuisance parameters μ_b, ϕ_b, we obtain the following joint posterior distribution for the item and ability parameters:

$$\begin{aligned} &p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b} \mid \mathbf{u}, \nu_b, \lambda_b, \nu_1, \ldots, \nu_n, \omega_1, \ldots, \omega_n) \propto L(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \theta_i^2\right) \\ &\qquad \cdot \left\{ \lambda_b + \sum_{j=1}^{n} (b_j - b\cdot)^2 \right\}^{-(n + \nu_b - 1)/2} \left\{ \prod_{j=1}^{n} a_j^{\nu_j - 1} \exp\left(\frac{-a_j^2}{2\omega_j}\right) \right\}. \end{aligned} \tag{15}$$

Here $b\cdot$ denotes the mean of $\mathbf{b}$, and $L(\cdot)$ is the likelihood function. If non-informative priors are chosen for θ_i and b_j, the posterior distribution reduces to

$$p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b} \mid \mathbf{u}, \nu_1, \ldots, \nu_n, \omega_1, \ldots, \omega_n) \propto L(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b})\left\{\prod_{j=1}^{n} a_j^{\nu_j - 1} \exp\left(\frac{-a_j^2}{2\omega_j}\right)\right\}. \tag{16}$$

The posterior distribution given by (15) or (16) contains all the information required for making probability statements about the item and ability parameters. However, given its complexity, it is not in a readily usable form. Hence, it is necessary to obtain point estimates of the item and ability parameters.

Estimation of Parameters

Following Lindley and Smith (1972), and Novick, Jackson and Lewis (1973), we shall take the joint modal estimates of the parameters as the Bayesian estimates. The logarithm of the joint posterior distribution (15) can be expressed as

$$\begin{aligned} \log p(\boldsymbol{\theta}, \mathbf{a}, \mathbf{b} \mid \mathbf{u}) = \log L(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}) \\ + \log f_1(\boldsymbol{\theta}) + \log f_2(\mathbf{b}) \\ + \log f_3(\mathbf{a}) + \text{constant}, \end{aligned} \tag{17}$$

where $f_1(\boldsymbol{\theta})$, $f_2(\mathbf{b})$, and $f_3(\mathbf{a})$ are the contributions of the priors on $\boldsymbol{\theta}$, $\mathbf{b}$, and $\mathbf{a}$ to the posterior distribution. The "modal" equations are then given by

$$\frac{\partial\{\log L + \log f_1(\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{18}$$

$$\frac{\partial\{\log L + \log f_2(\mathbf{b})\}}{\partial \mathbf{b}} = \mathbf{0}, \tag{19}$$

$$\frac{\partial\{\log \mathrm{L} + \log f_3(\mathbf{a})\}}{\partial \mathbf{a}} = \mathbf{0}. \tag{20}$$

Solutions to these modal equations, or alternatively, the values of the parameters that maximize the joint posterior distribution, are not available in closed forms and hence numerical procedures have to be employed. The Newton-Raphson procedure is applicable in this situation and easily implemented.

If $\mathbf{x}^{(k)}$ is the kth approximation of the vector of values that maximizes the joint posterior distribution, then a better approximation, $\mathbf{x}^{(k+1)}$, is given by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [H^{-1}(\mathbf{x}^k)]g(\mathbf{x}^k) \tag{21}$$

where $H(\mathbf{x}^k)$ is the matrix of second derivatives and $g(\mathbf{x}^k)$ is the vector of first derivatives of the logarithm of the posterior distribution evaluated at $\mathbf{x}^k$. Since $H(\mathbf{x}^k)$ is a square matrix of order $2n + N$, for ease of implementation, the off-diagonal elements of $H(\mathbf{x}^k)$ can be ignored, in which case,

$$x_m^{(k+1)} = x_m^{(k)} - \frac{g[x_m^{(k)}]}{h[x_m^{(k)}]} \tag{22}$$

where x_m is the mth parameter and g and h are the first and second derivatives of the logarithm of the posterior distribution. It is easily seen that $g(\ )$ and $h(\ )$ have two additive components: one that results from the contribution of the likelihood function while the other from the contribution of the prior distribution. These components are conveniently summarized in Table 1.

