DOCUMENT RESUME

ED 476 471 TM 034 943

AUTHOR	Johnson, Matthew S.; Sinharay, Sandip
TITLE	Calibration of Automatically Generated Items Using Bayesian Hierarchical Modeling.
PUB DATE	2003-05-12
NOTE	32p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Bayesian Statistics; *Constructed Response; *Educational Assessment; Estimation (Mathematics); Markov Processes; Monte Carlo Methods; Multiple Choice Tests; *Test Items
IDENTIFIERS	*Calibration; Family Clusters

ABSTRACT

For complex educational assessments, there is an increasing use of "item families," which are groups of related items. However, calibration or scoring for such an assessment requires fitting models that take into account the dependence structure inherent among the items that belong to the same item family. C. Glas and W. van der Linden (2001) suggest a Bayesian hierarchical model to analyze data involving item families with multiple choice items. This paper extends the model to take into account item families with constructed response items, and designs a Markov chain Monte Carlo algorithm for the Bayesian estimation of model parameters. The hierarchical model, which accounts for the dependence structure inherent among the items, implicitly defines the "family response function" (FRF) for the score categories. This paper suggests a way to combine the FRFs over the score categories to obtain a "family score function" (FSF), which is a quick graphical summary of the expected score of an individual with a certain ability to an item randomly generated from an item family. The paper also suggests a method for the Bayesian estimation of the FRF and FSF. This work is a significant step towards building a tool to analyze data involving item families, and it may be very useful practically, for example, in automatic item generation systems that create tests involving item families. (Contains 1 table, 10 figures, and 25 references.) (Author/SLD)

Educational Testing Service

Calibration of Automatically Generated Items Using Bayesian Hierarchical Modeling

Matthew S. Johnson
Department of Statistics
Baurch College
55 Lexington Avenue
New York, NY 10010

Sandip Sinharay
Center for Statistical Theory and Practice
Educational Testing Service
Rosedale Road; MS 12-T
Princeton, NJ 08541

May 12, 2003

Note: Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service

Abstract

For complex educational assessments, there is an increasing use of *item families*, which are groups of related items. However, calibration or scoring for such an assessment requires fitting models that take into account the dependence structure inherent among the items that belong to the same item family. Glas and van der Linden (2001) suggest a Bayesian hierarchical model to analyze data involving item families with multiple choice items. This paper extends the model to take into account item families with constructed response items, and designs a Markov chain Monte Carlo (MCMC) algorithm for the Bayesian estimation of the model parameters. The hierarchical model, which accounts for the dependence structure inherent among the items, implicitly defines the *family response function* (FRF) for the score categories. This paper suggests a way to combine the FRFs over the score categories to obtain a *family score function* (FSF), which is a quick graphical summary of the expected score of an individual with a certain ability to an item randomly generated from an item family. This paper also suggests a method for the Bayesian estimation of the FRF and FSF. This work is a significant step towards building a tool to analyze data involving item families and may be very useful practically, for example, in automatic item generation systems that create tests involving item families.

Key words: Hierarchical model; Markov chain Monte Carlo; Automatic item generation;; Family response function; Family score function; Item score function

Acknowledgements

The research was supported by Educational Testing Service Research Allocation project 884.12. The authors thank Andreas Oranje, Isaac Bejar, Randy Bennett, Paul Holland, and Shelby Haberman for their helpful comments about the work.

In addition, the authors gratefully acknowledge the National Center for Education Statistics for allowing access to the NAEP MathOnline data.

1. Introduction

The operation of large-scale high-stakes testing programs demands a large pool of high-quality items from which items can be sampled. Large pools are especially important for assessment programs that offer flexible assessment times such as the Graduate Records Examination (GRE), where concerns over item exposure and potential disclosure are the greatest. While efforts to populate item pools are laborious for pools consisting entirely of multiple choice items, the same efforts for complex constructed response tasks are even more challenging. In response to the effort, expense, and occasionally inconsistent item quality associated with traditional item production there is an increasing interest in using item models to guide production of items, automatically or manually, with similar conceptual and statistical properties. Irvine and Kyllonen (2002) surveys some current areas of investigation for item modeling and generation. Items produced from a single item model, whether by automatic item generation (AIG) systems or by rigorous manual procedures, are related to one another through the common generating model, and therefore constitute a family of related items.

Naturally, it is necessary and beneficial to use calibration models that account for the dependence structure among the items from the same item family. The works by Janssen, Tuerlinckx, Meulders, and De Boeck (2000) and Wright (2002) are initial attempts at building such models. Glas and van der Linden (2001) suggest one such model for multiple choice items that is more general. The model assumes the item parameters of a three-parameter logistic model (3PL; Lord, 1980) to be normally distributed—the mean vector and the variance matrix of the normal distribution depend on the item model from which the item is generated.

Glas and van der Linden's model has some similarity to the testlet model of Bradlow, Wainer, and Wang (1999). Both models describe an extra level of dependence in the observed assessment data. However the testlet model describes the extra "local" dependence between a single examinee's item responses within a testlet, whereas the item family model explains the dependence between all examinees' responses to the same single member from an item family.

Glas and van der Linden's model has the limitation that it cannot take into account families with constructed response items. This paper generalizes Glas and van der Linden's model to take into account item families with constructed response items. Further, this work designs a method to estimate the joint posterior distribution of the model parameters using the Markov chain Monte Carlo (MCMC; Gilks, Richardson, & Spiegelhalter, 1996) algorithm.

The hierarchical model implies a *family response function* (FRF) for each category for an item family. The FRF for category k for an item family gives the probability that an individual with a given ability will score k on an item randomly generated from the item family. The idea is similar to that behind the *family expected response function* defined by Sinharay, Johnson, and Williamson (2003) who deal with families with dichotomous items. This paper suggests a way to combine the FRFs over the score categories to obtain a *family score function* (FSF), which is a quick graphical summary of the expected score of an individual with a given ability to an item randomly generated from an item family. This work also suggests a way to compute estimates of the FRF and FSF and an approximate prediction interval around them using the Monte Carlo method and the output of the MCMC algorithm. The paper examines the performance of the hierarchical model using both simulated data and real data examples .

The next section provides a broad overview of the existing techniques for the analysis of data involving item family. Section 3 describes in detail one such model—the model generalizes Glas and van der Linden's hierarchical model to account for constructed response item family data as well. Section 4 discusses estimation of the model parameters and the family expected response functions using the Markov chain Monte Carlo (MCMC) algorithm. Section 5 reports the results from a simulation study. Section 6 discusses the application of the model to a data set from the National Assessment of Educational Progress (NAEP). Finally, the paper concludes with a summary of the findings and thoughts on possible future directions.

2. Models for item families

There are three approaches for modeling data involving item families. The models all build from standard item response theory (IRT) models for dichotomous and polytomous data. This paper uses the three-parameter logistic model (3PL; Birnbaum 1968) to describe the response behavior of examinees to multiple choice items, and the generalized partial credit model (GPCM; Muraki 1992) to describe the response behavior of examinees to constructed response items.

