A Rasch Hierarchical Measurement Model
Author(s): Kimberly S. Maier
Source: *Journal of Educational and Behavioral Statistics*, Vol. 26, No. 3 (Autumn, 2001), pp. 307-330
Published by: American Educational Research Association and American Statistical Association
Stable URL: http://www.jstor.org/stable/3648160
Accessed: 12/10/2011 12:13

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Educational Research Association and *American Statistical Association* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Educational and Behavioral Statistics.*

http://www.jstor.org

Journal of Educational and Behavioral Statistics
Fall 2001, Vol. 26, No. 3, pp. 307–330

A Rasch Hierarchical Measurement Model

Kimberly S. Maier
University of Chicago

In this article, a hierarchical measurement model is developed that enables researchers to measure a latent trait variable and model the error variance corresponding to multiple levels. The Rasch hierarchical measurement model (HMM) results when a Rasch IRT model and a one-way ANOVA with random effects are combined (Bryk & Raudenbush, 1992; Goldstein, 1987; Rasch, 1960). This model is appropriate for modeling dichotomous response strings nested within a contextual level. Examples of this type of structure include responses from students nested within schools and multiple response strings nested within people. Model parameter estimates of the Rasch HMM were obtained using the Bayesian data analysis methods of Gibbs sampling and the Metropolis-Hastings algorithm (Gelfand, Hills, Racine-Poon, & Smith, 1990; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The model is illustrated with two simulated data sets and data from the Sloan Study of Youth and Social Development. The results are discussed and parameter estimates for the simulated data sets are compared to parameter estimates obtained using a two-step estimation approach.

Keywords: *Gibbs sampling, hierarchical linear models, item response theory*

Both item response theory and hierarchical linear modeling are used in a variety of social science research applications. The use of item response theory (IRT) allows connections to be made between observed categorical responses provided by students and an underlying unobservable trait, such as ability or attitude (Hambleton & Swaminathan, 1985; Lord & Novick, 1968). Hierarchical linear modeling (HLM) allows the natural multilevel structure present in so much social science data to be represented formally in data analysis (Bryk & Raudenbush, 1992; Goldstein, 1987; Longford, 1993). In some cases, a researcher may wish to study the effects of covariates on the latent trait of interest. These covariates may include information about the respondents, as well as contextual information. This article will present both a model that integrates an IRT and a hierarchical linear model and a method of estimating model parameter values that does not rely on large-sample theory and normal approximations.

The author wishes to acknowledge the Alfred P. Sloan Foundation, and the principal investigators of the Sloan Study of Youth and Social Development.

Item response theory models and hierarchical linear models can be combined to model the effect of multilevel covariates on a latent trait. We may wish to examine relationships between person-ability estimates and person-level and contextual-level characteristics that may affect these ability estimates. Alternatively, we may wish to model data obtained from the same individuals across repeated questionnaire administrations. We may even wish to study the effect of person characteristics on ability estimates over time.

In particular, the model resulting from the integration of a hierarchical linear model and a one-parameter logistic item response model will be presented in this article. This model will be referred to as a Rasch hierarchical measurement model (HMM). The particular Rasch HMM developed in this study incorporates a Rasch model (Rasch, 1960) and a two-level hierarchical linear model having a random intercept at the first level, with no additional fixed or random covariates at either level. This form of a hierarchical linear model is known as a one-way analysis of variance with random effects. The Rasch model is appropriate for modeling dichotomous responses and models the probability of an individual's correct response on a dichotomous item. The logistic item characteristic curve, a function of ability, forms the boundary between the probability areas of answering an item incorrectly and answering the item correctly. This one-parameter logistic model assumes that the discriminations of all items are assumed to be equal to one.

The model resulting from the integration of a hierarchical linear model and a Rasch model allows one to estimate all model parameters simultaneously and therefore incorporate the standard errors of the latent trait estimates into the total variance of the model. In the Rasch HMM, the expected value of the latent trait parameter is replaced with a one-way ANOVA with random effects. The Rasch HMM can allow one, for example, correctly to model the variances of person-level and school-level error while estimating latent trait parameters of student ability estimates or student attitudes from student responses to a questionnaire of dichotomous items.

Researchers have expanded traditional IRT models in a number of ways that are appropriate in a variety of applications. Person-level characteristics have been included in IRT models to help improve estimation of item difficulty parameters, or to model the effects of person characteristics upon the estimated latent trait measures (Mislevy, 1987; Patz & Junker, 1999a; Patz & Junker, 1999b). The IRT model has also been reformulated as a two-level model consisting of items nested within people in order to model measurement error among and between these two levels (Adams, Wilson, & Wu, 1997; Kamata, 1998). Cheong and Raudenbush (2000) take this last example a step further by including a third contextual level.[1]

A variety of methods have been used to estimate the parameters of these expanded IRT models. A two-step approach has sometimes been used. In this strategy, an IRT model is used to estimate latent trait parameters for each person, which are then with a hierarchical linear model. The standard errors of the latent trait esti-

mates are not modeled in the second step, resulting in biased parameter estimates. The extent of this bias can be especially large when the total sample size is small or when the hierarchical structure is sparsely populated. Others have used methods that rely on large-sample approximations or empirical Bayes approaches (Adams, Wilson, & Wu, 1997; Cheong & Raudenbush, 2000; Kamata, 1998; Mislevy, 1987; Zwinderman, 1991, 1997). The use of these particular estimation methods, because they depend on normal distribution theory, introduces constraints on the minimum allowable sample size or the degree to which the hierarchical structure can be sparsely populated. In addition, complex integrations are usually required within the context of the solution strategy.

Bayesian methods, a third approach to estimating model parameters of an expanded IRT model, do not rely on normal approximations. Bayesian methods allow an easier solution strategy that produces unbiased estimates and eliminates the need for directly computing complex integrations (Bayes, 1763; Gelman, Carlin, Stern, & Rubin, 1995). Values for the parameters of the Rasch hierarchical measurement model will be estimated using Bayesian data analysis methods.

The Bayesian paradigm assumes the model parameters are random quantities having distributions. The distributions characterizing these unknown parameters are conditional on the observed data, which are assumed to be fixed. Bayesian inference supplements the likelihood equation with prior beliefs the analyst may have about the distributions of the parameters, via prior distributions. The likelihood and prior distributions are combined according to Bayes' theorem to produce the posterior distribution of the model parameters to be estimated. In contrast, normal theory or the frequentist method postulates that the true values of the parameters are fixed and the data are random, and rely on large-sample approximations to produce estimates of model parameters. Empirical Bayesian methods make use of both paradigms. A subset of parameters is estimated and treated as fixed and known values in a subsequent Bayesian data analysis technique to estimate the remaining unknown parameters. Typically, estimates of the first subset of model parameters are obtained using frequentist methods that rely on approximations.

Markov Chain Monte Carlo (MCMC) techniques are particular Bayesian data analysis methods that are used to estimate model parameters. In contrast to frequentist methods that produce a model parameter estimate and a standard error of the estimate, MCMC techniques can be used to produce the entire posterior distribution of the model parameter estimate. Gibbs sampling, a specific MCMC technique, is a method for generating random variables from a distribution by sampling from the collection of full conditional distributions of the complete posterior distribution (Gelfand et al., 1990). In complex models such as the Rasch hierarchical measurement model, a complicated posterior distribution can be represented as a collection of conditional probability distributions having standard distributional forms. A single sampled data point is drawn from the conditional probability distribution of each parameter, conditional on the values of the col-

lection of remaining parameters and the data. The marginal probability distributions of the parameters can be constructed from the random draws after the Markov chain has converged.

