Journal of Educational Statistics
Fall 1982, *Volume* 7, *Number* 3, *pp. 175–191*

BAYESIAN ESTIMATION IN THE RASCH MODEL

HARIHARAN SWAMINATHAN AND JANICE A. GIFFORD
University of Massachusetts

Key words: *Rasch model*, *Bayesian estimation*, *hierarchical model*

ABSTRACT. Bayesian estimation procedures based on a hierarchical model for estimating parameters in the Rasch model are described. Through simulation studies it is shown that the Bayesian procedure is superior to the maximum likelihood procedure in that the estimates are (a) more accurate, at least in small samples; and (b) meaningful in that parameters corresponding to perfect item and ability responses can be estimated.

In recent years there has been considerable interest among measurement theorists and practitioners in latent trait theory because it offers the potential for improving educational and psychological measurement practices. However, before latent trait theory can be successfully applied to solve existing measurement problems, the problem of estimating parameters in latent trait models has to be addressed.

The literature in latent trait theory abounds with procedures for the estimation of parameters. The estimation procedures that have been developed over the past 30 years range from heuristic procedures such as those given by Urry (1974) and Jensema (1976) to conditional as well as unconditional maximum likelihood procedures (Andersen, 1970, 1972, 1973a, 1973b; Bock, 1972; Lord, 1968, 1974; Samejima, 1969, 1972; Wright & Douglas, 1977; Wright & Panchapakesan, 1969). With the exception of the "conditional" maximum likelihood procedure provided by Andersen (1970) for the one-parameter model, the maximum likelihood estimators of the parameters in the latent trait models are less than optimal as a result of the problem of estimating "structural parameters" in the presence of "incidental parameters" (Andersen, 1970; Zellner, 1971, pp. 114–154). The "structural parameters" in latent trait models are the item parameters while the "incidental parameters" are the ability parameters because these increase without bound as the number of examinees is increased to provide stable estimates of the parameters.

When several parameters have to be estimated simultaneously, and when, as in the present case, both structural and incidental parameters have to be estimated, a Bayesian solution to the estimation problem may be appropriate (Zellner, 1971, pp. 114–119). This is particularly true if prior information or belief about the parameters is available, because in this case, the incorporation of this information will certainly increase the "accuracy" or the meaningfulness of the estimates. An example of this was encountered by Lord (1968), where to prevent estimates of the item discrimination parameter from drifting out of bounds, it was necessary to impose limits on the range of values the parameter could take. Although the estimation procedure employed by Lord (1968) was not Bayesian, this illustrates the role of prior information in obtaining meaningful estimates.

Bayesian procedures have been successfully applied in a variety of situations. For a sampling of these applications the reader is referred to Novick and Jackson (1974) and Zellner (1971). Birnbaum (1969) obtained Bayes estimates of the ability parameter in the one- and two-parameter logistic models under the assumption that the item parameters were known. He chose, for mathematical tractability, the prior pdf of θ_i, the ability of the ith examinee, to be the logistic density function; that is,

$$p(\theta_i) = \exp(-D\theta_i)/[1 + \exp(-D\theta_i)]^2,$$

where $D = 1.7$ is a scaling factor. Owen (1975), in applying the latent trait model in an adaptive testing context, obtained Bayes estimates of ability, θ_i, under the assumption that the prior pdf of θ_i was normal with zero mean and unit variance.

The Bayesian procedures suggested by Birnbaum (1969) and Owen (1975) require rather exact specification of prior belief. An alternative and a more powerful procedure has been suggested by Lindley (1971). He has shown that if the information that is available can be considered *exchangeable*, then a hierarchical Bayesian model can be effectively employed for the estimation of parameters.

The hierarchical Bayesian model has been successfully employed by Lindley and Smith (1972), Novick, Lewis, and Jackson (1973), and Zellner (1971), to name a few. However, this approach has not been employed for estimating parameters in latent trait models. This paper provides a Bayesian estimation procedure, in the sense of Lindley, for estimating parameters in the one-parameter latent trait model.

The Model

Let X_{ij} denote a random variable that represents the binary response of an examinee i ($i = 1,\ldots,N$) on item j ($j = 1,\ldots,n$). If the examinee responds correctly to the item, $X_{ij} = 1$, while for an incorrect response, $X_{ij} = 0$. We assume that the complete latent space is unidimensional, and that the probability, $P(X_{ij} = 1)$, that an individual with ability parameter θ_i will correctly respond to an item with difficulty parameter, b_j, is given by the logistic model

$$P(X_{ij} = 1 \mid \theta_i) = \exp(\theta_i - b_j)/[1 + \exp(\theta_i - b_j)]. \tag{1}$$

On the other hand, the probability that the individual will respond incorrectly is given by

$$\begin{aligned} P(X_{ij} = 0 \mid \theta_i) &= 1 - P(X_{ij} = 1 \mid \theta_i) \\ &= 1/[1 + \exp(\theta_i - b_j)]. \end{aligned} \tag{2}$$

The probabilities given in Equations (1) and (2) can be combined to yield

$$P(X_{ij} = x_{ij} \mid \theta_i) = \exp[x_{ij}(\theta_i - b_j)]/[1 + \exp(\theta_i - b_j)], \tag{3}$$

where $x_{ij} = 1$ for a correct response and $x_{ij} = 0$ for an incorrect response.