Table 1
First and second derivatives of the logarithm of the joint posterior distribution

Parameter	Contribution	First derivative*	Second derivative
θ_i	Likelihood	$\sum_{j=1}^{n} a_j(u_{ij}-P_{ij})$	$-\sum_{j=1}^{n} a_j P_{ij}(1-P_{ij})$
	Prior	$-\theta_i$	-1
a_j	Likelihood	$\sum_{i=1}^{N} (\theta_i-b_j)(u_{ij}-P_{ij})$	$-\sum_{i=1}^{N} (\theta_i-b_j)^2 P_{ij}(1-P_{ij})$
	Prior	$+(\nu_j-1)/a_j - a_j/\omega_j$	$-(\nu_j-1)/a_j^2 - 1/\omega_j$
b_j	Likelihood	$-\sum_{i=1}^{N} a_j(u_{ij}-P_{ij})$	$-\sum_{i=1}^{N} a_j^2 P_{ij}(1-P_{ij})$
	Prior	$-(b_j-b.)/\sigma_b^2$	$-[\sigma_b^2(1-\frac{1}{n})-\{2(b_j-b.)^2/(\nu_b+n-1)\}]/(\sigma_b^2)^2$

$* \; P_{ij} = [1+\exp\{-a_j(\theta_i-b_j)\}]^{-1}$ $\qquad \sigma_b^2 = [\lambda_b + \sum_{j=1}^{n} (b_j-b.)^2/(\nu_b+n-1)]$

$b. = \sum b_j/n$

The first and second derivatives given in this table can easily be adjusted to reflect the different prior distributions that may be assumed. If, for example, non-informative priors are assumed for θ_i and b_j, the first and second derivatives will only have contributions from the likelihood.

The iterative scheme given in (22) can be employed to first estimate the ability parameters holding the item parameters fixed. When convergence is obtained, the process is repeated for each set of item parameters. This can be called a stage. The stages are then repeated until the required degree of accuracy is attained. Although the convergence of this iterative process cannot be proved rigorously, for large number of items and examinees, the process will converge within any stage of the iterative scheme.

This can be seen if we note that asymptotically the likelihood dominates the prior and for a given ability $\boldsymbol{\theta}$, the matrix of second derivatives of each pair of discrimination and difficulty parameters is negative definite. This follows since

$$\frac{\partial^2 \log L}{\partial a_j \partial b_j} = a_j(\theta_i - b_j)P_{ij}(1 - P_{ij}),$$

and the determinant of the negative of the matrix of second derivatives (using the expressions given in Lord, 1980, p. 191) given by

$$\left[\sum_{i=1}^{N}(\theta_i - b_j)^2 P_{ij}(1 - P_{ij})\right]\left[\sum_{i=1}^{N} a_j^2 P_{ij}(1 - P_{ij})\right] - \left[\sum_{i=1}^{N}(\theta_i - b_j)a_j P_{ij}(1 - P_{ij})\right]$$

is positive as a result of the Cauchy-Schwartz inequality.

The initial values for the iterative process may be taken as follows:

$$\theta_i^{(0)} = \log\left\{\frac{p_i}{(1 - p_i)}\right\},$$

$$b_j^{(0)} = \frac{\Phi^{-1}(q_j)}{r_j},$$

and

$$a_j^{(0)} = \frac{r_j}{(1 - r_j^2)^{1/2}}.$$

Here

$$p_i = \sum_j \frac{u_{ij}}{n},$$

$$q_j = \sum_i \frac{u_{ij}}{N},$$

$\Phi^{-1}(q_j)$ is the normal deviate that cuts off an area q_j to its right, and r_j is the point biserial correlation between item score and the total score. A problem occurs if p_i is zero or one, and in these cases, p_i may be taken as $(n - 1/2)/n$ or $1/2n$, respectively. In the corresponding situation for q_j, q_j may be taken as $(N - 1/2)/N$ and $1/2N$, respectively.