Bayesian data analysis methods were used to produce parameter estimates of the Rasch hierarchical measurement model. In particular, estimates of the parameters were found using Gibbs sampling. If the parameter does not have a conditional distribution of a common distributional form (the latent trait parameters and the item difficulty parameters), the Metropolis-Hastings algorithm has been used to generate a random draw from the conditional distribution (Hastings, 1970; Metropolis et al., 1953). Patz and Junker (1999b) estimate parameters for a two-parameter logistic model using the combination of these particular Bayesian methods, and provide a detailed description of these MCMC methods within the context of IRT models.

Construction of the Posterior and Full Conditional Distributions

As a first step in Bayesian data analysis, the prior distributions for all model parameters must be specified in order to form a posterior density. For the latent trait parameter, it is sensible to assume that the latent trait of individual n, θ_n, is drawn from a normal distribution (Lord & Novick, 1968) with unknown mean and variance,

$$p_\theta(\theta_n) \sim \text{Normal}\left(\mu_\theta, \sigma_\theta^2\right). \tag{1}$$

Specific prior distributions for the hyperparameters of the latent trait distribution will be assigned later; for now, these prior distributions will be noted merely as $p(\mu_\theta)$ and $p(\sigma_\theta^2)$.

Typically, item difficulty parameters range between –4 and +4 standard deviations and can be modeled by a unimodal symmetric distribution (Baker, 1992). Consequently, a normal prior distribution will be assigned to the item difficulty parameters,

$$p_\theta(\xi_i) \sim \text{Normal}\left(\mu_\xi, \sigma_\xi^2\right). \tag{2}$$

Upon examining the Rasch model, it is clear that the model is unidentified when estimates for all latent trait and item difficutly parameters are unknown. This difficulty can be addressed by assuming that the mean of the item parameters is zero and the variance is one. This constraint can be directly incorporated into the prior distribution for the item difficulty parameters (Box & Tiao, 1973).

The use of Gibbs sampling requires that all full conditional distributions of the model parameters be determined. Consider the case where students (level-1) are nested within classrooms (level-2), and the outcome variable matrix consists of the dichotomous response strings students provide on a test with I items. Given the $N \times I$ matrix $\boldsymbol{x}$ for $N = \Sigma_{k=1} n_k$ individuals answering I items, and assuming conditional independence among the responses, the likelihood of observing the response string $\boldsymbol{x}$ for N students nested within K classrooms is

$$\ell(\theta, \xi|\boldsymbol{x}) = p(\boldsymbol{x}|\theta, \xi) = \frac{\exp\left[\sum_{k=1}^{K}\sum_{j=1}^{n_k}\sum_{i=1}^{I} x_{ijk}(\theta_{jk} - \xi_i)\right]}{\prod_{k=1}^{K}\prod_{j=1}^{n_k}\prod_{i=1}^{I} 1 + \exp(\theta_{jk} - \xi_i)}. \tag{3}$$

This is very similar to the likelihood equation for a one-parameter logistic item response model, with the addition of an extra indexing variable, *k*. The unknown parameters in the likelihood include the latent trait variables θ and the item difficulty parameters ξ. Student *n*'s latent trait parameter can be modeled with a one-way ANOVA with random effects. The latent trait parameter of group *k* is expected to have a value α_k and a variance σ_ε^2. The random intercept α_k at the student level is in turn modeled by a linear equation, and is expected to have a mean value of γ_{00} and a variance of τ_{00},

$$\theta_{jk} = \alpha_k + \varepsilon_{jk}, \tag{4}$$

$$\varepsilon_{jk} \sim \text{Normal}\left(0, \sigma_\varepsilon^2\right), \tag{5}$$

$$\alpha_k = \gamma_{00} + \delta_{0k}, \tag{6}$$

$$\delta_{0k} \sim \text{Normal}\,(0, \tau_{00}). \tag{7}$$

The posterior distribution of the Rasch HMM is the product of the likelihood equation and the prior distributions of all unknown parameters,

$$p\left(\xi, \theta, \alpha, \gamma_{00}, \sigma_\varepsilon^2, \tau_{00}, \sigma_\xi^2|\boldsymbol{x}\right) \propto$$

$$\prod_{k=1}^{K}\prod_{j=1}^{n_k} p(\xi)p\left(\theta|\alpha_k, \sigma_\varepsilon^2\right)p(\alpha_k|\gamma_{00}, \tau_{00})p\left(\sigma_\varepsilon^2\right)p(\gamma_{00})p(\tau_{00})p\left(\sigma_\xi^2\right)p(\boldsymbol{x}|\theta, \xi). \tag{8}$$

The hierarchical linear model is incorporated into the hierarchical measurement model via the prior distribution for the latent trait parameter, $p(\theta|\alpha, \sigma_\varepsilon^2)$. Based on the normal distribution specified earlier for the latent trait parameter, the prior distribution for the students' latent trait parameters is conditional upon the level-1 random intercept and the level-1 error variance of the hierarchical linear model,

$$p\left(\theta|\alpha, \sigma_\varepsilon^2\right) \propto \prod_{k=1}^{K}\prod_{j=1}^{n_k} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left(\theta_{jk} - \alpha_k\right)^2\right]. \tag{9}$$

The prior distribution for the level-1 random intercept can be constructed by assuming that the level-1 random intercept α_k is normally distributed with a mean of the level-2 fixed intercept γ_{00} and a variance of the level-2 error variance τ_{00},

$$p(\alpha_k|\gamma_{00}, \tau_{00}) \propto \frac{1}{\sqrt{2\pi\tau_{00}}} \exp\left[-\frac{1}{2\tau_{00}}(\alpha_k - \gamma_{00})^2\right]. \tag{10}$$

Combining the prior distribution for the latent trait parameter (Equation 10) and the likelihood equation (Equation 3), the full conditional posterior for the latent trait parameter of an individual student is

$$p\left(\theta_{jk}\middle|\xi, \sigma_\xi^2, \theta_{<j><k>}, \sigma_\varepsilon^2, \alpha, \tau_{00}, \gamma_{00}, \boldsymbol{x}\right) \propto$$
$$\frac{\exp\left[\sum_{i=1}^{I} \boldsymbol{x}_{ijk}(\theta_{jk} - \xi_i)\right]}{\prod_{i=1}^{I}[1 + \exp(\theta_{jk} - \xi_i)]} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(\theta_{jk} - \alpha_k)^2\right]. \tag{11}$$

The full conditional posterior distribution for an individual student is conditional on the remaining students' latent trait parameters, as indicated by the notation $\theta_{<j><k>}$ in Equation 11. Since Equation 11 is not the kernel of any standard probability distribution, this posterior conditional distribution cannot be directly sampled from, necessitating an alternate strategy to generate random draws.

The full conditional probability distributions of the level-1 random intercepts and the level-2 fixed intercept can be expressed as products of the likelihood equation and normal prior distributions. Using Equation 10 as the prior distribution for the level-1 random intercept α_k, the full conditional probability distribution for this parameter is

$$p\left(\alpha_k|\theta, \xi, \sigma_\varepsilon^2, \alpha_{<k>}, \sigma_\varepsilon^2, \gamma_{00}, \tau_{00}, \boldsymbol{x}\right) \propto$$
$$\exp\left[-\frac{1}{2\tau_{00}}(\alpha_k - \gamma_{00})^2\right] \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\sum_{j=1}^{n_k}(\alpha_k - \theta_{jk})^2\right]. \tag{12}$$

Since this is a case of normal data, with a normal prior distribution, the full conditional probability distribution can be reformulated as a normal distribution from which sampling is easy (Box & Tiao, 1973; Seltzer & Ang, 1999),

$$\alpha_k \sim \text{Normal}\left(\tilde{\alpha}_k, \tilde{V}_k\right), \tag{13}$$

$$\tilde{\alpha}_k = \lambda_k \bar{\theta}_k + (1 - \lambda_k)\gamma_{00}, \tag{14}$$

where

$$\lambda_k = \frac{n_k/\sigma_\varepsilon^2}{\left[n_k/\sigma_\varepsilon^2 + 1/\tau_{00}\right]}, \tag{15}$$

$$\tilde{V}_k = \frac{1}{\left[n_k / \sigma_\varepsilon^2 + 1 / \tau_{00} \right]}. \tag{16}$$