Because it depends only on one item parameter, difficulty, the above model is commonly known as the Rasch model or the one-parameter logistic model. For a detailed description of this model and its properties, the reader is referred to Wright (1977).

Estimation

Conditional Estimation of Ability

In some situations it may be of interest to estimate the ability θ_i of an examinee who takes a test that has been calibrated; that is, the difficulty parameters are known. Moreover, because the problem of estimating ability when the item parameters are known is simpler to deal with and provides an illustration of the basic ideas involved, this case will be dealt with in detail first.

The model given by Equation (3) should in the strict sense be expressed as

$$P(X_{ij} = x_{ij} \mid \theta_i, b_j) = \exp[x_{ij}(\theta_i - b_j)]/[1 + \exp(\theta_i - b_j)]. \tag{4}$$

Although there are several ways to write the model, the expression given by (4) is the most convenient for the present situation.

It follows from the principle of local independence that the joint probability of responses of the N examinees on n items is given by

$$\begin{aligned} P(X_{11} = x_{11}, X_{12} = x_{12}, \ldots, X_{ij} = x_{ij}, \ldots,& \\ X_{Nn} = x_{Nn} \mid \theta_1, \theta_2, \ldots, \theta_N; b_1, b_2, \ldots, b_n)& \\ = \prod_{i=1}^{N} \prod_{j=1}^{n} \exp[x_{ij}(\theta_i - b_j)]/[1 + \exp(\theta_i - b_j)].& \end{aligned} \tag{5}$$

Once the responses of the N examinees on the n items are observed, the above expression ceases to have the probability interpretation and becomes the likelihood function, $L(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b})$. On simplification,

$$L(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b}) = \exp\left[\sum_i \sum_j x_{ij}(\theta_i - b_j)\right] \Big/ \prod_i \prod_j \left[1 + \exp(\theta_i - b_j)\right]$$

$$= \exp\left[\sum_i r_i\theta_i - \sum_j q_j b_j\right] \Big/ \prod_i \prod_j \left[1 + \exp(\theta_i - b_j)\right], \quad (6)$$

where $r_i = \Sigma_j x_{ij}$, and $q_j = \Sigma_i x_{ij}$. Because the item parameters are known constants, the likelihood function is strictly a function of $\boldsymbol{\theta}$ and, hence, can be expressed as

$$L(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b}) \propto \exp\left(\sum_i r_i\theta_i\right) \Big/ \prod_i \prod_j \left[1 + \exp(\theta_i - b_j)\right]. \quad (7)$$

To obtain the posterior density function of $\boldsymbol{\theta}$ given the observations and the item parameters, it is necessary to specify the prior distribution of $\boldsymbol{\theta}$. To this end, in the first stage of the hierarchical model, we assume a priori, that the ability parameters, θ_i, are independently and identically normally distributed; that is,

$$\theta_i \mid \mu, \phi \sim N(\mu, \phi). \quad (8)$$

The assumption that the thetas are independently and identically distributed follows from the assumption of exchangeable prior information about the thetas. The assumption of normality also appears to be reasonable and has been made by numerous authors, for example, Lord and Novick (1968).

To complete the hierarchical Bayesian model, we have to specify prior distributions for μ and ϕ. This is the second stage. At this level, we assume that, a priori, μ and ϕ are independently distributed, and that μ has the uniform distribution. Thus,

$$p(\mu, \phi) \propto p(\phi). \quad (9)$$

The uniform distribution is not a proper distribution, although this choice can be justified to some extent (Zellner, 1971, pp. 41–43). It may, however, be more appropriate to specify a "nondiffuse" prior (this possibility will be explored further in a later paper).

It now remains to specify the form of $p(\phi)$. Because ϕ is the variance of θ_i, ϕ can be assumed to have the inverse chi-square distribution; that is,

$$p(\phi \mid \nu, \lambda) \propto \phi^{-(\nu/2+1)} \exp(-\lambda/2\phi). \quad (10)$$

The quantities ν and λ are parameters of the inverse chi-square distribution and must be specified a priori. We shall return to this in a later section.

From Bayes' theorem, it follows that the joint posterior distribution of $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_N)$ given $\mathbf{b}$ and the item responses is

$$p(\boldsymbol{\theta}, \mu, \phi \mid \mathbf{b}, \mathbf{x}) \propto L(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b})\, p(\boldsymbol{\theta} \mid \mu, \phi)\, p(\mu)\, p(\phi \mid \nu, \lambda). \quad (11)$$

The likelihood function $L(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b})$ is given by Equation (7), $p(\mu, \phi)$ by Equation (9), and

$$p(\boldsymbol{\theta} \mid \mu, \phi) = \prod_{i=1}^{N} \phi^{-1/2} \exp\left[-\frac{1}{2}(\theta_i - \mu)^2/\phi\right]$$

$$= \phi^{-N/2} \exp\left[-\sum_{i=1}^{N} (\theta_i - \mu)^2 \Big/ 2\phi\right]. \tag{12}$$