The 3PL model assumes that the probability an individual with ability θ_i correctly responds to item j is defined by the following equation

$$P_{j1}(\theta) = P(X_{ij} = 1|\theta_i, a_j, \beta_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp\{a_j(\beta_j - \theta)\}}. \quad (1)$$

The parameter c_j is called the asymptote, a_j is the item's discrimination, and β_j is the difficulty of the item. $P_{j0}(\theta) = 1 - P_{j1}(\theta)$ is the probability that an examinee with ability θ_i incorrectly responds to item j.

The GPCM assumes that the adjacent category logits are linear in the examinee's proficiency θ_i. Mathematically the GPCM probabilities for an item with K score categories $(0, 1, \ldots, K-1)$ are defined by the following:

$$\begin{aligned} P_{jk}(\theta_i) &\equiv P(X_{ij} = k | \theta_i, a_j, \beta_j, \delta_{j0}, \delta_{j1}, \ldots \delta_{j\overline{K-1}}) \\ &= \frac{\exp\{k a_j(\theta_i - \beta_j) - a_j \sum_{\ell=0}^{k} \delta_{j\ell}\}}{\sum_{k=0}^{K-1} \exp\{k a_j(\theta_i - \beta_j) - a_j \sum_{\ell=0}^{k} \delta_{j\ell}\}} \end{aligned} \quad (2)$$

for $k = 0, 1, \ldots, K-1$, where a_j is the item discrimination, β_j is the overall difficulty for the item and δ_{jk} is the k-th item-category step parameter. To ensure identifiability, set $\delta_{j0} = 0$ and $\sum_{\ell=0}^{K-1} \delta_{j\ell} = 0$, which results in the need of estimating only $(K-2)$ parameters out of $\delta_{j1}, \delta_{j2}, \ldots \delta_{j\overline{K-1}}$ to fit the GPCM. Note that the two-parameter logistic model (2PL; Birnbaum, 1968) is a special case of this model with K=2 and no δ-parameter.

Unrelated Siblings Model

The gold standard approach for modeling item response functions is to assume that each item is independent of all other items, regardless of whether they are siblings or not. We call this approach the *Unrelated Siblings Model* (USM). The USM assumes that X_{ij}, examinee i's score to item j, follows a multinomial or binomial distribution with probabilities defined by the generalized partial credit model in (2) for constructed response items and probabilities defined by the three-parameter logistic model in (1) for multiple choice items. The USM makes no other assumption about the item parameters.

Although standard IRT software can fit the USM, the model has the disadvantage that each item has to be individually calibrated. In addition, this approach ignores the relationship between siblings in an item family and hence will provide standard errors of item parameters that are too large, and will require larger sample sizes for acceptable calibration precision.

Identical Siblings Model

Hombo and Dresher (2001) study the results of a model that assumes the same item response function for all items in the same item family. We call this approach the *Identical Siblings*

Model (ISM). This model assumes the item parameters of the 3PL and GPCM remain constant for all items in the family (e.g. $\beta_j = \beta_{\mathcal{I}(j)}$). While this model can also be fit by standard software like PARSCALE (Muraki and Bock, 1997) or BILOG (Mislevy and Bock, 1982), it has the limitation that it ignores any variation between siblings and hence, in the face of such variations, provides incorrect estimates of the item parameters, and examinee scores.

Related Siblings Model

One way to overcome the limitations of the above-mentioned two methods is to apply the *Related Siblings Model* (RSM), a hierarchical model that assumes a separate item response function for each item, but relates siblings within a family using a hierarchical component (Glas and van der Linden 2001). The method uses a mixing distribution to describe the relationship between items within the same item family, in much the same way that the mixing distribution on the student parameter θ in an IRT model is used to describe the dependence between item responses from the same examinee.

One important point to note here is that the ISM and USM are limiting cases of the RSM. If the mixing distribution approaches to a point mass (or the variances of the mixing distribution go to 0), then the RSM approaches in the limit to the ISM. On the other hand, if the mixing distribution approaches to the Lebesgue measure (or the variances of the mixing distribution go to ∞), then the RSM approaches in the limit to the USM (Sinharay, Johnson, & Williamson, 2003).

While the advantage of the RSM is that it properly accounts for the variability among the items for the same item model, it has the disadvantage that there is no standard software for fitting this model. We use our own C++ program to fit the RSM in this work.

Janssen, Tuerlinckx, Meulders, and De Boeck (2000) and Wright (2002) provide examples of such models. Glas and van der Linden (2001) suggest one such model for multiple choice items that is more general; the model starts from a three-parameter logistic model (3PL; Lord, 1980) and uses a normal mixing distribution to relate the item parameters belonging to the same family. The mean vector and the variance matrix of the normal mixing distribution depend on the item family from which the item is generated.

All the above-mentioned models have the limitation that they cannot take into account families with constructed response items. The next section discusses an example of an RSM that encompasses constructed response items as well. The model is an extension of the model by Glas and van der Linden (2001), is a Bayesian hierarchical model, and uses a normal mixing distribution to relate siblings.

3. An RSM for Constructed Response Item Families

Suppose there are J items denoted by $j = 1, 2, \ldots J$ in a test, and that the j-th item is scored on a scale from 0 to $(K_j - 1)$. Consider that the test is given to N examinees. Let $\mathcal{I}(j)$ be the item family of which item j is a member. Items j and k are *siblings* if they are members of the same item family, i.e., if $\mathcal{I}(j) = \mathcal{I}(k)$.

We model multiple choice (MC) items using the 3PL model defined in (1) and constructed response (CR) items using the GPCM defined in (2). To be able to use a normal mixing distribution on the item parameters, we apply the transformations $\alpha_j = \log\{a_j\}$ and $\gamma_j = \log\left\{\frac{c_j}{1-c_j}\right\}$. Assuming normality of α_j and γ_j, both of which range from $-\infty$ to ∞ (whereas a_j ranges from 0 to ∞ and c_j ranges from 0 to 1) is quite reasonable. Recall that fitting a 3PL model requires estimating a_j, β_j, and c_j for each item while fitting a GPCM requires estimating a_j, β_j, and any $(K_j - 2)$ out of $\delta_{j1}, \delta_{j2}, \ldots \delta_{j\overline{K_j-1}}$. Let

$$\boldsymbol{\eta}_j = \begin{cases} (\alpha_j, \beta_j, \gamma_j)^t & \text{if item } j \text{ is an MC item} \\ (\alpha_j, \beta_j, \delta_{j1}, \delta_{j2}, \ldots \delta_{j\overline{K_j-2}})' & \text{if item } j \text{ is a CR item} \end{cases} \tag{3}$$

be the item parameter vector for item j. Then the hierarchical model defining the likelihood of the related siblings model is

$$\left.\begin{array}{rcl} X_{ij} & \sim & \text{Multinomial}(1; P_{j0}(\theta_i), \ldots, P_{j\overline{K_j-1}}(\theta_i)) \\ \theta_i & \sim & \mathcal{N}(\mu, \sigma^2) \\ \boldsymbol{\eta}_j & \sim & \mathcal{N}(\boldsymbol{\lambda}_{\mathcal{I}(j)}, \boldsymbol{T}_{\mathcal{I}(j)}) \end{array}\right\} \tag{4}$$