In the case of a one-way ANOVA with random effects, the expected value of the level-1 random intercept is the level-2 fixed intercept γ_{00}. A uniform prior distribution for the level-2 fixed intercept will be assumed. The full conditional probability distribution of this parameter as a function of the conditional distributions of the level-1 random intercepts is

$$p(\gamma_{00}|\theta, \xi, \alpha, \sigma_\varepsilon^2, \tau_{00}, \boldsymbol{x}) \propto p(\alpha|\gamma_{00}, \tau_{00}) p(\gamma_{00}) \tag{17}$$

$$\propto \exp\left[-\frac{1}{2\tau_{00}} \sum_{k=1}^{K} (\gamma_{00} - \alpha_k)^2 \right]. \tag{18}$$

As a consequence of the manipulation of Equation 18, the full conditional probability distribution for the level-2 fixed intercept γ_{00} is a normal distribution,

$$\gamma_{00} \sim \text{Normal}\left(\bar{\alpha}, \tau_{00} / K \right), \tag{19}$$

with a mean of the average of all level-1 random intercepts $\bar{\alpha}$, taken across all K level-1 groups.

Several prior distributions for the level-1 error variance σ_ε^2 and the level-2 error variance τ_{00} will be considered here. In particular, both informative and noninformative prior distributions will be assumed. Using the uniform distribution as a prior distribution provides the least amount of prior information possible. The use of this prior distribution suggests that any value for the estimate of the parameter is equally likely and yields a full conditional probability that is dependent only upon the likelihood equation. The inverse of the full conditional probability distributions for the level-1 error variance σ_ε^2 and the level-2 error variance τ_{00}, assuming uniform prior distributions, are kernels of gamma probability distributions,

$$p(\sigma_\varepsilon^2|\theta, \xi, \sigma_\xi^2, \alpha, \gamma_{00}, \tau_{00}, \boldsymbol{x}) \propto \left(\frac{1}{\sigma_\varepsilon^2} \right)^{\frac{N}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\theta_{jk} - \alpha_k)^2 \right], \tag{20}$$

$$p(\tau_{00}|\theta, \xi, \sigma_\varepsilon^2, \alpha, \sigma_\varepsilon^2, \gamma_{00}, \boldsymbol{x}) \propto \left(\frac{1}{\tau_{00}} \right)^{\frac{K}{2}} \exp\left[-\frac{1}{2\tau_{00}} \sum_{k=1}^{K} (\alpha_k - \gamma_{00})^2 \right]. \tag{21}$$

The conditional probability distributions for the variances can be rewritten as gamma probability distributions,

$$\frac{1}{\sigma_\varepsilon^2} \sim \text{Gamma}\left[\frac{N-2}{2}, \frac{2}{\sum_{k=1}^{K}\sum_{j=1}^{n_k}(\theta_{jk}-\alpha_k)^2}\right], \tag{22}$$

$$\frac{1}{\tau_{00}} \sim \text{Gamma}\left[\frac{K-2}{2}, \frac{2}{\sum_{k=1}^{K}(\alpha_k-\gamma_{00})^2}\right]. \tag{23}$$

When we wish to incorporate prior knowledge we may have about a particular parameter, the scaled inverse chi-square distribution is an informative prior distribution for the level-1 or level-2 error variances. When this prior distribution is assumed for the level-1 and level-2 error variances, the inverses of these conditional probability distributions are found to be kernels of a gamma distribution,

$$\frac{1}{\sigma_\varepsilon^2} \sim \text{Gamma}\left[\frac{N+v}{2}, \frac{2}{S+\sum_{k=1}^{K}\sum_{j=1}^{n_k}(\theta_{jk}-\alpha_k)^2}\right], \tag{24}$$

$$\frac{1}{\tau_{00}} \sim \text{Gamma}\left[\frac{K+v}{2}, \frac{2}{S+\sum_{k=1}^{K}(\alpha_k-\gamma_{00})^2}\right]. \tag{25}$$

The scaled inverse chi-square distribution can be scaled to reflect the increasingly informative prior information one may have about the error variances. As the parameter v becomes larger, this distribution becomes more concentrated at the mean, $S/(v-2)$. For values of v between 1 and 4, the variance of the scaled inverse chi-square distribution $2S^2/[(v-2)^2(v-4)]$ is infinite, and this prior information then becomes weak relative to the data.

The full conditional distributions for the item difficulty parameters remain to be developed. Recall that the prior distribution for the item difficulty parameters is a standard normal distribution. Consequently, the full conditional distribution for item i is similar to, but slightly simpler in form than, that of the latent trait θ_{jk},

$$p(\xi_i|\xi_{<i>}, \theta, \sigma_\varepsilon^2, \alpha, \tau_{00}, \gamma_{00}, \boldsymbol{x}) \propto \frac{\exp\left[\sum_{k=1}^{K}\sum_{j=1}^{n_k} \boldsymbol{x}_{ijk}(\theta_{jk}-\xi_i)\right]}{\prod_{k=1}^{K}\prod_{j=1}^{n_k}[1+\exp(\theta_{jk}-\xi_i)]}\exp(-\xi_i^2). \tag{26}$$

As with the conditional distribution of the latent trait parameter, the full conditional distribution for the item difficulty parameters does not have a common distributional form.

The full set of conditional probability distributions developed previously forms the basis for the Rasch hierarchical measurement model. Aside from the latent trait and item difficulty parameters, the conditional probability distributions for the remaining model parameters are proportional to common distributions and thus easy to sample from directly. The conditional probability distributions for the latent trait and the item parameters cannot be directly sampled from, and the Metropolis-Hastings (M-H) algorithm will be employed to draw samples from these conditional probability distributions.

Examples: Balanced and Unbalanced Data Sets

Three data sets, two simulated, were analyzed using the Rasch hierarchical measurement model. The first simulated data set is balanced and represents an ideal data situation, with each level-2 group containing an equal number of level-1 units. A practical example of this data set occurs when a researcher gathers the same number of repeated measurements on a sample of students. The second data set is an unbalanced sparse data set that would occur when a small and unequal number of measurements are made on a sample of students. This data set illustrates a more realistic situation that a researcher may experience, and was used to evaluate the effectiveness of the model for challenging data situations. The third data set is a subset of data from a longitudinal study of adolescents. The spare structure of this data set was replicated to create the unbalanced simulated data set.

The structures of the three data sets are somewhat similar to one another. The balanced simulated data set consists of $N = 742$ response strings to 10 items, with a grouping structure that consists of $K = 53$ level-2 groups, each with $n_k = 14$ level-1 response string units. The unbalanced simulated data set has the sample total number of response strings, but with a different grouping structure. The level-1 response string units of this data set are sparsely dispersed within level-2 groups with three-quarters of the level-2 groups containing two to three level-1 response string units, and the remaining level-2 groups containing between four and six level-1 response string units. The Sloan data has a grouping structure identical to that of the simulated unbalanced data but consists of response strings to 7 items.

The response strings for each of the two simulated data sets were generated in the following manner. First, values for the item difficulty parameters were generated from a standard normal distribution. Next, values of the latent trait parameters were generated using values of the level-2 intercept and level-1 and level-2 error variances based on results from descriptive and IRT analyses of the data set constructed for the Maier (2000) study. The actual values used for the data simulation were 0.2835 for the level-1 error variance, 0.7099 for the level-2 error variance, and –0.0001 for the level-2 fixed intercept. Finally, the probability that a level-1 unit would answer an item correctly was calculated using the Rasch IRT model and the generated latent trait and item difficulty parameter values. To prevent the model from fitting the data perfectly, overdispersion was built into the simulation of the response strings: A unit's response for a particular test item was assigned a value of one if the calculated probability of a correct response exceeded a randomly generated uniform number.