Combining these expressions, we have

$$p(\boldsymbol{\theta}, \mu, \phi \mid \mathbf{b}, \mathbf{x}, \nu, \lambda) \propto \left\{\exp\left[\sum_i r_i\theta_i\right] \Big/ \prod_i \prod_j \left[1 + \exp(\theta_i - b_j)\right]\right\}$$

$$\left\{\phi^{-N/2} \exp\left[-\sum_i (\theta_i - \mu)^2 \Big/ 2\phi\right]\right\}\left[\phi^{-(\nu/2+1)} \exp(-\lambda/2\phi)\right]. \tag{13}$$

The above expression depends on the "nuisance" parameters μ and ϕ and hence these have to be integrated out. Because $\Sigma(\theta_i - \mu)^2 = \Sigma(\theta_i - \theta.)^2 + N(\theta. - \mu)^2$, where $\theta.$ is the mean of the thetas, and

$$\int_{-\infty}^{\infty} \exp -\left[N(\theta. - \mu)^2/2\phi\right] d\mu \propto \phi^{1/2}, \tag{14}$$

integration with respect to μ yields

$$p(\boldsymbol{\theta}, \phi \mid \mathbf{b}, \mathbf{x}, \nu, \lambda) \propto L(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b})\phi^{-(N+\nu+1)/2} \exp\left\{-\left[\lambda + \sum (\theta_i - \theta.)^2\right] \Big/ 2\phi\right\}. \tag{15}$$

Noting that

$$\int_0^{\infty} \phi^{-m} \exp(-k/\phi)\, d\phi \propto k^{-(m-1)}, \tag{16}$$

and integrating with respect to ϕ, we obtain

$$p(\boldsymbol{\theta} \mid \mathbf{b}, \mathbf{x}, \nu, \lambda) \propto L(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{b})\left[\lambda + \sum_i (\theta_i - \theta.)^2\right]^{-(N+\nu-1)/2} \tag{17}$$

$$= \left\{\exp\left[\sum_i r_i\theta_i\right] \Big/ \prod_i \prod_j \left[1 + \exp(\theta_i - b_j)\right]\right\}$$

$$\cdot \left[\lambda + \sum_i (\theta_i - \theta.)^2\right]^{-(N+\nu-1)/2}. \tag{18}$$

The posterior density function given by (18) contains all the information required for making probability statements about the parameters. However, given its complexity, it is not in a usable form. Point estimates of the parameters are useful in such situations. Following Lindley and Smith (1972) and Novick, Lewis, and Jackson (1973), we will take the joint modal estimates of the parameters as the appropriate estimates.

The joint posterior modes are obtained by differentiating $\log p(\boldsymbol{\theta} \mid \mathbf{b}, \mathbf{x}, \nu, \lambda)$ with respect to $\boldsymbol{\theta}$, setting these derivatives equal to zero, and solving the resulting equations:

$$\sum_{j=1}^{n} P_{ij} = r_i - (\theta_i - \theta.)/\sigma^2 \qquad (i = 1, \ldots, N), \tag{19}$$

where

$$P_{ij} = \exp(\theta_i - b_j)/\left[1 + \exp(\theta_i - b_j)\right],$$

and

$$\sigma^2 = \left[\lambda + \sum_{i=1}^{N} (\theta_i - \theta.)^2\right] \Big/ (\nu + N - 1).$$

Because this system of equations is nonlinear, numerical procedures must be employed. The Newton-Raphson iterative procedure is ideally suited for this situation. Let

$$f(\theta_i) = \sum_{j=1}^{n} P_{ij} + (\theta_i - \theta.)/\sigma^2 - r_i. \tag{20}$$

Then

$$f'(\theta_i) = \sum_{j=1}^{n} P_{ij}(1 - P_{ij}) + \left[\sigma^2\left(1 - \frac{1}{N}\right) - 2(\theta_i - \theta.)^2/(\nu + N - 1)\right] \Big/ (\sigma^2)^2. \tag{21}$$

If $\theta_i^{(k)}$ is the value of θ_i at the kth iteration, then $\theta_i^{(k+1)}$ is given by

$$\theta_i^{(k+1)} = \theta_i^{(k)} - f\left[\theta_i^{(k)}\right]/f'\left[\theta_i^{(k)}\right] \tag{22}$$

The starting value, $\theta_i^{(o)}$, given by Wright & Douglas (1977) is

$$\theta_i^{(o)} = b. + \left(1 + s_b^2/2.89\right)\log(r_i/n - r_i), \tag{23}$$

where

$$b. = \sum b_j/n, \quad \text{and} \quad s_b^2 = \sum \left(b_j - b.\right)^2/(n - 1).$$

The starting values given above are clearly invalid when the raw score, r_i, is zero or n. Moreover, neither conditional nor unconditional maximum likelihood estimators exist in these cases. Bayes estimators do exist in these situations. For the case when $r_i = 0$ or n, starting values may be obtained by setting $r_i = \frac{1}{2}$ or $n - \frac{1}{2}$, respectively.

It should also be pointed out that the Newton-Raphson scheme given above is not the vector version of the procedure since for this procedure the matrix of derivatives $\{\partial f/\partial\theta_i\partial\theta_j\}$ has to be computed and inverted. The procedure described here worked sufficiently well, converging in as few as three to four iterations.