As noted earlier, an identifying condition was incorporated in the specification of prior information for θ. Experience suggests that convergence is accelerated if at every stage the ability parameters are rescaled to have a mean of zero and a standard deviation of one. The item parameters are rescaled accordingly.

Specification of Prior Belief

The parameters for the prior distribution of a_j, the item discrimination parameter, are the degrees of freedom v_j and the scale parameter ω_j of the chi distribution. The values of these parameters may be chosen by stating the end points of a $(1 - \alpha)$th percent credibility interval for a_j since there is a one-to-one correspondence between v_j, ω_j, and the highest density credibility interval. Unfortunately, limits for the credibility intervals for the chi distribution cannot be obtained analytically and hence approximate methods have to be employed. The normal approximation to the chi distribution discussed in Appendix A provides a simple method to obtain the $(1 - \alpha)$th percent credibility interval for a_j.

For large v_j, the end points of the $(1 - \alpha)$th percent credibility interval are given as

$$\omega_j^{1/2}\left\{(v_j - \tfrac{1}{2})^{1/2} \pm \frac{z_{1/2\alpha}}{2^{1/2}}\right\}$$

Where $z_{1/2\alpha}$ is the upper $\frac{1}{2}\alpha$ percentage points of the unit normal curve. For example, $v_j = 10$ and $\omega_j = .1$ yield the approximate 99% central credibility interval:

$$.40 < a_j < 1.55.$$

Although this interval is not exact, the approximation is reasonable for $v \geq 10$.

In general, if H denotes the upper limit and L denotes the lower limit for the $(1 - \alpha)$th percent credibility interval, then

$$\omega_j^{1/2}\left\{(v_j - \tfrac{1}{2})^{1/2} + \frac{z_{1/2\alpha}}{2^{1/2}}\right\} = H$$

and

$$\omega_j^{1/2}\left\{(v_j - \tfrac{1}{2})^{1/2} - \frac{z_{1/2\alpha}}{2^{1/2}}\right\} = L.$$

Solving these two equations, we obtain

$$v_j = \frac{1}{2}\left[1 + z_{1/2\alpha}^2\left\{\frac{H + L}{H - L}\right\}^2\right]$$

and

$$\omega_j = \frac{1}{2}\left\{\frac{H - L}{z_{1/2\alpha}}\right\}^2.$$

The degrees of freedom v_j could be rounded off to the nearest integer, if so desired. Thus, with a specification of the end points of an interval for a_j and the "degree of confidence," it is possible to specify the prior distribution of a_j. Experience with data suggests that $v_j = 10$, and $\omega_j = .1$ provide a reasonably good description of the prior distribution for a_j, and for convenience, may be taken to be the same for all a_j.

If non-vague priors are specified for θ_i and b_j, then values of v_b and λ_b have to be given. The mean value for the inverse chi-square distribution (Novick & Jackson, 1974, p. 191) is $\lambda/(v - 2)$. Unless $v > 4$, the variance of the distribution is not defined. Furthermore, if $v < 7$, the mode of the distribution does not lie within the interquartile range. Thus, it seems necessary to specify a value of v at least as large as 7. However, as v gets large, the distribution of ϕ becomes concentrated around the mean implying that precise information concerning ϕ is available. To avoid this, v should be set at a reasonable value, in the vicinity of 7. The choice of λ is governed by a similar argument—it should be reasonably large, in the vicinity of 10 since, λ being a scale parameter, for small values of λ the distribution becomes concentrated.

A consideration of the information provided by the credibility interval for ϕ generated with the specified values of v and λ may be helpful. For example, the 99% credibility interval for ϕ_b given $v = 10$ and $\lambda = 10$ is [.3, 4.0]. This implies that the 99% credibility interval for b_j have widths ranging from 1.2 (when the lower value of ϕ_b, .3, is taken) to 16.0 (when the larger value of ϕ_b, 4.0, is taken). For an extreme value of b_j around, say, 2.0, the credibility intervals, corresponding to the limits of the interval for ϕ_b, range from [1.4, 2.6] to [−6, 10]. The information provided by such credibility intervals is not too precise and not too vague. Thus, $v = 10$ and $\lambda = 10$ may be considered reasonable values. Experience with artificially generated data supports this and reveals that the estimation is robust with respect to these hyperparameters, the reason being that the *form* of the prior distribution contains most of the information.