The third data set is a subset of data from the Sloan Study of Youth and Social Development and is similar to the data set used by Maier (2000). This data consists of responses of 313 adolescents collected while they were engaged in a mathematics classroom. The response set consists of seven questionnaire items that loaded on a common factor that was identified as mood (Hektner, 1996). The adolescents were asked to indicate their levels (high or low) of feeling happy, strong, active, sociable, proud, involved, and excited. Table 1 reveals that adolescents were more likely to record high levels of feeling happy, strong, involved, proud or sociable, and low levels of feeling active or excited.

TABLE 1
Frequencies for Mood Components

	Frequency	
Mood Component	0	1
Happy	277	465
Strong	354	388
Active	415	327
Sociable	290	452
Proud	340	402
Involved	287	455
Excited	460	282

Implementation of Gibbs Sampling and the Metropolis-Hastings Algorithm

All three data sets were analyzed and the posterior distributions of the model parameters were produced using Gibbs sampling. The M-H algorithm was used to draw samples from the conditional distributions of the latent trait and item difficulty parameters. For all the data sets, two analyses were completed to produce a total of four complete analyses: One analysis assumed uniform prior distributions for the level-1 and level-2 variances while the other assumed scaled inverse chi-square prior distributions for the error variances. In particular, a scaled inverse chi-square prior distribution having $v = 10$ degrees of freedom and a mean $S = 2.268$ was assumed for the level-1 variance σ_ε^2. The scaled inverse chi-square prior distribution assumed for the level-2 error variance τ_{00} had the same degrees of freedom, but a mean of $S = 5.689$. Swaminathan and Gifford (1982) suggested choosing $5 \leq v \leq 15$ when utilizing this prior distribution with a Rasch IRT model.

The candidate-generating density $q(x, y)$ of the M-H algorithm used to simulate the latent trait and item difficulty parameters was chosen to be a normal distribution having a mean of the current state of the chain x and a standard deviation c_n. The form of this candidate-generating density produces the random-walk M-H algorithm. Since the candidate-generating density is symmetric $[(q(z) = q(-z))]$, the probability of the chain moving from the current value x to the proposed value y reduces to

$$\alpha(x, y) = \min\left[\frac{\pi(y)}{\pi(x)}, 1\right].$$

The standard deviation c_n of the candidate-generating density was fixed to achieve an acceptance proportion of roughly 0.5 (Gelman, Roberts, & Gilks, 1996; Patz & Junker, 1999b). For the balanced simulated data set, in the case of uniform priors for the level-1 and level-2 error variances, the acceptance proportion was 0.5377 for the item difficulty parameters specifying a standard deviation $c_n = 0.15$ and 0.4500 for the latent trait parameters specifying $c_n = 1.0$. Assuming scaled inverse chi-square priors for the level-1 and level-2 error variances and using the same values of c_n, the proportion of acceptance was 0.5388 for the item difficulty parameters and 0.4441 for the latent trait parameters. For the unbalanced simulated data set, assuming uniform priors for the error variances, the acceptance proportion was 0.5418 for the item difficulty parameters and 0.4871 for the latent trait parameters. The acceptance proportions for the same data set, assuming scaled inverse chi-square priors for the error variances, were 0.5410 for the item difficulty parameters and 0.4785 for the latent trait parameters. In the case of the Sloan data set and assuming uniform priors, the acceptance proportions were 0.5735 and 0.6347 for the item difficulty and latent trait parameters, respectively; assuming scaled inverse chi-square priors, these respective proportions were 0.5743 and 0.6274.

The values of the Markov chains of each model parameter were used to generate the corresponding marginal distribution for each of the parameter estimates. The starting values used for the analyses of the data sets appear in Table 2. The initial value used for the latent trait parameter of each level-1 unit was simply the raw score averaged across test items. For all analyses, 30,000 iterations of the algorithm were run. The first 1,000 iterations were considered to be the burn-in iterations and these corresponding deviates were discarded. The resulting 29,000 iterations formed the basis for parameter estimation.

Results of Analysis

The results of the analyses are shown in Tables 3–7 for the simulated balanced and Sloan data sets. For the simulated data sets, the true values of each of the model parameters are listed in the second columns of the tables. These values can be compared to the mean of the deviates over 29,000 iterations. The variance of the posterior distribution is also calculated. The time-series standard error of the estimate

TABLE 2
Starting Model Parameter Values

Model Parameter	Starting Value
Level-1 Variance, σ^2_ε	0.35
Level-1 Intercept, α_k	0.50
Level-2 Variance, τ_{00}	1.00
Level-2 Intercept, γ_{00}	0.50
Latent Trait, θ_{jk}	Average raw score
Item Parameters, ξ_i	0.00

TABLE 3
Item Difficulty Parameter Estimates Under Uniform and Scaled Inverse Chi-Square Prior Distributions for Level-1 and Level-2 Error Variances, Simulated Balanced Data Set

Model Parameter	True Value	*Mean*	Time-series *SE* of *Mean*	Variance	95% Credibility Interval
ξ_0	−1.95903				
Uniform		−1.89	0.00142	0.00912	(−2.090, −1.710)
Inverse χ^2		−1.89	0.00136	0.00922	(−2.080, −1.700)
ξ_1	0.75865				
Uniform		0.791	0.00101	0.00674	(0.632, 0.953)
Inverse χ^2		0.788	0.00106	0.00658	(0.627, 0.945)
ξ_2	1.22899				
Uniform		1.32	0.00117	0.00766	(1.150, 1.490)
Inverse χ^2		1.32	0.00114	0.00803	(1.140, 1.490)
ξ_3	0.46361				
Uniform		0.394	0.00094	0.00618	(0.241, 0.549)
Inverse χ^2		0.392	0.00096	0.00612	(0.238, 0.547)
ξ_4	−0.61123				
Uniform		−0.548	0.00100	0.00605	(−0.703, −0.395)
Inverse χ^2		−0.548	0.00096	0.00594	(−0.700, −0.396)
ξ_5	−1.07601				
Uniform		−1.05	0.00102	0.00679	(−1.210, −0.887)
Inverse χ^2		−1.04	0.00106	0.00666	(−1.210, −0.885)
ξ_6	−0.09302				
Uniform		−0.177	0.00091	0.00593	(−0.326, −0.025)
Inverse χ^2		−0.177	0.00090	0.00594	(−0.329, −0.026)
ξ_7	−0.26429				
Uniform		−0.198	0.00088	0.00573	(−0.346, −0.050)
Inverse χ^2		−0.196	0.00092	0.00585	(−0.345, −0.044)
ξ_8	0.62427				
Uniform		0.595	0.00093	0.00612	(0.441, 0.749)
Inverse χ^2		0.591	0.00096	0.00616	(0.437, 0.744)
ξ_9	0.92816				
Uniform		0.766	0.00102	0.00654	(0.609, 0.925)
Inverse χ^2		0.766	0.00102	0.00645	(0.611, 0.924)

Note. 95% credibility interval for the deviates in parentheses.

of the mean can be used as an estimate of the Monte Carlo error. The final columns of the tables specify the 95% credibility interval for the deviates.