Joint Estimation of Item and Ability Parameters

The case considered above, where the item parameters were assumed to be known, provides the necessary background for the Bayesian estimation procedure. However, this situation may not be realistic and therefore it is necessary to develop a procedure for the joint estimation of the item and ability parameters.

We proceed in the manner indicated for the case of known item parameters. In addition to making the assumptions about the ability parameters, we must make assumptions regarding the item parameters. Again, as in the previous case, we specify prior beliefs about the parameters in two stages: In the first stage, for the model given in (3), we assume:

$$\theta_i \mid \mu_\theta, \phi_\theta \sim N(\mu_\theta, \phi_\theta) \qquad (i = 1, \ldots, N), \tag{24a}$$

$$b_j \mid \mu_b, \phi_b \sim N(\mu_b, \phi_b) \qquad (j = 1, \ldots, n). \tag{24b}$$

In addition, we assume that, a priori, θ_i and b_j are independent, θ_k and θ_l $(k \neq l)$ are independent, and b_k and b_l are independent.

As with the ability parameters, the specification of normal prior for b_j seems reasonable. This assumption has been made by several authors (Lord & Novick, 1968; Wright & Douglas, 1977). Furthermore, as a result of the hierarchical Bayesian model, departures from this assumption appear to have a negligible effect on the estimates of b_j.

For the second stage, we assume that a priori, the hyperparameters are independent, and that the prior information about the parameters, μ_θ and μ_b, is uniform.

$$\begin{aligned} p(\mu_\theta, \phi_\theta) &\propto p(\phi_\theta) \\ &\propto \phi_\theta^{-(\nu_\theta/2+1)} \exp(-\lambda_\theta/2\phi_\theta), \end{aligned} \tag{25}$$

and

$$\begin{aligned} p(\mu_b, \phi_b) &\propto (p\phi_b) \\ &\propto \phi_b^{-(\nu_b/2+1)} \exp(-\lambda_b/2\phi_b). \end{aligned} \tag{26}$$

The joint posterior pdf of $\boldsymbol{\theta}$ and $\mathbf{b}$ is given by

$$p(\boldsymbol{\theta}, \mathbf{b}, \mu_\theta, \phi_\theta, \mu_b, \phi_b \mid \mathbf{x}, \nu_\theta, \lambda_\theta, \nu_b, \lambda_b)$$

$$\propto L(\boldsymbol{\theta}, \mathbf{b} \mid \mathbf{x}) \left[\prod_{i=1}^{N} p(\theta_i \mid \mu_\theta, \phi_\theta) \prod_{j=1}^{n} p(b_j \mid \mu_b, \phi_b) \right] p(\phi_\theta \mid \nu_\theta, \lambda_\theta) p(\phi_b \mid \nu_b, \lambda_b), \tag{27}$$

where $L(\boldsymbol{\theta}, \mathbf{b} \mid \mathbf{x})$ is the likelihood function given by (6).

On integrating with respect to ϕ_θ and μ_θ, we have, from (17),

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\left[\prod_{i=1}^{N} p(\theta_i \mid \mu_\theta, \phi_\theta)\right] p(\phi_\theta \mid \nu_\theta, \lambda_\theta)\, d\mu_\theta\, d\phi_\theta \propto \left[\lambda_\theta + \sum_{i=1}^{N} (\theta_i - \theta.)^2\right]^{-(N+\nu_\theta-1)/2}. \quad (28)$$

Similarly,

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\left[\prod_{j=1}^{n} p(b_j \mid \mu_b, \phi_b)\right] p(\phi_b \mid \nu_b, \lambda_b)\, d\mu_b\, d\phi_b \propto \left[\lambda_b + \sum_{j=1}^{n} (b_j - b.)^2\right]^{-(n+\nu_b-1)/2}. \quad (29)$$

Combining (27), (28), and (29), we obtain the joint posterior density of $\boldsymbol{\theta}$ and $\mathbf{b}$:

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{b} \mid \mathbf{x}, \nu_\theta, \lambda_\theta, \nu_b, \lambda_b) &= \left\{\left[\exp\left(\sum_{i=1}^{N} r_i\theta_i\right)\right]\left[\lambda_\theta + \sum_{i=1}^{N} (\theta_i - \theta.)^2\right]^{-(N+\nu_\theta-1)/2}\right\} \\ &\cdot \left\{\left[\exp\left(-\sum_{j=1}^{n} q_j b_j\right)\right]\left[\lambda_b + \sum_{j=1}^{n} (b_j - b.)^2\right]^{-(n+\nu_b-1)/2}\right\} \\ &\cdot \left\{\prod_{i=1}^{N}\prod_{j=1}^{n}\left[1 + \exp(\theta_i - b_j)\right]\right\}^{-1}. \end{aligned} \quad (30)$$