where the probabilities $P_{j0}, \ldots, P_{jK_j-1}$ are defined in (1) or (2) depending on whether the item is an MC or CR item,

$$\boldsymbol{\lambda}_{\mathcal{I}(j)} = \begin{cases} (\lambda_{a\mathcal{I}(j)}, \lambda_{b\mathcal{I}(j)}, \lambda_{g\mathcal{I}(j)})' & \text{if item } j \text{ is an MC item} \\ (\lambda_{a\mathcal{I}(j)}, \lambda_{b\mathcal{I}(j)}, \lambda_{d_1\mathcal{I}(j)}, \lambda_{d_2\mathcal{I}(j)}, \ldots \lambda_{d_{\overline{K_j-2}}\mathcal{I}(j)})' & \text{if item } j \text{ is a CR item} \end{cases} \tag{5}$$

is the mean item parameter vector for family $\mathcal{I}(j)$, and $T_{\mathcal{I}(j)}$ is the within-family item parameter covariance matrix for family $\mathcal{I}(j)$. We will call the λ_as as the family discrimination parameters, and the λ_bs as the family difficulty parameter. To fix the origin and scale (ensure identifiability), let $\mu = 0$ and $\sigma^2 = 1$ (an alternative method would be to force sum-to-zero constraints on the first and second components of the item family mean parameter vectors $\boldsymbol{\lambda}_{\mathcal{I}(j)}$). Note that expectations are not invariant under transformations ($E[a_j] \neq E[e^{\alpha_j}]$). However, because the transformation is monotone increasing, the quantiles (including the median) are invariant under the transformation.

A fully Bayesian formulation of the model requires the specification of prior distributions for the model parameters $\boldsymbol{\lambda}_{\mathcal{I}(j)}$ and $\boldsymbol{T}_{\mathcal{I}(j)}$. We employ the use of conjugate prior distributions for these parameters as in Glas and van der Linden (2001). Assume independent multivariate normal prior distributions for the family mean item parameter vectors $\boldsymbol{\lambda}_{\mathcal{I}(j)}$,

$$\boldsymbol{\lambda}_{\mathcal{I}(j)} \sim \mathcal{N}\left(\mu_\lambda, V_\lambda\right) \tag{6}$$

and independent inverse-Wishart prior distributions on the $\boldsymbol{T}_{\mathcal{I}(j)}$'s,

$$\boldsymbol{T}^{-1}_{\mathcal{I}(j)} \sim \text{Wishart}(W_1, W_2). \tag{7}$$

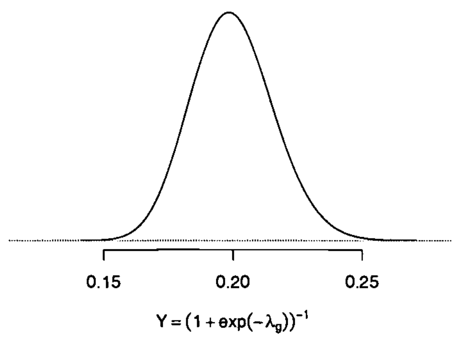
The notation $M \sim \text{Wishart}(W_1, W_2)$ implies that the density function of the $p \times p$ matrix M is proportional to

$$|M|^{(W_1-p-1)/2} \exp\left(-\frac{1}{2}\text{tr}\left(W_2^{-1}M\right)\right).$$

The prior in (7) implies that the prior mean of $\boldsymbol{T}^{-1}_{\mathcal{I}(j)}$ is W_1W_2, and that *a priori* there is information that is equivalent to W_1 observations of the item parameter vector $\boldsymbol{\eta}_j$. In most cases we suggest using a diagonal matrix for V_λ, the prior covariance matrix of the mean item parameter vectors; in the absence of prior information, the diagonal elements of the matrix should be large, e.g., $V_\lambda = 100\boldsymbol{I}_K$, where $\boldsymbol{I}_K$ is the $K \times K$ identity matrix.

One situation where it is sensible to use an informative prior is when the item family is a multiple-choice family. In that situation, a good choice for a prior distribution would be one that places its mass around the point $\mu_\lambda = (\mu_a, \mu_b, \text{logit}\{\frac{1}{\text{No. of choices}}\})^t$. For example in the case where the item family has five choices we suggest using the prior mean $\mu_g = \text{logit}\{0.2\} = -1.386$, and $\sigma_g = 0.1$. Figure 1 contains the density function of the transformed random variable $Y = \frac{1}{1+\exp(-\lambda_g)}$, where $\lambda_g \sim \mathcal{N}(-1.386, 0.1^2)$. Note that the density is centered around $0.2 = \frac{1}{\text{No. of choices}}$, and has almost all of its mass in the interval $(0.15, 0.25)$.

Figure 1: **Probability density function for the random variable $Y = \frac{1}{1+\exp(-\lambda_g)}$, where $\lambda_g \sim \mathcal{N}(-1.386, 0.1^2)$.**



Family Response Function and Family Score Function

Note that by integrating the individual item parameter vectors $\boldsymbol{\eta}_j$ out of the item response functions $\boldsymbol{P}_j(\theta)$, the RSM defines a new set of *family response functions* (FRF). Let $P_k(\theta \mid \mathcal{I})$ denote family $\mathcal{I}$'s response function for category k, $k = 0, 1, \ldots, K_{\mathcal{I}} - 1$. The family response function for category k is defined by

$$P_k(\theta \mid \mathcal{I}) = \int_{\boldsymbol{\eta}} P_k(\theta \mid \boldsymbol{\eta}) d\Phi(\boldsymbol{\eta} \mid \boldsymbol{\lambda}_{\mathcal{I}}, T_{\mathcal{I}}), \quad k = 0, 1, \ldots, K_{\mathcal{I}} - 1, \tag{8}$$

where $\Phi(.|.,.)$ is the cumulative density function of the multivariate normal distribution and $P_k(\theta \mid \boldsymbol{\eta})$ is the k^{th} category response function and is defined by (1) or (2) depending on whether the item is an MC item or CR item. The family response function $P_k(\theta \mid \mathcal{I})$ defines the probability that an individual with proficiency θ will score k on a randomly selected item from family $\mathcal{I}$. Sinharay, Johnson, and Williamson (2003), in the context of multiple choice items, call the posterior expected function of FRF as family expected response function (FERF).

Notice that an item family containing items with K score categories has the same number of FRF's. It is often desirable to examine a single function for each item family. Therefore, we define the *family score function* (FSF) $m(\theta|\mathcal{I})$, which describes the expected score on a randomly selected item from the $\mathcal{I}$-th item family for an examinee with proficiency θ, as

$$m(\theta|\mathcal{I}) = \sum_{\ell=0}^{K_{\mathcal{I}}-1} \ell \times P_\ell(\theta|\mathcal{I}), \tag{9}$$

where $P_\ell(\theta|\mathcal{I})$ is defined in (8). A similar quantity exists for each individual item in each of the families. The *item score function* (ISF) for item j, defined as

$$m_j(\theta) = \sum_{\ell=0}^{K_{\mathcal{I}}-1} \ell \times P_{j\ell}(\theta), \tag{10}$$

describes the expected score on item j of an individual with proficiency θ, where $P_{j\ell}(\theta)$s are defined in (1) or (2) depending on whether the item is an MC item or CR item.