Comparison Studies

In order to illustrate the applicability of the Bayesian procedure, a simulation study was carried out. In the present context, such a study is important since the true values of the parameters can be compared with the estimates.

Item responses for N individuals ($N = 50, 100, 200, 500$) and n items ($n = 15, 25, 35$)

were artificially generated using the computer program DATGEN (Hambleton & Rovinelli, 1973) according to the two-parameter logistic model.

The prior distributions for θ_i and b_j were chosen to be uniform, to demonstrate the efficacy of the Bayesian procedure. In this case estimates of ability and item parameters do not exist when p_i(or q_i) is either zero or one. An advantage of the Bayesian procedure is clearly lost with this prior specification. The prior distribution for a_j was chosen as the chi distribution with $v_j = 10$ and $\omega_j = .1$. As indicated earlier, these produced the 99% credibility intervals, [.40, 1.55], for each a_j. It should be noted that the same prior distribution is specified for each a_j merely for convenience. The accuracy of the estimates can be considerably improved by specifying a more appropriate distribution for each a_j.

The data were generated in such a way as to not reproduce the prior distributions. Since the prior distributions for θ_i and b_j were taken as uniform, these parameters were generated as samples from a *normal* distribution with zero mean and unit variance. The values for a_j, on the other hand, were generated from a *uniform* distribution in the interval [.6, 1.90] since an informative prior was employed for a_j. It was hoped that these data generation considerations would also attest to the robustness of the Bayesian procedure.

The estimates for the item and ability parameters were obtained using the LOGIST program (Wood, Wingersky, & Lord, 1976) which yielded maximum likelihood estimates of the parameters and a program specifically written for the Bayes procedure. The two sets of estimators, Maximum Likelihood (ML) and Bayes, were compared with respect to accuracy in terms of: (a) the correlation between the true values and the estimates, (b) the mean squared difference between the true values and the estimated values, and (c) the scatter plot of estimates against the true values. Mean squared deviation as measured by the squared deviations from the *mean* was not employed as a criterion since it was felt that as a result of regression towards the mean, this would clearly favor the Bayesian estimation procedure.

The results of the simulation studies are summarized in Table 2. The Bayesian estimates are superior to the ML estimates in that (a) the correlations between the true values and estimates are higher, and (b) the mean square differences are lower. The most dramatic improvement that results from the Bayesian procedure is that no estimates drift out of bounds. This is clearly evident with the discrimination parameter. In order to prevent the drift of the estimates of the discrimination parameter, the LOGIST program requires that an upper limit be set on the values taken by the estimates. For the purpose of this study, an upper limit of ten was specified for the estimates of the discrimination parameter. The number in parentheses in Table 2 indicates the number of estimates of the discrimination parameter that reached the maximum value of 10 (with difficulty and ability estimates, the number in parentheses indicates the number of estimates that were beyond ± 5.0). Clearly, the Bayesian procedure arrests this drift.

An examination of the mean squared differences reveals an interesting fact. Since the mean squared differences (MSD) is the average of the squares of the differences between the true values and the estimates, a small value indicates the accuracy of estimation. The Bayesian procedure yields consistently smaller MSD than the ML procedure. This trend reveals that the Bayesian estimates are less biased than the ML estimates, a surprising fact indeed. This may be partly explained by the fact that the discrimination parameter is estimated more accurately by the Bayesian procedure, and this in turn yields more accurate estimates of ability and difficulty parameters. However, as the number of items and examinees increase, the two procedures yield similar results, a trend that is clearly evident with increasing number of items. This result can be anticipated since, in large samples, the likelihood dominates the prior and the Bayesian estimates are indistinguishable from the ML estimates.