Now

$$L(\boldsymbol{\theta}, \mathbf{b} \mid \mathbf{x}) = \prod_i \prod_j \exp(\theta_i - b_j) \Big/ \left[1 + \exp(\theta_i - b_j)\right],$$

and, hence, is bounded. In fact,

$$\mid L(\boldsymbol{\theta}, \mathbf{b} \mid \mathbf{x}) \mid \leqslant 1.$$

Therefore, it follows that

$$\int \mid (p(\boldsymbol{\theta}, \mathbf{b}) \mid \ldots) \mid d\boldsymbol{\theta}\, d\mathbf{b} < \int \left[\lambda_\theta + \sum_{i=1}^{N} (\theta_i - \theta.)^2\right]^{-(N+\nu_\theta-1)/2} d\boldsymbol{\theta} \cdot \int \left[\lambda_b + \sum_{j=1}^{n} (b_j - b.)^2\right]^{-(n+\nu_b-1)/2} d\mathbf{b}.$$

The integrals on the right of the inequality clearly exist since the kernels are those of multivariate t densities. Hence, the posterior pdf, $p(\boldsymbol{\theta}, \mathbf{b} \mid \mathbf{x}, \nu_\theta, \lambda_\theta, \nu_b, \lambda_b)$, is a proper pdf, although the normalizing constant cannot be evaluated explicitly.

As before, we take the joint posterior modes as the estimates of θ_i and b_j $(i = 1,\ldots,N;\ j = 1,\ldots,n)$. These are obtained by setting equal to zero the derivatives of log $p(\boldsymbol{\theta}, \mathbf{b} \mid \ldots)$, and solving the resulting equations:

$$\sum_{j=1}^{n} P_{ij} = r_i - (\theta_i - \theta.)/\sigma_\theta^2 \qquad (i = 1,\ldots,N), \tag{31}$$

$$\sum_{i=1}^{N} P_{ij} = q_j + (b_j - b.)/\sigma_b^2 \qquad (j = 1,\ldots,n), \tag{32}$$

where

$$\sigma_\theta^2 = \left[\nu\lambda_\theta + \sum_i (\theta_i - \theta.)^2\right] \Big/ (\nu_\theta + N - 1),$$

and

$$\sigma_b^2 = \left[\nu\lambda_b + \sum_j (b_j - b.)^2\right] \Big/ (\nu_b + n - 1).$$

Because the system of equations is nonlinear, the Newton-Raphson procedure can again be employed to solve the equations iteratively. To accomplish this, we let

$$f(\theta_i) = \sum_{j=1}^{n} P_{ij} + (\theta_i - \theta.)/\sigma_\theta^2 - r_i, \tag{33}$$

and

$$h(b_j) = \sum_{i=1}^{N} P_{ij} - (b_j - b.)/\sigma_b^2 - q_j. \tag{34}$$

Then

$$f'(\theta_i) = \sum_{j=1}^{n} P_{ij}(1 - P_{ij}) + \left[\sigma_\theta^2\left(1 - \frac{1}{N}\right) - 2(\theta_i - \theta.)^2/(\nu_\theta + N - 1)\right] \Big/ (\sigma_\theta^2)^2, \tag{35}$$

and

$$h'(b_j) = -\sum_{i=1}^{N} P_{ij}(1 - P_{ij}) - \left[\sigma_b^2\left(1 - \frac{1}{n}\right) - 2(b_j - b.)^2/(\nu_b + n - 1)\right] \Big/ (\sigma_b^2)^2. \tag{36}$$

As before, if $\theta_i^{(k)}$ and $b_j^{(k)}$ denote the values of θ_i and b_j at the kth iteration, then

$$\theta_i^{(k+1)} = \theta_i^{(k)} - f\left(\theta_i^{(k)}\right)/f'\left(\theta_i^{(k)}\right), \tag{37}$$

and

$$b_j^{(k+1)} = b_j^{(k)} - h\left(b_j^{(k)}\right)/h'\left(b_j^{(k)}\right). \tag{38}$$

The initial values $\theta_i^{(0)}$ $(i = 1,\ldots,N)$ are given by (23), while

$$b_j^{(0)} = \log\left[(N - q_j)/q_j\right] \qquad (j = 1,\ldots,n).$$

When $q = 0$ or N, starting values are obtained by setting $q_j = \frac{1}{2}$ and $N - \frac{1}{2}$, respectively. With these values $\boldsymbol{\theta}$ is estimated. These values of $\boldsymbol{\theta}$ are then used to obtain revised estimates of **b**. This process is repeated with the revised estimates of **b** being used to obtain revised estimates of $\boldsymbol{\theta}$. The process is terminated when the convergence criterion is reached.

The model given by (1) is clearly unidentified when θ_i and b_j are unknown. To identify the model, we must set $\theta.$ (or $b.$) equal to zero. As Zellner (1971, pp. 253–258) has pointed out, this identification condition can be incorporated into the specification of prior information. Thus, setting $\theta.$ (or $b.$) equal to zero has the effect of specifying a prior distribution for θ_i (or b_j) with mean zero.