Note that for a dichotomous item, the FSF becomes $P_1(\theta|\mathcal{I})$ and the ISF becomes $P_{j1}(\theta)$.

4. Bayesian Estimation for the Related Siblings Model

Maximum likelihood estimation of the related siblings model requires the calculation of the joint likelihood function of the family parameters $\boldsymbol{\lambda}_{\mathcal{I}(j)}$s and $\boldsymbol{T}_{\mathcal{I}(j)}$s given the observed data. Consistent estimation of these parameters would require marginalizing the likelihood with respect to both the examinee parameters $\boldsymbol{\theta}$ and the individual item parameters $\boldsymbol{\eta}$. The calculation of the family response function (FRF) in (8) demonstrated how the item parameters could be integrated out of the response function. Suppose $\boldsymbol{\Lambda}$ denotes the collection of all $\boldsymbol{\lambda}_{\mathcal{I}(j)}$s, and $\boldsymbol{T}$ denotes the collection of all $\boldsymbol{T}_{\mathcal{I}(j)}$s Now define the conditional likelihood of an examinee with proficiency θ given the family parameters $\boldsymbol{\lambda}_{\mathcal{I}(j)}$s and $\boldsymbol{T}_{\mathcal{I}(j)}$s by taking the product over all item families

$$L(\theta \mid \boldsymbol{X}, \boldsymbol{\Lambda}, \boldsymbol{T}) = \prod_{\mathcal{I}} P_{x_{\mathcal{I}}}(\theta \mid \mathcal{I}),$$

where $x_{\mathcal{I}}$ is the examinee's score to an item from family $\mathcal{I}$, $\boldsymbol{X}$ is the vector of the scores of the examinee to all items, and $P_{x_{\mathcal{I}}}(\theta \mid \mathcal{I})$ is define in (8).

It is not enough to simply integrate θ out of the examinee likelihood and take the product of the resulting terms to define the likelihood for the item family parameters. Doing so would require that item responses from different individuals to the different members of an item family are independent, when in fact they should be considered related to one another. However, by integrating the individual item parameters $(\alpha_j, \beta_j, \gamma_j)^t$ out of the true joint likelihood, the resulting model correctly accounts for the fact that responses of two individuals to different items from the same item family are correlated even when conditioning on the family parameter $\boldsymbol{\lambda}_{\mathcal{I}(j)}$ and $\boldsymbol{T}_{\mathcal{I}(j)}$. Maximizing the correct likelihood for the related siblings model would be an extremely difficult task requiring complex numerical integration techniques.

We prefer to perform a Bayesian estimation of the model. Bayesian estimation requires the determination of the joint posterior distribution of all model parameters given the observed data. Because the posterior distribution requires the the evaluation of an intractable integral, we employ Markov Chain Monte Carlo (MCMC; Gilks, et al., 1996) techniques, specifically the Metropolis-Hastings algorithm (MH; Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953; Hastings, 1970) within a Gibbs sampler (Geman and Geman, 1984). The algorithm generates a sample of parameter values from a Markov Chain that approximates the joint posterior distribution of the parameters of the model by drawing iteratively from the conditional posterior distribution of each model parameter.

Recall the definition of $\boldsymbol{\eta}_j$ in (3). Item parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and the ability variables $\boldsymbol{\theta}$ are drawn from their respective conditional distributions as described in Patz and Junker (1999). Conditional on the item parameters

$$\boldsymbol{\eta} = \text{collection of } \boldsymbol{\eta}_j\text{s},$$

the $\mathcal{I}$-th item family mean vector $\boldsymbol{\lambda}_{\mathcal{I}}$ and covariance matrix $\boldsymbol{T}_{\mathcal{I}}$ are independent of $\boldsymbol{\theta}$ and the observed data $\boldsymbol{X}$. The conditional distributions of the $\boldsymbol{\lambda}_{\mathcal{I}}$'s, which are independent over the families (i.e., over k), are given by

$$\boldsymbol{\lambda}_{\mathcal{I}} \mid \boldsymbol{\eta}, \boldsymbol{T}_{\mathcal{I}} \;\sim\; \mathcal{N}_3\left(\boldsymbol{V}_{\mathcal{I}}\left\{V_\lambda^{-1}\mu_\lambda + J_{\mathcal{I}}\boldsymbol{T}_{\mathcal{I}}{}^{-1}\overline{\eta}_{\mathcal{I}}\right\}, \boldsymbol{V}_{\mathcal{I}}\right), \tag{11}$$

where

$$\boldsymbol{V}_{\mathcal{I}} = \left(J_{\mathcal{I}}\boldsymbol{T}_{\mathcal{I}}{}^{-1} + V_\lambda^{-1}\right)^{-1},$$

μ_λ and V_λ are the prior mean and variances of $\boldsymbol{\lambda}_{\mathcal{I}}$'s respectively,

$$\overline{\eta}_{\mathcal{I}} = \frac{1}{J_{\mathcal{I}}} \sum_{j:\ \mathcal{I}(j)=\mathcal{I}} \boldsymbol{\eta}_j,$$

and $J_{\mathcal{I}}$ is the number of members in item family $\mathcal{I}$.

The conditional distributions of the $\boldsymbol{T}_{\mathcal{I}}$'s, which are also independent over the families (i.e., over $\mathcal{I}$), are given by

$$\boldsymbol{T}_{\mathcal{I}} \mid \boldsymbol{\eta}, \boldsymbol{\lambda} \sim \text{Inv-Wishart}\left(J_{\mathcal{I}} + W_1, \left\{\sum_{j:\ \mathcal{I}(j)=\mathcal{I}} (\boldsymbol{\eta}_j - \boldsymbol{\lambda}_{\mathcal{I}(j)})(\boldsymbol{\eta}_j - \boldsymbol{\lambda}_{\mathcal{I}(j)})^t + W_2^{-1}\right\}^{-1}\right). \tag{12}$$

Hence the addition of the hierarchical component in the model amounts to additional sampling from normal and inverse Wishart distributions, which are both straightforward. Hence, the hierarchy of the model does not pose significant difficulties in the Bayesian estimation procedure.