Table 2

Accuracy of Estimation in the Two-Parameter Model

		Difficulty				Discrimination				Ability			
		Bayes		ML		Bayes		ML		Bayes		ML	
n	N	r	MSD	r	MSD	r	MSD	r	MSD	r	MSD	r	MSD
	50	.990	.041	.763	7.984(2)*	.706	.041	.580	14.519(3)	.934	.129	.853	1.261(3)
	100	.992	.033	.899	.735	.580	.055	.214	10.234(2)	.930	.139	.876	.339
15	200	.996	.014	.978	.057	.673	.047	.201	9.973(2)	.927	.145	.907	.185
	500	.998	.005	.985	.052	.896	.028	.324	4.959(1)	.923	.154	.901	.198
	50	.980	.048	.927	.163	.548	.067	.505	9.419(3)	.951	.096	.880	.235
	100	.989	.026	.968	.067	.677	.056	.412	5.808(2)	.956	.087	.921	.157
25	200	.995	.012	.994	.031	.896	.026	.862	.077	.955	.089	.952	.096
	500	.998	.005	.996	.008	.941	.012	.909	.058	.953	.093	.951	.097
	50	.977	.057	.674	6.709(2)	.545	.067	.404	6.378(3)	.965	.068	.935	.198
	100	.984	.032	.971	.074	.790	.045	.700	.356	.967	.066	.959	.082
35	200	.990	.021	.980	.040	.884	.022	.842	.103	.970	.060	.965	.072
	500	.996	.009	.991	.020	.935	.013	.934	.028	.966	.068	.964	.076

*The number in parentheses indicates the number of estimates with absolute values exceeding five.
n = Number of items.
N = Number of examinees.
r = Correlation between estimates and true values
MSD = Mean Squared Differences between estimates and true values