Specification of Prior Belief

The rth moment of the inverse chi-square distribution, μ_r, is given by

$$\mu_r = \left(\frac{\lambda}{2}\right)^r \Gamma\left(\frac{\nu}{2} - r\right)\Big/\Gamma\left(\frac{\nu}{2}\right).$$

It follows then that for the rth moment to be defined, $\nu > 2r$. In addition, as documented in Novick and Jackson (1974, p. 191), the descriptive statistics of the inverse chi-square distribution are:

$$\text{Mean} = \lambda(\nu - 2)^{-1}$$

and

$$\text{Standard Deviation} = \lambda(\nu - 2)^{-1}[2/(\nu - 4)]^{1/2};$$

for these to be defined, $\nu > 4$. Thus, it seems necessary to specify ν at least as large as 5. However, as ν gets large, the distribution of ϕ becomes concentrated around the mean, implying that precise information concerning ϕ is available. To avoid this, ν should be set at a reasonable value between 5 and 15. The choice of λ is governed by a similar argument. The expressions for the mean and standard deviation given above show that λ has a scaling effect on the distribution of ϕ. For values of λ close to zero, the standard deviation becomes small, concentrating the distribution around the mean. A large value of λ makes the distribution of ϕ diffuse while at the same time it increases the value of the mean, indicating that the distribution of θ_i (or b_j) has a large variance. In general, for a choice of ν between 5 and 15, the value, $\lambda = 10$, appears to reflect the distribution of θ_i and b_j in testing situations. The studies reported here indicate that the estimation procedure is robust when $\lambda = 10$ and ν is chosen to have a value between 5 and 15. Other simulation studies, not reported here, carried out with different values of λ also indicate that $\lambda = 10$ is an optimal value.

Comparison Studies

To study the efficacy of the Bayesian procedure described above and to compare the Bayesian estimates with the maximum likelihood estimates, studies with simulated data and real data were conducted. Although simulation studies may not be realistic in some situations, they can be justified in the present context since through a simulation study one estimation procedure can be compared with another.

Artificial data, representing the responses of N individuals on n items, were generated using DATGEN (Hambleton & Rovinelli, 1973) according to the one-parameter logistic model. In generating the values of θ_i and b_j $(i = 1,\ldots,N;$ $j = 1,\ldots,n)$, it was assumed that θ_i and b_j were independently and identically distributed. In addition, data were generated in such a way as to not reproduce the prior distribution. Because θ_i and b_j were assumed to be normal, these parameters were generated as samples from uniform distributions, with zero mean and unit variance.

The design of the comparison study was conceptualized in terms of the following, completely crossed, factors: estimation procedure (Bayesian, maximum likelihood), number of examinees, N, number of items, n. This design was carried out for (a) conditional estimation of $\boldsymbol{\theta}$, and (b) joint estimation of $\boldsymbol{\theta}$ and $\mathbf{b}$.

To compare the two estimation procedures, the following statistics were computed: (a) the correlation between the true values and the estimates, and (b) the mean squared error difference between the true values and the estimates. The latter statistic is particularly useful in providing an assessment of the bias in estimation for the procedures.

In the first study, the conditional Bayesian and maximum likelihood estimation procedures for estimating ability were compared. The results are indicated in Table I.

In terms of the correlations with the true values, the difference between ML and Bayes estimates is negligible for relatively large values of N and n. However, for small values of N and/or n, the Bayes estimates correlate better with true values than the ML estimates.

The most dramatic difference between the Bayesian estimates and the ML estimates is with respect to the mean squared deviations of the estimates from the true values. In general, the mean squared deviations are much smaller for

TABLE I

Conditional Estimation of θ: Comparison of the Bayesian Estimate and Maximum Likelihood Estimate

		$\Sigma(\hat{\theta} - \theta_t)^2/N$		Correlation	
Number of Examinees	Number of Items	*ML*	Bayes	*ML*	Bayes
20	15	.479	.137	.885	.912
	25	.175	.083	.951	.950
	40	.106	.056	.959	.955
	50	.138	.048	.979	.981
50	15	.440	.117	.915	.928
	25	.282	.129	.950	.944
	40	.231	.099	.976	.977
	50	.246	.091	.982	.985

Note. Priors were set as: $\nu_\theta = \lambda_\theta = 10$; $\hat{\theta}$–estimate; θ_t–true value.

the Bayesian estimates than for the ML estimates. The difference is particularly noticeable with small *N* and *n*. In some cases, the mean squared deviations for the ML estimates is almost four times as large as that for the Bayesian estimates. This result is rather surprising since the Bayesian estimates can be expected to be regressed toward the mean or "biased." The only explanation for this finding is that the ML procedure is severely biased for small *n* and *N*, even more so than the Bayesian procedure.

In the second study the Bayesian and maximum likelihood procedures for obtaining joint estimates of the item and ability parameters were compared. Because the joint estimation of item and ability parameters is usually of more interest than the conditional estimation of one set of parameters, the effect of prior distribution on the estimates was also investigated. Two examinee samples, $N = 50$, 75, and three test lengths, $n = 15$, 25, 50 were chosen for study. To examine the effect of priors, the scale parameters, λ_θ and λ_b, were set at 10, while four values for the degrees of freedom, ν_θ and ν_b were studied. The results of the simulation study are reported in Table II.

The results obtained for the conditional estimation of ability appear to hold for the joint estimation of item and ability parameters. The correlations between the estimates and the true values are identical for the maximum likelihood and the Bayes estimates of the difficulty parameters. In the estimation of ability, however, the Bayes estimates show slightly higher correlations with the true values. As with the conditional estimation of ability, the two procedures differed in terms of the mean squared deviations. The Bayes procedure is clearly superior, particularly in the estimation of ability. As can be expected, the Bayes procedure shows the greatest improvement with small examinee sample size and short tests.