Examining the results for the item difficulty parameter estimates of the simulated data sets first reveals that the agreement between the mean of the posterior distribution of the estimate and the true value for the parameter is quite good. For both data sets, the true value lies within the 95% credibility interval for all but one of the item difficulty parameters. The true value of Item 9 lies just outside the 95% credibility interval of the estimate, but within the 97.5% credibility interval. The standard error of the estimate of the mean of the item difficulty parameter estimates

TABLE 4
Posterior Distribution of Item Difficulty Parameters Under Uniform and Scaled Inverse Chi-Square Prior Distributions for Level-1 and Level-2 Error Variances, Simulated Unbalanced Data Set

Model Parameter	True Value	*Mean*	Time-series *SE* of *Mean*	Variance	95% Credibility Interval
ξ_0	−1.95903				
Uniform		−1.96	0.00155	0.0104	(−2.160, −1.760)
Inverse χ^2		−1.95	0.00143	0.0104	(−2.150, −1.750)
ξ_1	0.75865				
Uniform		0.771	0.00104	0.00638	(0.614, 0.929)
Inverse χ^2		0.769	0.00103	0.00637	(0.613, 0.925)
ξ_2	1.22899				
Uniform		1.25	0.00109	0.00733	(1.080, 1.420)
Inverse χ^2		1.24	0.00110	0.00694	(1.080, 1.410)
ξ_3	0.46361				
Uniform		0.470	0.00095	0.00613	(0.315, 0.622)
Inverse χ^2		0.465	0.00092	0.00601	(0.312, 0.617)
ξ_4	−0.61123				
Uniform		−0.725	0.00099	0.00653	(−0.887, −0.566)
Inverse χ^2		−0.721	0.00096	0.00643	(−0.879, −0.564)
ξ_5	−1.07601				
Uniform		−1.04	0.00108	0.00714	(−1.210, −0.879)
Inverse χ^2		−1.04	0.00115	0.00726	(−1.210, −0.869)
ξ_6	−0.09302				
Uniform		−0.108	0.00095	0.00604	(−0.263, 0.045)
Inverse χ^2		−0.109	0.00096	0.00601	(−0.261, 0.042)
ξ_7	−0.26429				
Uniform		−0.262	0.00098	0.00607	(−0.414, −0.108)
Inverse χ^2		−0.260	0.00097	0.00613	(−0.414, −0.105)
ξ_8	0.62427				
Uniform		0.620	0.00094	0.00615	(0.467, 0.774)
Inverse χ^2		0.619	0.00091	0.00594	(0.466, 0.770)
ξ_9	0.92816				
Uniform		0.985	0.00102	0.00681	(0.821, 1.150)
Inverse χ^2		0.980	0.00105	0.00664	(0.822, 1.140)

Note. 95% credibility interval for the deviates in parentheses.

range from a high value of 0.00142 to a low value of 0.00088, estimated by dividing the square root of the spectral density estimate by the sample size. These statistics were calculated using CODA software (Best, Cowles, & Vines, 1995).

The results of the Sloan data appear in Table 7. The estimates for the item difficulty parameters show that that the sampled adolescents indicated mood levels slightly above average. The variability of mood level was greater between people than within person, with the variability between people almost twice the variability within persons. Upon examination of the item difficulty parameters, the results indicate that it is much easier for this sample of adolescents to agree that they have

TABLE 5
Hierarchical Parameter Estimates Under Uniform and Scaled Inverse Chi-Square Prior Distributions for Level-1 and Level-2 Error Variances, Simulated Balanced Data Set

Model Parameter	True Value	*Mean*	Time-series *SE* of *Mean*	Variance	95% Credibility Interval
γ_{00}	–0.000125				
Uniform		–0.0893	0.00087	0.01232	(–0.306, 0.129)
Inverse χ^2		–0.0888	0.00089	0.01210	(–0.303, 0.126)
σ^2_ε	0.283486				
Uniform		0.282	0.00144	0.00225	(0.196, 0.380)
Inverse χ^2		0.266	0.00132	0.00187	(0.188, 0.355)
τ_{00}	0.709858				
Uniform		0.600	0.00146	0.0196	(0.380, 0.920)
Inverse χ^2		0.572	0.00116	0.0137	(0.383, 0.840)

Note. 95% credibility interval for the deviates in parentheses.

high levels of happiness and feeling sociable; it is much harder for them to agree that they are feeling high levels of excitement and feeling active.

The particular number of burn-in iterations was chosen based on examination of autocorrelation values and time series plots. Examination of these plots and statistics showed that almost all of the Markov chains exhibited common behavior that was indicative of a rapidly mixing Markov chain. The notable exception is the level-1 error variance σ^2_ε, which showed a somewhat different mixing pattern and will be addressed below. Aside from this particular parameter, the time series plots of the remaining parameter estimates show similar mixing patterns. Figure 1 shows time-series plots of the Markov chain for Item 2, the simulated unbalanced data set, assuming uniform priors for the level-1 and level-2 error variances, and Figure 2 shows the corresponding plot for the simulated balanced data set assuming scaled inverse chi-square priors. Figures 3–6 show the corresponding plots for the level-2 and the level-

TABLE 6
Posterior Distribution of Hierarchical Parameters Under Uniform and Scaled Inverse Chi-Square Prior Distributions for Level-1 and Level-2 Error Variances, Simulated Unbalanced Data Set

Model Parameter	True Value	*Mean*	Time-series *SE* of *Mean*	Variance	95% Credibility Interval
γ_{00}	–0.000125				
Uniform		–0.0722	0.00076	0.00325	(–0.040, 0.183)
Inverse χ^2		–0.0722	0.00078	0.00312	(–0.038, 0.182)
σ^2_ε	0.283486				
Uniform		0.409	0.00204	0.00493	(0.282, 0.556)
Inverse χ^2		0.373	0.00187	0.00412	(0.255, 0.508)
τ_{00}	0.709858				
Uniform		0.580	0.00180	0.0078	(0.420, 0.769)
Inverse χ^2		0.576	0.00157	0.0064	(0.431, 0.745)

Note. 95% credibility interval for the deviates in parentheses.

TABLE 7
Model Parameter Estimates Under Uniform and Scaled Inverse Chi-Square Prior Distributions for Level-1 and Level-2 Error Variances, Sloan Data Set

Model Parameter	*Mean*	Time-series *SE* of *Mean*	Variance	95% Credibility Interval
ξ_0, Happy				
Uniform	–0.627	0.00115	0.00824	(–0.808, –0.450)
Inverse X^2	–0.621	0.00117	0.00803	(–0.797, –0.448)
ξ_1, Strong				
Uniform	0.075	0.00112	0.00772	(–0.096, 0.248)
Inverse X^2	0.074	0.00113	0.00766	(–0.098, 0.246)
ξ_2, Active				
Uniform	0.617	0.00115	0.00771	(0.447, 0.790)
Inverse X^2	0.610	0.00116	0.00787	(0.435, 0.782)
ξ_3, Sociable				
Uniform	–0.506	0.00123	0.00823	(–0.686, –0.330)
Inverse X^2	–0.500	0.00116	0.00815	(–0.677, –0.322)
ξ_4, Proud				
Uniform	–0.053	0.00121	0.00774	(–0.227, 0.118)
Inverse X^2	–0.049	0.00115	0.00790	(–0.223, 0.127)
ξ_5, Involved				
Uniform	–0.534	0.00120	0.00789	(–0.712, –0.361)
Inverse X^2	–0.528	0.00119	0.00803	(–0.705, –0.354)
ξ_6, Excited				
Uniform	1.030	0.00128	0.00830	(0.850, 1.210)
Inverse X^2	1.010	0.00129	0.00839	(0.836, 1.190)
γ_{00}, Uniform	0.234	0.00124	0.01323	(0.011, 0.461)
Inverse X^2	0.229	0.00121	0.01232	(0.012, 0.449)
σ^2_ε, Uniform	1.470	0.00541	0.04080	(1.110, 1.900)
Inverse X^2	1.330	0.00505	0.03460	(0.995, 1.730)
τ_{00}, Uniform	2.970	0.00881	0.16080	(2.250, 3.820)
Inverse X^2	2.730	0.00779	0.12888	(2.080, 3.490)

Note. 95% credibility interval for the deviates in parentheses.