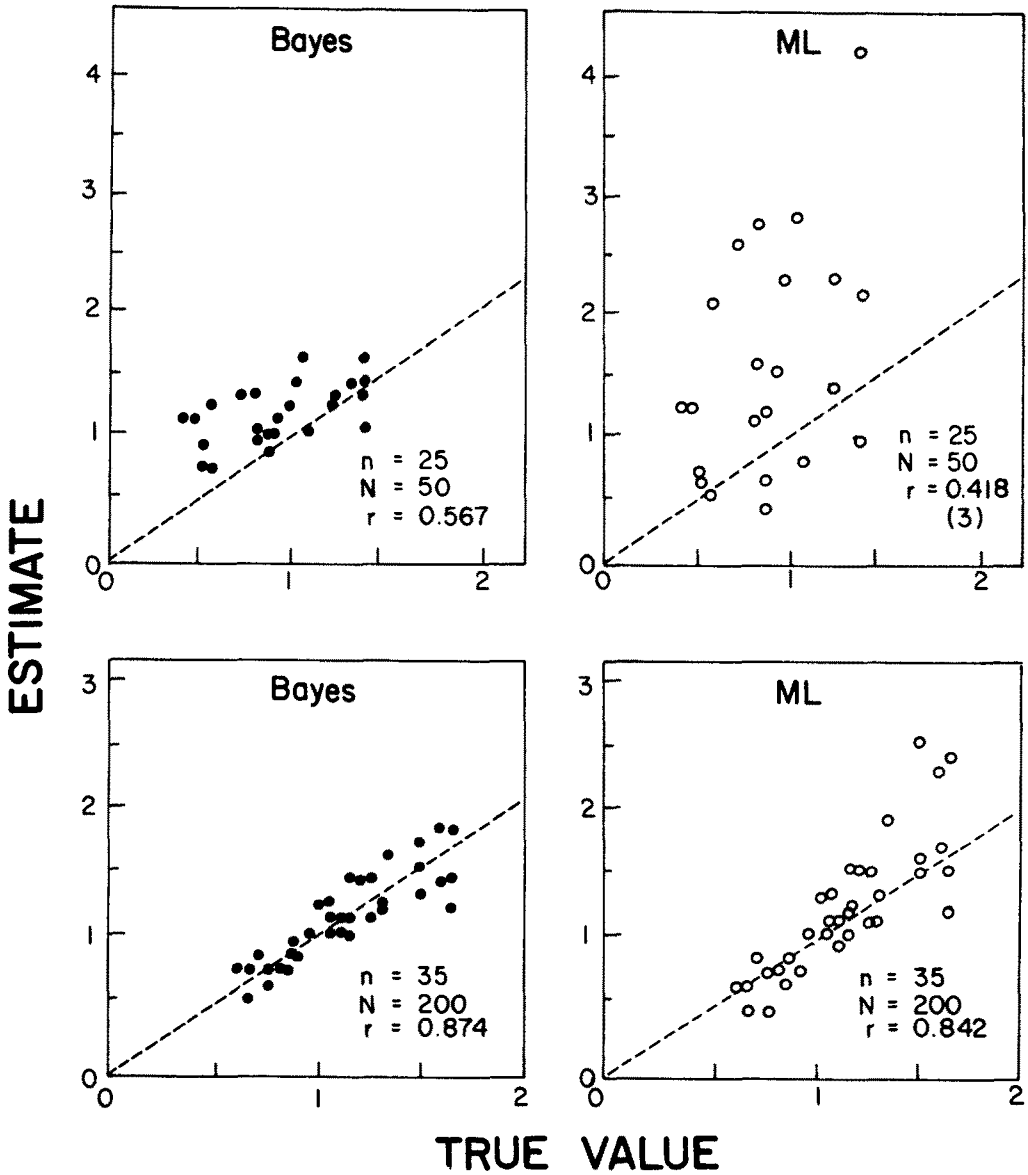

FIGURE 1
Plot of estimated values against true values of the difficulty parameter.

The scatter plots depicted in Figures 1–2 confirm the results provided in Table 2. The scatter of the Bayesian estimates against the true values is in general less than that observed with the ML estimates, particularly in small samples, and specifically for the discrimination parameter. As the number of items and examinees increases, the scatter about the line $y = x$ decreases, indicating that the estimators are consistent, a finding that agrees with the result obtained by Swaminathan and Gifford (1983).

Conclusion

The Bayesian procedure described in this paper for estimating parameters in the two-parameter logistic item response model provides an attractive alternative to the max-