TABLE II

Comparison of Bayes and Maximum Likelihood Estimators of Parameters in the Rasch Model

		Number of Items					
		15		25		50	
		Number of Examinees					
Statistic[a]		50	75	50	75	50	75
Difficulty							
Correlation *ML*		.952	.983	.970	.978	.974	.980
Correlation Bayes[b]		.952	.983	.970	.978	.975	.980
$\Sigma(\hat{b} - b_t)^2/n$	*ML*	.0867	.0978	.0606	.0491	.0592	.0390
	$\nu = 5$	.0707	.0406	.0437	.0416	.0475	.0350
	$\nu = 8$	.0725	.0427	.0446	.0428	.0479	.0355
	$\nu = 15$	.0775	.0490	.0479	.0463	.0495	.0367
	$\nu = 25$	.0867	.0609	.0543	.0523	.0528	.0389
Ability							
Correlation *ML*		.924	.943	.949	.953	.978	.971
Correlation Bayes		.928	.946	.956	.957	.980	.972
$\Sigma(\hat{\theta} - \theta_t)^2/N$	*ML*	.1686	.1836	.1637	.1279	.0571	.0777
	$\nu = 5$	.1534	.1413	.0802	.0953	.0384	.0601
	$\nu = 8$	.1593	.1477	.0797	.0967	.0387	.0602
	$\nu = 15$	.1748	.1652	.0800	.1007	.0398	.0605
	$\nu = 25$	.2006	.1947	.0838	.1082	.0424	.0616

[a] $\hat{b}$–Estimate of difficulty; b_t–True value; $\hat{\theta}$–Estimate of ability; θ_t–True value.

[b] The correlation between the estimates and the true values for given item-examinee combination were unaffected by the priors. Hence, only one correlation coefficient is reported for each set of priors.

TABLE III

Comparisons of Maximum Likelihood and Bayes Estimates for NAEP Data

	Difficulty						Ability				
Item No.	*ML*	Bayes $\lambda = 10$				Raw Score	*ML*	Bayes $\lambda = 10$			
		$\nu = 5$	$\nu = 8$	$\nu = 15$	$\nu = 25$			$\nu = 5$	$\nu = 8$	$\nu = 15$	$\nu = 25$
11	-2.15	-1.96	-1.95	-1.93	-1.90	2	-2.71	-1.89	-1.88	-1.85	-1.79
12	-2.09	-1.90	-1.89	-1.88	-1.85	4	-2.16	-1.61	-1.60	-1.57	-1.53
3	-1.97	-1.80	-1.79	-1.77	-1.75	5	-1.95	-1.48	-1.47	-1.44	-1.40
14	-1.91	-1.75	-1.74	-1.73	-1.70	6	-1.76	-1.35	-1.34	-1.32	-1.28
13	-1.91	-1.75	-1.74	-1.73	-1.70	8	-1.42	-1.10	-1.09	-1.07	-1.05
22	-1.72	-1.58	-1.58	-1.56	-1.54	9	-1.25	-.98	-.97	-.96	-.93
2	-1.67	-1.54	-1.54	-1.53	-1.51	10	-1.09	-.86	-.85	-.84	-.82
15	-1.59	-1.47	-1.47	-1.46	-1.44	11	-.93	-.74	-.73	-.72	-.70
16	-1.38	-1.29	-1.29	-1.28	-1.26	12	-.77	-.62	-.61	-.60	-.59
1	-1.21	-1.13	-1.13	-1.12	-1.11	13	-.61	-.50	-.50	-.49	-.47
10	-1.15	-1.09	-1.08	-1.08	-1.06	14	-.45	-.38	-.38	-.37	-.36
17	-.54	-.54	-.53	-.53	-.53	15	-.30	-.26	-.26	-.25	-.25
5	-.50	-.50	-.50	-.50	-.50	16	-.14	-.14	-.14	-.13	-.13
19	-.46	-.47	-.47	-.47	-.47	17	.02	-.01	-.01	-.01	-.01
8	-.46	-.47	-.47	-.47	-.47	18	.19	.12	.12	.11	.11
20	-.21	-.25	-.25	-.24	-.24	19	.37	.25	.25	.24	.24
21	-.06	-.12	-.12	-.12	-.12	20	.56	.39	.38	.38	.37
25	.02	-.04	-.04	-.05	-.05	21	.77	.53	.53	.52	.51
4	.10	.03	.03	.03	.02	22	1.02	.69	.69	.67	.66
6	.30	.20	.20	.20	.19	23	1.34	.87	.86	.84	.82
23	.40	.29	.29	.28	.28	24	1.80	1.06	1.05	1.03	1.00
7	.43	.32	.32	.31	.31	25	—	1.29	1.28	1.24	1.20
24	.59	.45	.45	.45	.44						
9	.68	.53	.53	.52	.51						
8	1.21	.96	.96	.95	.93						
M	-.69	-.68	-.67	-.67	-.66	*M*	.00	.00	.00	.00	.00
SD	1.02	.90	.89	.88	.87	*SD*	.78	.61	.60	.59	.57

Note. Based on 25 items and 200 examinees.