1 error variances. Figures 7 and 8 show the trace plot for the level-1 and level-2 error variances of the Sloan data set, assuming inverse chi-square priors. The first four figures provide good examples of the type of rapid mixing that occurred with the Markov chains of most of the remaining parameter estimates. However, the time-series plots for the level-1 error variance for all the data sets show a lower rate of mixing, perhaps indicating that the Markov chain may not have converged.

The autocorrelation values of the Markov chains for most of the parameters rapidly approach zero as the lag increases. Table 8 shows autocorrelation values corresponding to lags of 1, 5, 10, and 50 for the Markov chains of all parameter esti-

(text continues on page 326)

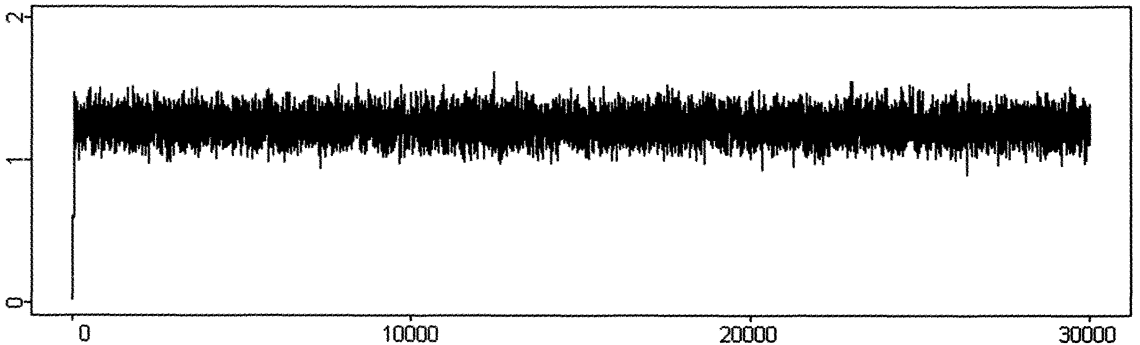

FIGURE 1. *Time-series plot of item 2, simulated unbalanced data set, uniform priors for level-1 and level-2 error variances.*

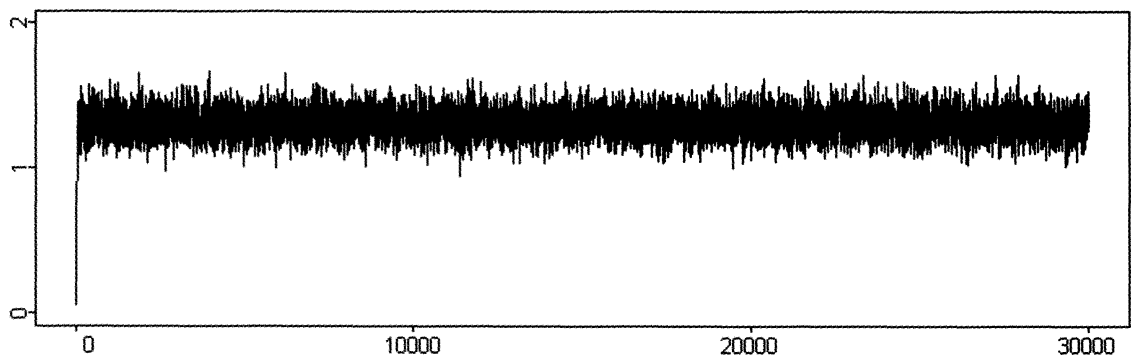

FIGURE 2. *Time-series plot of item 2, simulated balanced data set, scaled inverse chi-square priors for level-1 and level-2 error variances.*

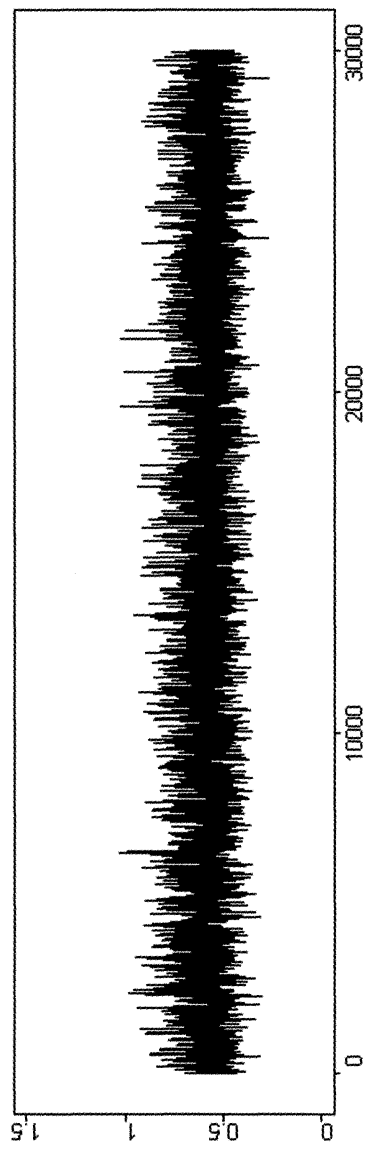

FIGURE 3. *Time-series plot of τ_{00}, simulated unbalanced data set, uniform priors for level-1 and level-2 error variances.*

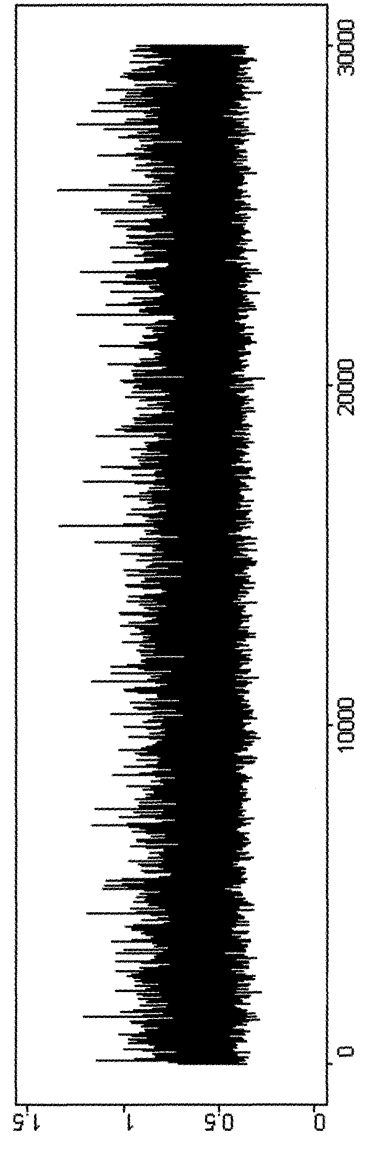

FIGURE 4. *Time-series plot of τ_{00}, simulated balanced data set, scaled inverse chi-square priors for level-1 and level-2 error variances.*

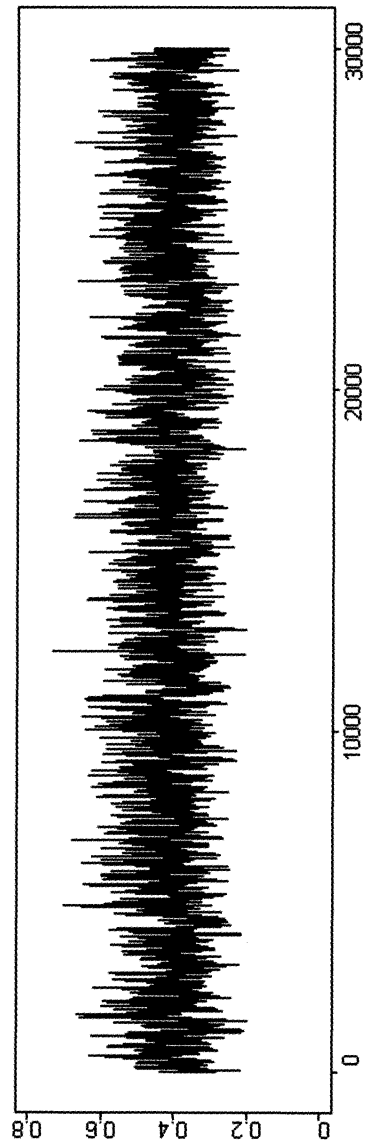

FIGURE 5. *Time-series plot of* σ_ε^2*, simulated unbalanced data set, uniform priors for level-1 and level-2 error variances.*