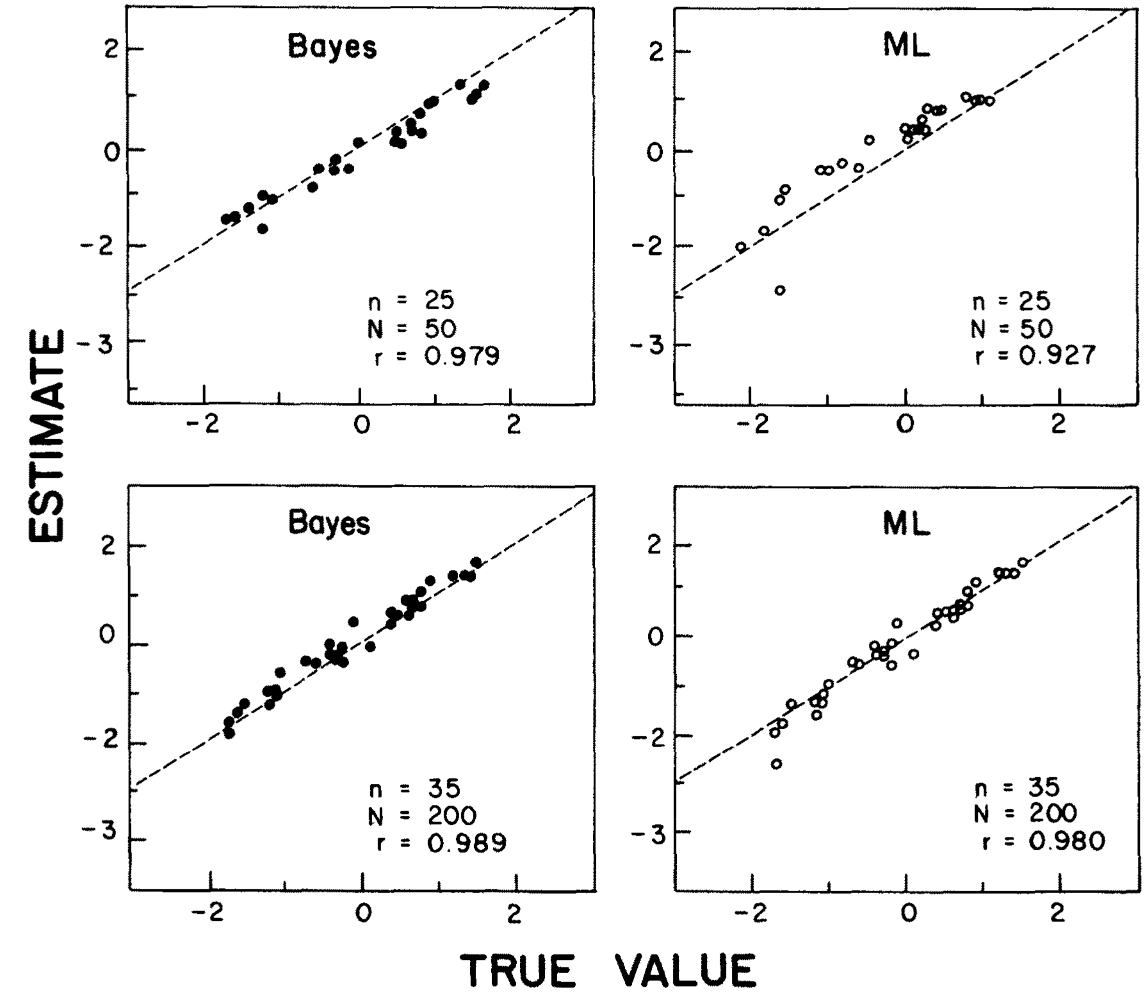

FIGURE 2
Plot of estimated values against true values of the discrimination parameters.

imum likelihood procedure. The most important feature of the Bayesian procedure is that the estimates of the item and ability parameters are well behaved and do not drift out of bounds.

In the illustrative study that was conducted, non-informative (uniform) prior distributions were specified for the difficulty and ability parameters. When an informative prior is specified for the discrimination parameter, the estimation of all the parameters proceeds smoothly. Hence, specification of non-informative priors for difficulty and ability parameters with an informative prior for the discrimination parameter appears to be a reasonable approach. Since this procedure decreases the regression effect for the estimation of ability parameters, this procedure appears even more attractive.

Specification of prior information for the discrimination parameter can be accomplished in a straightforward manner, given the accuracy of the normal approximation to the chi distribution. Furthermore, initial investigations have revealed that poor specification of v and ω have little effect on the estimation of the discrimination parameter. In this sense, the specification of prior information on the discrimination parameter may be viewed as specifying a penalty function, the main purpose of which is to prevent drift of the estimates into unacceptable regions.

The Bayesian procedure is particularly effective with small samples. However, the Bayesian procedure may be successfully employed in larger samples since even in large samples, outward drift of ML estimates has been observed. The Bayesian procedure can be considered a modification of the ML procedure and the computational expense involved is minor compared to the advantages that result from it.

Appendix A

The chi distribution is defined as the square root of the chi-square distribution. Its density function can be shown to be

$$f(\chi \mid \nu, \omega)\, d\chi = \left[\omega^{\nu/2} 2^{\nu/2-1} \Gamma\left(\frac{\nu}{2}\right)\right]^{-1} \chi^{\nu-1} \exp\left(\frac{-\chi^2}{2\omega}\right) d\chi, \qquad (\chi > 0) \tag{A1}$$

where ν is the degrees of freedom and ω is a scale parameter. The rth moment of χ is

$$\mu_r' = \frac{2^{r/2}\omega^{r/2}\Gamma\left(\frac{\nu+r}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \tag{A2}$$

with the mode at $[\omega(\nu - 1)]^{1/2}$. In particular,

$$\mu_1' = \frac{2^{1/2}\omega^{1/2}\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{A3}$$

and

$$\mu_2' = \omega\nu. \tag{A4}$$

Approximations for the first moment have been given by Johnson and Welch (1939) and have been discussed by Kendall and Stuart (1969, p. 371). With these approximations, it is relatively simple to evaluate the mean, μ and standard deviation, σ, of the χ distribution given (approximately) as

$$\mu = [\omega(\nu - \tfrac{1}{2})]^{1/2} \tag{A5}$$

and

$$\sigma = \left(\frac{\omega}{2}\right)^{1/2} \tag{A6}$$

For $\omega = 1$, and for ν as small as 10, these approximations have relative errors less than .1% and 1.5%, respectively. Furthermore, as demonstrated by Kendall and Stuart (1969, p. 372), for sufficiently large ν,

$$\chi_{\nu,\omega} \sim N(\mu, \sigma). \tag{A7}$$

This normal approximation is reasonably good even for ν as small as 10. (Figure A1 demonstrates the convergence to normality). More important is the fact that this normal