Estimating the family response function. We use Monte Carlo integration to estimate the FRF defined in (8) and to estimate a 95% prediction interval for the family. The following steps describe the Monte Carlo procedure used for the estimation of the FRF and 95% prediction interval for the $\mathcal{I}$-th family of items.

i) Generate a sample of size M from the joint posterior distribution of the hyper-parameters $\boldsymbol{\lambda}_{\mathcal{I}}$ and $\boldsymbol{T}_{\mathcal{I}}$. That is, draw

$$[\boldsymbol{\lambda}_{\mathcal{I}}{}^{(t)}, \boldsymbol{T}_{\mathcal{I}}{}^{(t)}] \sim F(\boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{T}_{\mathcal{I}} \mid \boldsymbol{X}), t = 1, \ldots, M\cdot$$

ii) For each of the above M values of the hyper-parameters $[\boldsymbol{\lambda}_{\mathcal{I}}^{(t)}, \boldsymbol{T}_{\mathcal{I}}^{(t)}]$, draw n values of the item parameter vector $\boldsymbol{\eta}_j$ from the conditional (prior) distribution of $\boldsymbol{\eta}_j$ given $\boldsymbol{\lambda}_{\mathcal{I}}{}^{(t)}$, and $\boldsymbol{T}_{\mathcal{I}}{}^{(t)}$,

$$\boldsymbol{\eta}_j^{((t-1)n+r)} \sim \Phi(\boldsymbol{\eta}_j | \boldsymbol{\lambda}_{\mathcal{I}}, \boldsymbol{T}_{\mathcal{I}})$$

for $r = 1, \ldots, n$ and $t = 1, \ldots, M$.

iii) Set the ability at θ.

iv) For each of the Mn draws $\boldsymbol{\eta}_j^{(t)}$ obtained in step ii), compute the probability for category ℓ, $p_\ell^{(t)} = P_\ell(\theta|\boldsymbol{\eta}_j^{(t)})$, for each of the item categories $\ell = 1, \ldots, K_{\mathcal{I}}$, where $K_{\mathcal{I}}$ is the number of score categories for item family $\mathcal{I}$. In addition calculate, the expected score function $m^{(t)} = \sum \ell \times p_\ell^{(t)}$.

v) The averages of the above probabilities and expected score functions are Monte Carlo estimates of the posterior means of the category FRF's and the FSF,

$$\begin{aligned} E\left[P_\ell(\theta|\mathcal{I}) \mid \boldsymbol{X}\right] &\approx \sum_{t=1}^{Mn} p_\ell^{(t)}, \\ E\left[m(\theta|\mathcal{I}) \mid \boldsymbol{X}\right] &\approx \sum_{t=1}^{Mn} m^{(t)}. \end{aligned}$$

Sinharay, Johnson, and Williamson (2003) call the above estimated posterior means of FRFs as *estimated FERFs* while dealing with item families consisting of dichotomous items.

vi) The 2.5^{th} and 97.5^{th} percentiles of the Mn probabilities $p_\ell^{(t)}$ and expected score functions $m^{(t)}$ form an approximate 95% prediction interval to attach with the estimates obtained in step (iv). This prediction interval reflects the within family variance as well as the uncertainty in the FRFs.

Steps iii) to vi) are repeated for a number of values of θ to obtain the estimated FRF for each category ℓ in item family $\mathcal{I}$. We use $M = 1000$, $n = 10$, and 100 equidistant values of θ in the interval (-4,4) to estimate the function. Changing M and n does not result in any noticable difference.

Note that the MCMC algorithm described earlier in this section provides a sample from an approximation of the posterior distribution. To obtain a sample of size M from the posterior distribution of $\boldsymbol{\lambda}_{\mathcal{I}}$ and $\boldsymbol{T}_{\mathcal{I}}$ in step i) above draw a sub-sample of size M from the output of the MCMC algorithm. Step ii) is simple because it only requires sampling from a multivariate normal distribution using the draws of $\boldsymbol{\lambda}_{\mathcal{I}}$ and $\boldsymbol{T}_{\mathcal{I}}$ from step (i). So the estimation of the FRF is quite straightforward given the output from the MCMC and takes little additional time.

5. Example 1: A Simulation Study

In this section, a simulation study examines the performance of the RSM on data generated according to a family structure.

Generating Data with Family Structure

The data generation process is very simular to that in Sinharay, Johnson, and Williamson (2003). We generate 16 item families: five families of two-category constructed response items (2PL), five families of multiple choice items (3PL), and six families of polytomous constructed response items (GPCM); two with three category items, two with four category items, and two with five category items. Each item family contains ten items (siblings). We generate the values of the individual proficiency parameters θ_is for $N = 5000$ examinees from a normal distribution with mean 0 and variance 1; each examinee receives one of ten "forms" of the test. Each form contains 16 items; one item from each of the 16 item families. The first items in each family make up the first form; the second item from each family make up the second form, et cetera. Exactly randomly selected examinees respond to each form. Because the items within one family are independent of the items in all other families, this design is not biased in any way.

This study generates item families in such a way that the resulting item parameters reflect the range of values typically observed in real-life assessments. For example, from our experience, the discrimination parameters a_js usually fall between 0.5 and 1.5 and so the data generator draws item families so that the item parameters will remain in that range. The data generator draws individual item parameters so that the within-family variance is one-fifth as large as the between-family variance. Remembering that an RSM with very small within-family variance is essentially an ISM and an RSM with large within-family variance is essentially an USM (Sinharay, Johnson, & Williamson, 2003), the ratio of within-family variance and the between-family variance used here makes data distinguishable from data generated by an USM or ISM.

Analysis of the Simulated Data

To analyze the data we use a Wishart prior distribution for the family precision matrix $\boldsymbol{T}_{\mathcal{I}}^{-1}$ with parameters $W_1 = k_{\mathcal{I}} + 1$ and $W_2 = \frac{W_1}{10} I_{k_{\mathcal{I}}}$, where $k_{\mathcal{I}}$ is the number of item parameters in the model for an item from family $\mathcal{I}$ (e.g. $k_{\mathcal{I}} = 3$ for a three category item) and $I_{k_{\mathcal{I}}}$ is the $k_{\mathcal{I}} \times k_{\mathcal{I}}$ identity matrix. The prior mean of $\boldsymbol{T}_{\mathcal{I}}^{-1}$ implied by the above choice of W_1 and W_2 is $\frac{(k_{\mathcal{I}}+1)^2}{10} I_{k_{\mathcal{I}}}$.