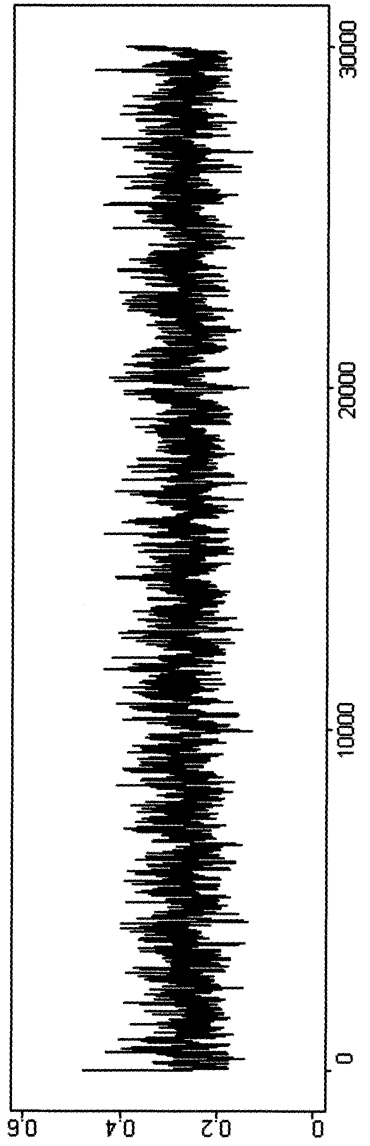

FIGURE 6. *Time-series plot of* σ_ε^2*, simulated balanced data set, scaled inverse chi-square priors for level-1 and level-2 error variances.*

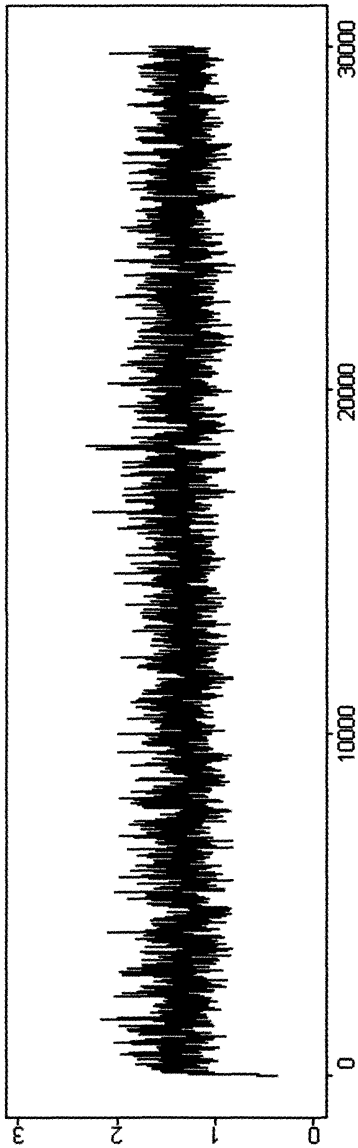

FIGURE 7. *Time-series plot of σ_ε^2, Sloan data set, scaled inverse chi-square priors for level-1 and level-2 error variances.*

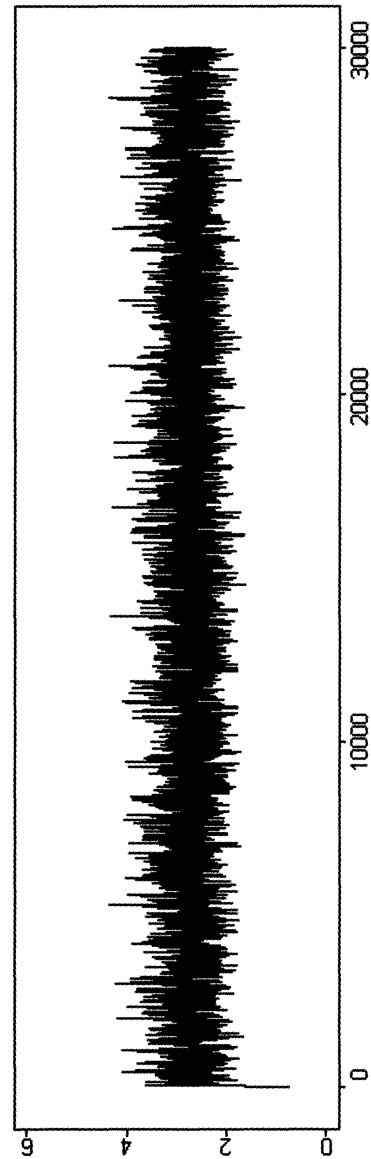

FIGURE 8. *Time-series plot of τ_{00}, Sloan data set, scaled inverse chi-square priors for level-1 and level-2 error variances.*

mates of the simulated balanced data set. The values of autocorrelation for the simulated unbalanced data set and the Sloan data set show the same pattern. As indicated in Table 8, the autocorrelation values rapidly approach zero for almost all of the parameter estimates, a property that indicates rapid Markov chain mixing. As with the time-series plots, the notable exception to this behavior is the level-1 error variance σ_ε^2.

Exploration of Level-1 Error Variance

Additional analyses were completed further to examine convergence and mixing rates of the Markov chains for the level-1 error variance. The Gelman and Rubin (1992) convergence diagnostic was calculated for corresponding Markov chains. Since this diagnostic requires multiple Markov chains, three separate sets of Markov chains were run for each of the two simulated data sets, assuming three different starting values for the level-1 error variance. These starting values were

TABLE 8
Autocorrelation Values as a Function of Lag, Simulated Balanced Data Set

Model Parameter	Level-1 & Level-2 Prior	Autocorrelation			
		Lag 1	Lag 5	Lag 10	Lag 50
ξ_0	Uniform	0.69300	0.19200	0.06220	0.00811
	Inverse χ^2	0.68900	0.19200	0.05730	–0.01010
ξ_1	Uniform	0.66400	0.13900	0.02050	0.01200
	Inverse χ^2	0.65900	0.13900	0.03940	0.00181
ξ_2	Uniform	0.67000	0.15200	0.02890	0.00624
	Inverse χ^2	0.67400	0.15100	0.03710	0.00417
ξ_3	Uniform	0.65300	0.11100	0.01990	0.00976
	Inverse χ^2	0.65200	0.13700	0.01700	–0.00305
ξ_4	Uniform	0.65200	0.12900	0.02600	–0.00388
	Inverse χ^2	0.65200	0.12800	0.01580	0.00237
ξ_5	Uniform	0.66500	0.15900	0.02660	–0.01520
	Inverse χ^2	0.66700	0.14700	0.02940	–0.00038
ξ_6	Uniform	0.65500	0.13600	0.01790	0.00751
	Inverse χ^2	0.64800	0.12600	0.01110	–0.00109
ξ_7	Uniform	0.63700	0.11400	0.01240	–0.00089
	Inverse χ^2	0.64400	0.12500	0.03240	0.00175
ξ_8	Uniform	0.65200	0.13700	0.01900	–0.00673
	Inverse χ^2	0.64600	0.13900	0.02920	–0.00078
ξ_9	Uniform	0.66500	0.14200	0.02180	0.00538
	Inverse χ^2	0.65900	0.14400	0.04990	0.00386
γ_{00}	Uniform	0.09960	0.03880	0.02070	0.00588
	Inverse χ^2	0.09790	0.04480	0.03570	–0.00230
σ_ε^2	Uniform	0.90100	0.76500	0.62600	0.12000
	Inverse χ^2	0.89600	0.75800	0.61400	0.15000
τ_{00}	Uniform	0.18600	0.08570	0.05030	0.01060
	Inverse χ^2	0.16600	0.08800	0.06140	0.01250

1.0, 3.0, and 5.0, while the starting values for the other parameters were retained from the first analysis. The Gelman and Rubin diagnostic was calculated for the three chains of the level-1 error variance using CODA software. All three Markov chains for the balanced and unbalanced data sets met the Gelman and Rubin criteria for convergence, suggesting that the Markov chains of the level-1 error variance converged to a stationary distribution.