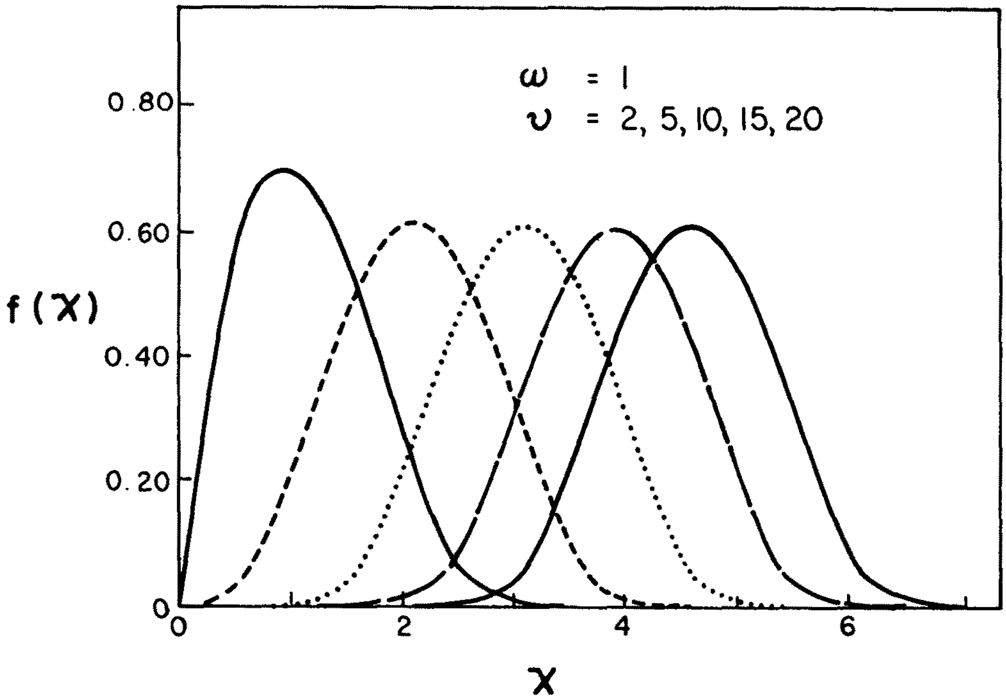

FIGURE A1
Chi distribution.

approximation enables the user to easily choose the values of ν and ω through specifying the end points of a credibility interval for the discrimination parameters.

References

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society* (Series B), *34*, 42–54.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*, 258–276.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443–460.

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

Haberman, S. (1975). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*, 815–841.

Hambleton, R. K., & Rovinelli, R. (1973). A FORTRAN IV program for generating examinees response data from logistic test models. *Behavioral Science*, *18*, 74.

Johnson, N. L., & Welch, B. L. (1939). On the calculation of the cumulants of the χ distribution. *Biometrika*, *31*, 216–218.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics*: *Vol. I*. New York: Hafner.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 887–890.

Lindley, D. V. (1971). The estimation of many parameters. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart, & Winston, pp. 435–455.

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

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*, 989–1020.

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

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179–193.

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

Novick, M. R., & Jackson, P. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Novick, M. R., Lewis, C., & Jackson, P. H. (1973). The estimation of proportions in *m* groups. *Psychometrika, 38*, 19–46.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351–356.

Slinde, J. A. & Linn, R. L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16*, 159–165.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175–191.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service. (Revised 1978)

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*, New York: John Wiley & Sons.

Manuscript received 10/30/81
Final version received 4/1/85