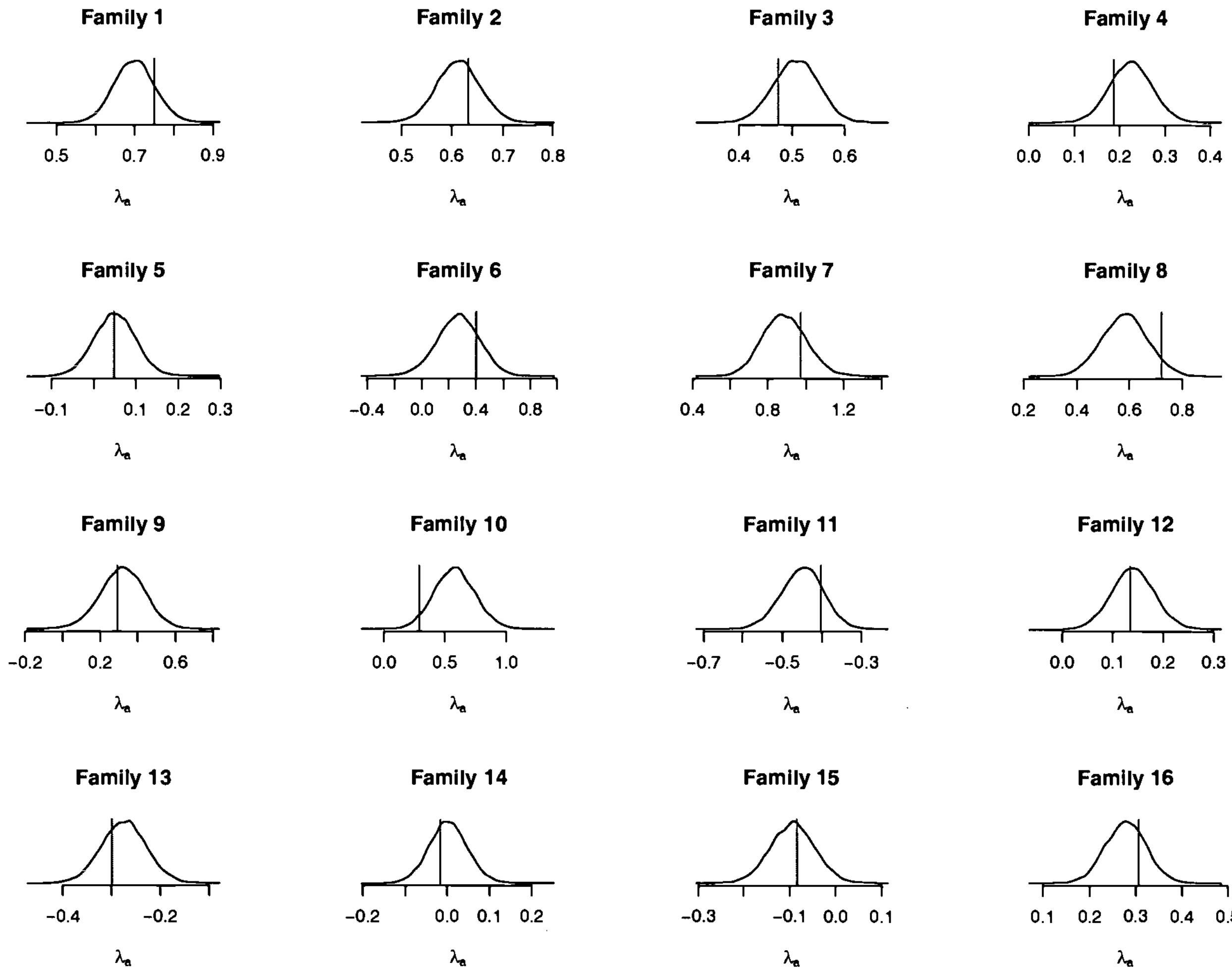
We apply the MCMC algorithm to approximate the posterior distributions of the model parameters. For this example, a number of convergence diagnostics (time-series plots, Gelman-Rubin convergence diagnostics, and Brooks-Gelman multivariace potential scale reduction factor) indicate that a chain with 50,000 iterations is sufficient to ensure convergence. We discard the first 10,000 iterations as burn-in and use every 10th draw from the Markov chain, leaving us with 4,000 draws from the approximated posterior distribution of the model parameters.

Figures 2 and 3 plot the approximated marginal posterior densities for the family parameters λ_a and λ_b respectively. Figure 4 contains the approximated posterior density for the family guessing parameter for the five 3PL item families. The vertical line in each panel of the figures represents the average value of the simulating item parameters in that family.

Figure 3 indicates that the MCMC estimation algorithm does an excellent job of recovering the simulating values of λ_bs, the family difficulty parameters. In each of the 16 item families the simulating value is contained within the 95% credible interval for that parameter.

Fourteen of the 95% credible intervals for the family discrimination parameters λ_as also contain the simulating parameter values. Families 8 and 10 are the only two families whose credible intervals do not contain the true value. The discrimination parameter of the eighth family is underestimated, and λ_a for the tenth family is overestimated. Further inspection reveals that these two families are multiple choice item families. And as is the case with simple item response models, multiple choice item families do not behave as well as constructed response item families.

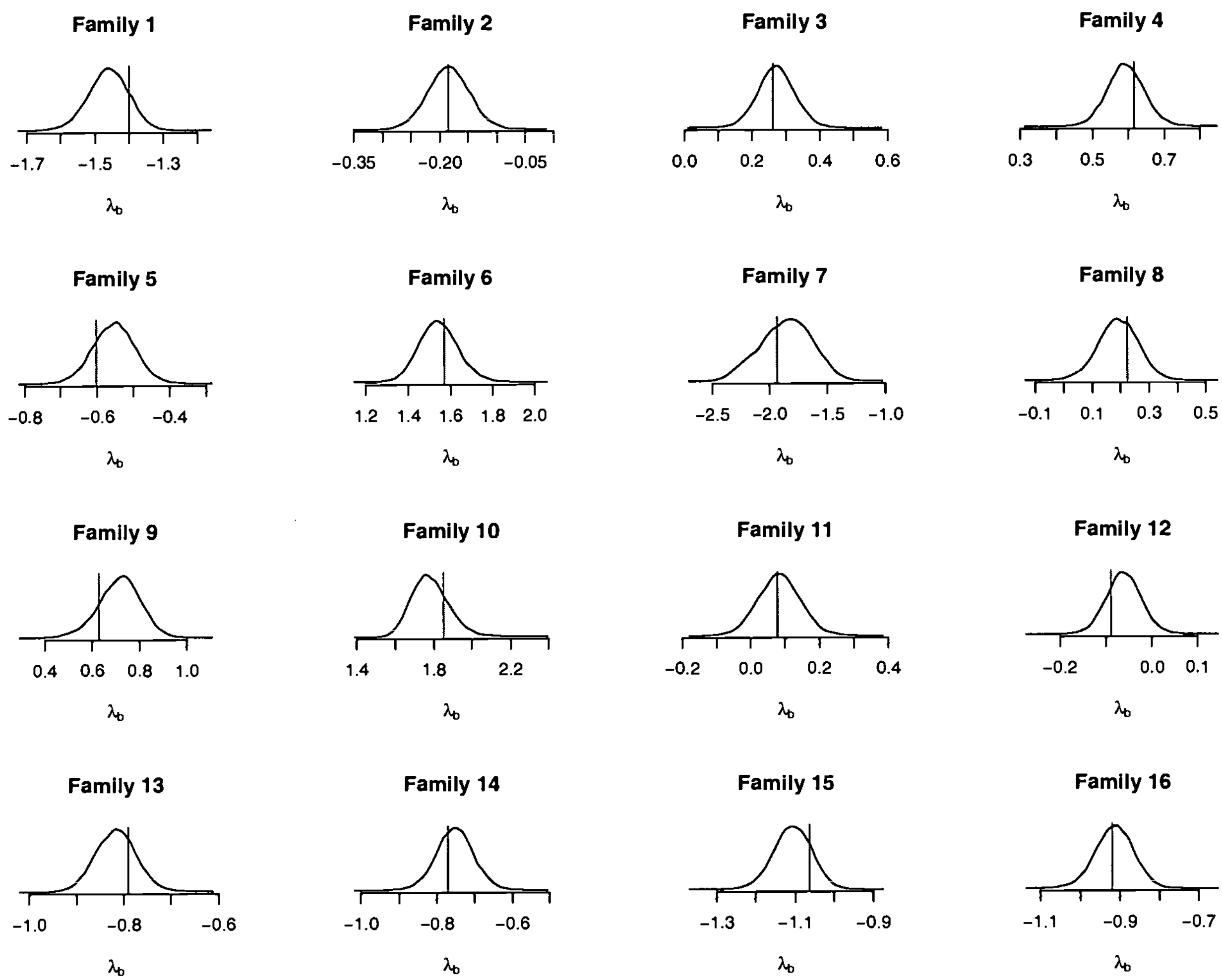
Figure 2: **Estimated posterior density functions of the family mean discriminations λ_as for the simulation study.**