Autocorrelation values of this additional set of Markov chains were examined to assess the rate of mixing. These values were comparable to that of the original Markov chains of all the model parameters. These findings suggest that applying a thinning interval to the Markov chain for the level-1 error variance may be an appropriate strategy to improve the mixing rate. Autocorrelation values for Markov chains with different thinning intervals were examined and a thinning interval of three was identified as the best option because it considerably reduced the autocorrelation without substantially increasing the Monte Carlo variance.

Over all, the additional set of Markov chains for the level-1 error variance behaved similarly to the chains originally simulated. As with the original Markov chains, the 95% credibility intervals (averaged from the three Markov chains) contain the true value of the level-1 error variance. In most cases, the posterior distributions are not centered on the true value of the parameter. For both data sets, the mean of the posterior distribution for the level-1 error variance in the unbalanced data set slightly overestimates the true value, while the posterior mean for the level-2 error variance slightly overestimates the true value of the parameter.

Both the original Markov chain and the additional set of Markov chains demonstrate similar behavior for the error variance estimates. Clearly this behavior was not a statistical artifact present only in the original Markov chains. It was decided to investigate whether this behavior was related to the true values of the error variance parameters, especially in the case of the level-1 variance, which is fairly close to zero. New balanced and unbalanced Rasch HMM data sets were simulated using a value of one for the level-1 variance σ_ε^2. Model parameters were estimated using the same MCMC algorithm as used for the original data sets. The first 1,000 iterations were discarded as the burn-in, and the Markov chains mixed adequately, as indicated by examination of the time-series plots. Again, the 95% credibility intervals contain the true values of the parameters; again, the posterior distributions are not centered on the true value of the parameters.

Since the results for the new data sets are similar to those of the original data sets, this pattern does not seem to be related to the true value of the level-1 variance. However, the pattern could conceivably be linked to the process used to simulate the data sets. As mentioned previously, overdispersion was built into the model by comparing the probability of a correct response to a randomly generated uniform deviate. This procedure may very well account for the discrepancies between posterior means and true values of the level-1 and level-2 variances. Also, although the Markov chains of the level-1 error variance seem to exhibit a lower rate of mixing, the chains meet the criterion of a variety of convergence diagnostics indicating stationarity had been reached.

Comparison to a Two-Step Approach

To illustrate how the Rasch HMM performs relative to a traditional two-step approach, the simulated balanced data set was reanalyzed. First, estimates of the latent trait parameter for each of the $N = 742$ response strings were produced according to a Rasch item response model, using the BIGSTEPS program (Wright & Linacre, 1993). The true values of the item difficulty parameters were given for this step, to make equating unnecessary. The resulting latent trait parameter estimates were then used as the outcome variable for a two-level HLM that used the same hierarchical structure as the simulated balanced data set. The hierarchical coefficients were estimated using the HLM program (Bryk, Raudenbush, & Congdon, 1996). The results of this analysis appear in Table 9. For this particular data set, the two-step analysis approach grossly overestimates the level-1 random error variance and underestimates the level-2 random error variance while correctly estimating the level-2 fixed intercept. Clearly, in this case the Rasch HMM models the data much better than the two-step strategy.

TABLE 9
Estimates of Hierarchical Parameters from Two-Step Analysis, Simulated Balanced Data Set

Model Parameter	Coefficient	*SE*	*T* ratio
γ_{00}	−0.109593	0.113543	−0.965
	$\sigma^2_\varepsilon = 0.951460$ ($SE = 0.78255$)		
	$\tau_{00} = 0.612390$ ($SE = 0.97543$)		

Implementation and Future Research

To obtain estimates of the hierarchical measurement model parameters, the Gibbs sampling algorithms were implemented in a computer program written in Visual C++. The object-oriented capabilities of C++ make this language a natural fit for the nested multi-parameter structure of the hierarchical measurement model. To produce estimates for the Rasch HMM, 18 and 24 minutes were required to run 30,000 iterations of the Gibbs sampling algorithm on a CPU with a 450 MHz processor and 192 MB of memory. CODA software (Best et al., 1995) was used as a post-Gibbs analysis tool. This software was used to calculate estimates of the mean, standard error of the mean, and the variance of the posterior distributions of the model parameters, as well as to generate time-series and autocorrelation values.

The Rasch HMM is very specialized because it appropriates for dichotomous responses only and does not allow incorporation of any level-1 or level-2 covariates. The usefulness of HMMs hinges on the degree to which these models can be generalized. Generalization can occur along at least three avenues. Different IRT models can be incorporated into the model. Work is currently being done that integrates a Partial Credit IRT model with a 2-level HLM, resulting in a Partial Credit HMM. Another way to expand the hierarchical measurement model is to consider

an alternative distribution for the level-1 random intercept or the level-2 error variances. A HMM is currently being investigated that uses a t-distribution for the latent trait parameters. This model would allow outlier level-1 groups to be modeled appropriately. Additionally, more complex item response models and hierarchical linear models will also be considered.

Note

[1]While this article was in press, Fox and Glas (2001) appeared. This article presents a HMM that combined a different HLM with a two-parameter normal ogive IRT model.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Dekker.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions to the Royal Society,* 330–418. Reprinted with biographical note by G. A. Barnard (1958), in *Biometrika, 45,* 293–315.

Best, N., Cowles, M. K., & Vines, K. (1995). *CODA: Convergence diagnosis and output analysis software for Gibbs sampling output Version 0.30* [computer program]. Cambridge: MRC Biostatistics Unit, Institute of Public Health.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* New York: John Wiley & Sons.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A., Raudenbush, S. W., & Congdon, R. (1996). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods, 5*(4), 477–495.

Fox, J. P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 269–286.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85*(412), 972–985.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting* (pp. 599–608). New York: Oxford.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Sciences, 7*(4), 457–472.

Goldstein, H. (1987). *Multilevel models in educational and social research.* New York: Oxford.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Hektner, J. (1996). Exploring optimal personality development: A longitudinal study of adolescents. Unpublished doctoral dissertation, University of Chicago.

Kamata, A. (1998). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Longford, N. T. (1993). *Random coefficient models.* New York: Oxford.

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

Maier, K. S. (2000). *Applying Bayesian methods to hierarchical measurement models.* Unpublished doctoral dissertation, University of Chicago.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics, 21,* 1087–1092.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*(1), 81–91.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational & Behavioral Statistics, 24*(4), 342–366.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago.

Seltzer, M. H., & Ang, A. (1999). *Estimation and inference in small-sample research settings: An introduction to Bayesian analysis via the Gibbs sampler.* Unpublished manuscript.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*(3), 175–191.

Wright, B. D., & Linacre, J. M. (1993). *BIGSTEPS* [computer program]. Chicago: University of Chicago.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56,* 589–600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245–256). New York: Springer.

Author

KIMBERLY S. MAIER is Postdoctoral Fellow, Alfred P. Sloan Center on Parents, Children, and Work, Rm. 352B, 1155 East 60th Street, Chicago, IL 60637; k-maier@uchicago.edu. Her research interests include hierarchical models, missing data methods, Bayesian data analysis, and gender differences in science and math.