A comparison of the Bayes estimates with respect to varying ν_θ and ν_b indicates that prior beliefs do not seem to affect the correlation between the true values and the estimates. The mean squared deviations indicate that the accuracy of estimation is affected to some degree, especially in small samples, by the prior beliefs. This trend is more evident in the estimation of ability than in the estimation of the item parameters. This finding can be explained by noting that, in general, the number of ability parameters is large in comparison with the number of items (with the exception of the situation where $n = 50$ and $N = 50$). The priors therefore have more of an effect when estimating ability parameters. As ν increases, the prior distribution becomes concentrated and results in the estimating being regressed toward the mean. This regression biases the estimates, and this is reflected in higher mean squared deviations. (The mean squared deviations for $n = N = 50$ bear this out). In general, the regression appears to be minimal for values of ν between 5 and 15. The specification of the values $\nu = 25$ and $\lambda = 10$ is unreasonable because this implies that the mean of the inverse chi-square distribution is .43 with standard deviation .13. With priors that reflect a more diffused distribution of ϕ, the Bayes estimates are reasonably robust.

An analysis of real data confirms the above findings. Test data obtained from an administration of the National Assessment of Educational Progress Mathematics Booklet One (for 13-year-olds in 1977–78) were analyzed using the Bayes and maximum likelihood procedures. The results are reported in Table III. For the Bayes estimation procedure, λ was set at 10, and the analysis was repeated with $\nu = 5, 8, 15, 25$. The difference among the Bayes estmates using $\nu = 5, 8, 15$, are minimal. (The specification $\nu = 25$ affects the estimates of ability more than the estimates of difficulty for reasons indicated earlier.) This coupled with the pairwise correlations among the four sets of Bayes estimates, which were 1.000, indicates that the prior specifications has little effect on the Bayes estimates.

The advantages of the Bayes procedure are clear. The most obvious is that an ability estimate corresponding to a perfect score, 25, is available. The maximum likelihood estimate clearly does not exist in this case and hence, to analyze the data, five examinees with perfect scores were eliminated by the LOGIST (Wood, Wingersky, & Lord, 1976) program. The Bayes estimates are clearly regressed, the extreme values being more regressed than those values that are around the mean.

Conclusion

The Bayesian procedure for estimating parameters in the one-parameter latent trait model is an attractive alternative to the maximum likelihood procedure. Bayesian procedures are conceptually more appealing because direct interpretations of probability statements involving the parameters are possible. Empirically, as the results of the comparison studies indicate, the Bayes procedure is superior to the maximum likelihood procedure in terms of (a) accuracy, at least when the number of items and the number of examinees are small; and (b) meaningfulness, since Bayes estimates are available for perfect item and examinee scores. In any Bayesian procedure, the question of the effect of prior distributions on the estimates arises. Studies reported in this paper with real and simulated data indicate that the Bayesian procedure described, being based on a hierarchical model, is relatively robust with respect to specification of prior information. In general, prior distributions that are neither too vague nor too specific are desirable. Values for the parameters that describe the distribution of hyperparameters, such as $\lambda = 10$ and $5 \leqslant \nu \leqslant 15$, result in robust estimation.

In summary, we note that the Bayesian procedure developed here is relatively easy to implement, computationally as simple as the maximum likelihood procedure, and has the potential for greatly improving the accuracy of the estimates. For large numbers of items and large numbers of examinees the maximum likelihood procedure and the Bayes procedure yield comparable results. Given this and that the Bayes procedure results in maximum improvement for small examinee samples and short tests, the Bayes procedure is clearly more attractive than the maximum likelihood procedure.

Acknowledgements

The research reported here was performed pursuant to Grant No. N0014–79–C–0039 with the office of Naval Research and to Grant No. FQ 7624–79–0014 with the Air Force Human Resource Laboratory. The opinions expressed here, however, do not reflect the positions or policies of these agencies.

The authors are grateful to Dr. Ronald K. Hambleton for making the NAEP data available.

References

Andersen, E. B. Asymptotic properties of conditional maximum likelihood estimates. *The Journal of the Royal Statistical Society*, Series B, 1970, *32*, 283–301.

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

Andersen, E. B. Conditional inference in multiple choice questionnaire. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 31–44. (a)

Andersen, E. B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, *28*, 123–140. (b)

Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, 1969, *6*, 250–276.

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

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

Jensema, C. J. A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, 1976, *36*, 705–715.

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

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

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

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, 1974, *39*, 247–264.

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

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

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

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

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

Samejima, F. A general model for free-response data. *Psychometric Monograph*, 1972, No. 18.

Urry, V. W. Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, 1974, *34*, 253–269.

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

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

Wright, B. D., & Douglas, G. A. Best procedure for sample-free item analysis. *Applied Psychological Measurement*, 1977, *1*, 281–295.

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

Zellner, A. *An introduction to Bayesian inference in econometrics*. New York: Wiley, 1971.

Authors

SWAMINATHAN, HARIHARAN. Associate Dean and Associate Professor, School of Education, University of Massachusetts, Amherst, MA 01003. *Specializations*: Psychometric theory, Multivariate statistics.

GIFFORD, JANICE A. Research Associate,School of Education, University of Massachusetts, Amherst, MA 01003. *Specializations*: Measurement, Multivariate statistics.