Notice in Figure 4 that the family guessing parameter for the eighth family is underestimated and the guessing parameter for the tenth family is overestimated. Because of the near indeterminacy (especially for difficult items) of the 3PL model parameters, the overall effect of the under or overestimation of the family parameters for Families 8 and 10 is minimal. This is evident in Figure 5, which contains the family score functions (FSFs) for the 16 simulated families, along with the simulating item score functions (ISF) and the 95% prediction intervals for the families.

The FSFs and 95% prediction intervals for families eight and ten track the simulating ISFs quite well; only small portions of a single ISF extend beyond the 95% intervals in each of

Figure 3: **Estimated posterior density functions of the family mean difficulties λ_bs for the simulation study.**

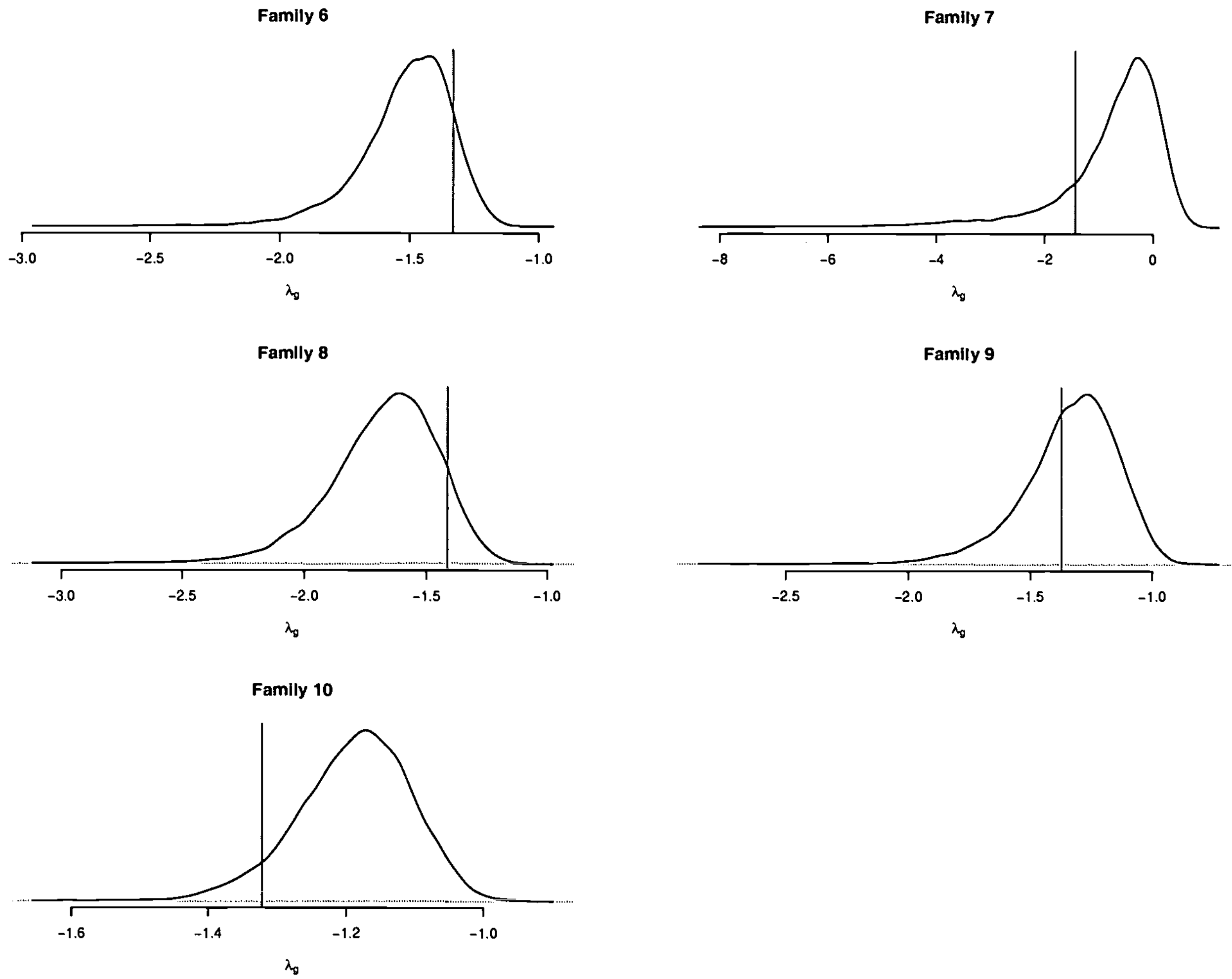


these two families. The two families that cause the greatest concern are the seventh and fourteenth item families.

The FSF for Family 7 clearly has an asymptote that is different from the simulating ISFs. Once again this can be explained by the near indeterminacy between the difficulty and guessing parameters in the 3PL when the item (or family) is an easy one and noting that Family 7 consists of very easy items. The model is unable to distinguish between an item that is easy and one that is easy to guess.

The model performs poorly for Family 14. There is a single item whose ISF is almost completely outside of the 95% prediction interval for the family. It is quite clear that the within

Figure 4: **Estimated posterior density functions of the family mean asymptotes λ_gs for the simulation study.**



family variance for this item is underestimated. Figure 6 contains the estimated posterior densities of the within family variance of the difficulty parameters (denoted as τ_bs) along with the simulating variance.

The simulating variance for the fourteenth family is the largest variance across all item families, but the approximated posterior density for this family is not substantially different from the other fifteen families. This might be an indication that the amount of information in the observed data about this variance component is small relative to the amount of information provided by the prior distribution. This is not surprising given that the simulated data has only ten items per item family, which is probably too few for estimating the within family variances.

Figure 5: The estimated FSFs (solid bold lines), corresponding 95% prediction intervals (solid lines), and the simulating ISFs (dashed lines) for the 16 item families in the simulated data set.

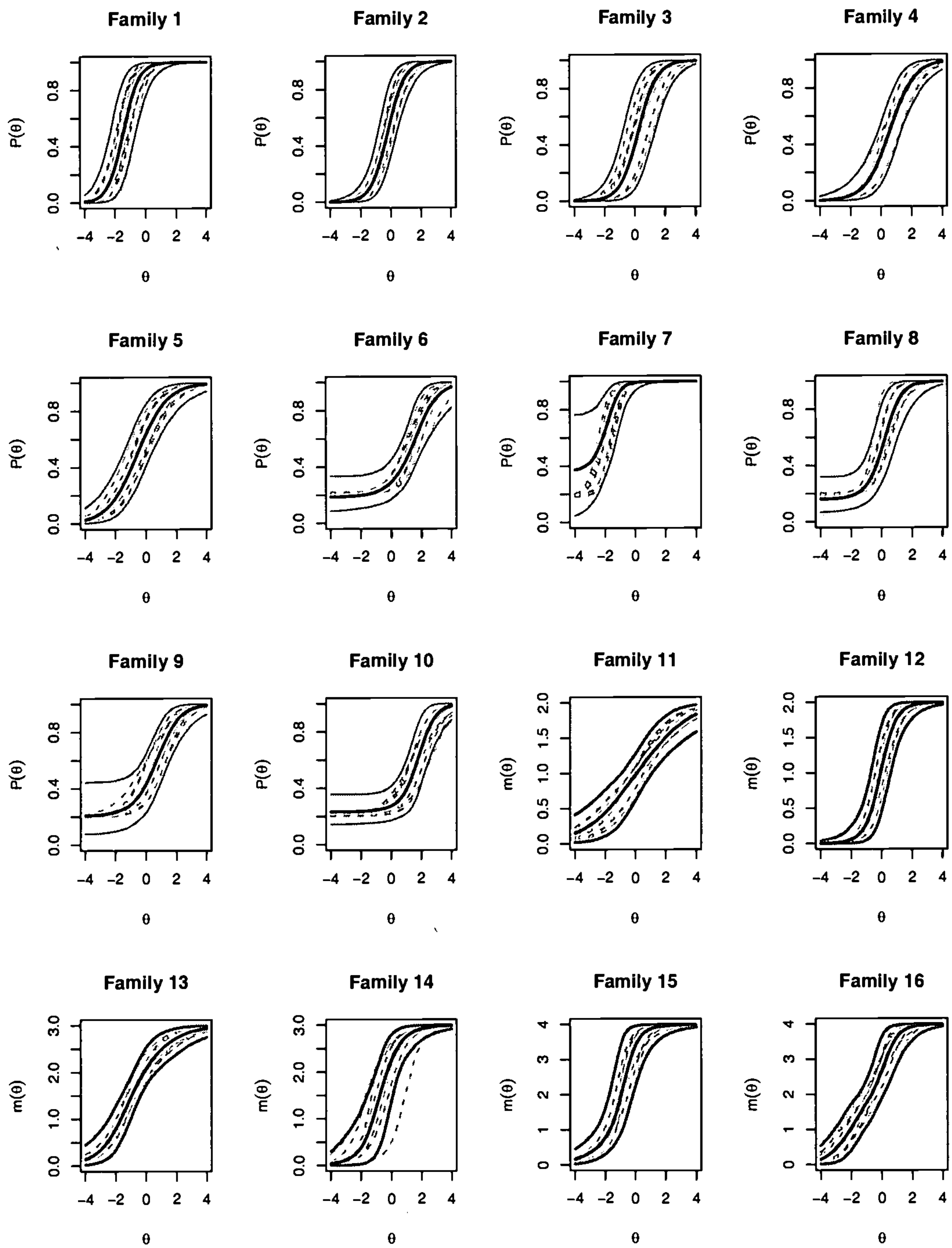
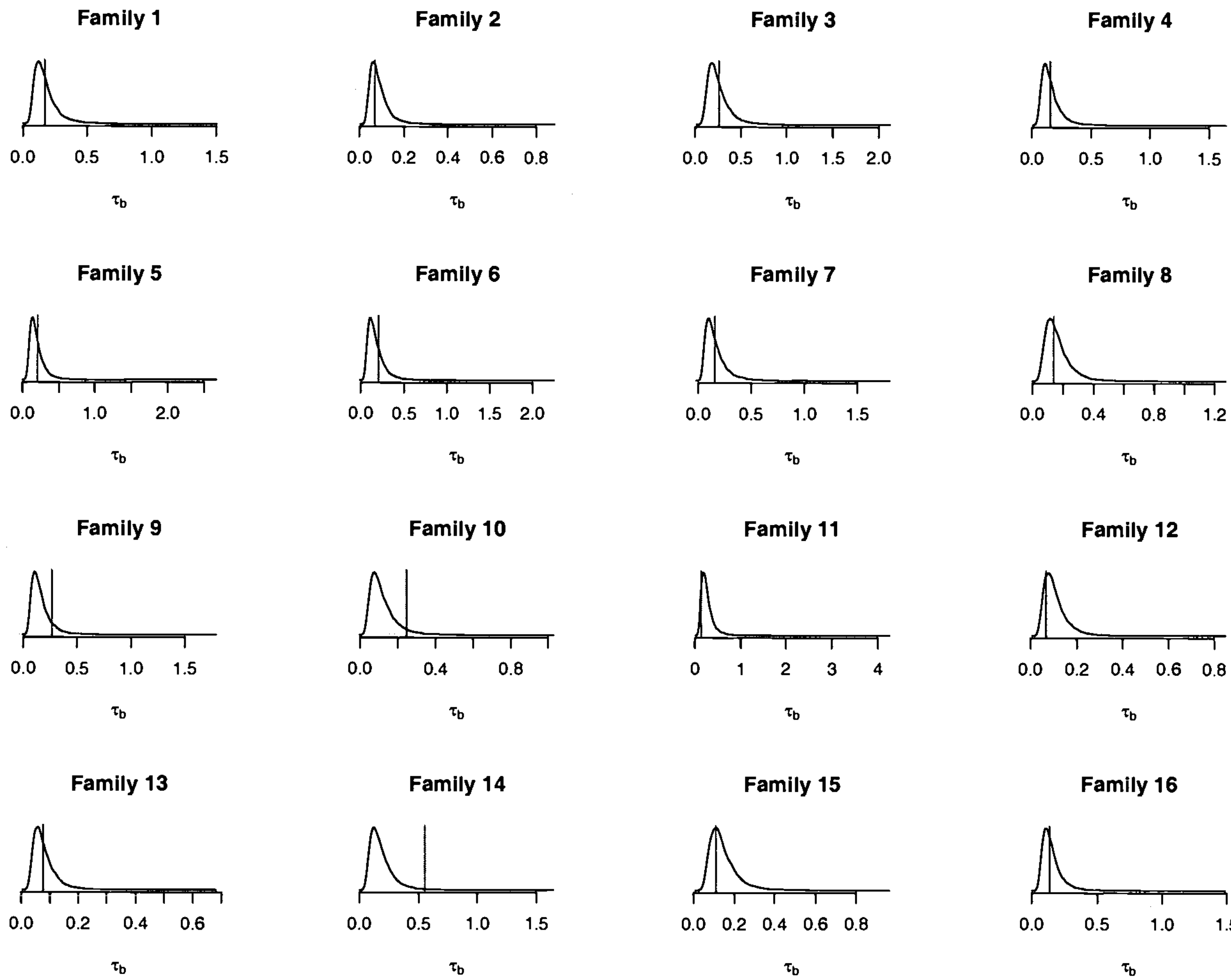
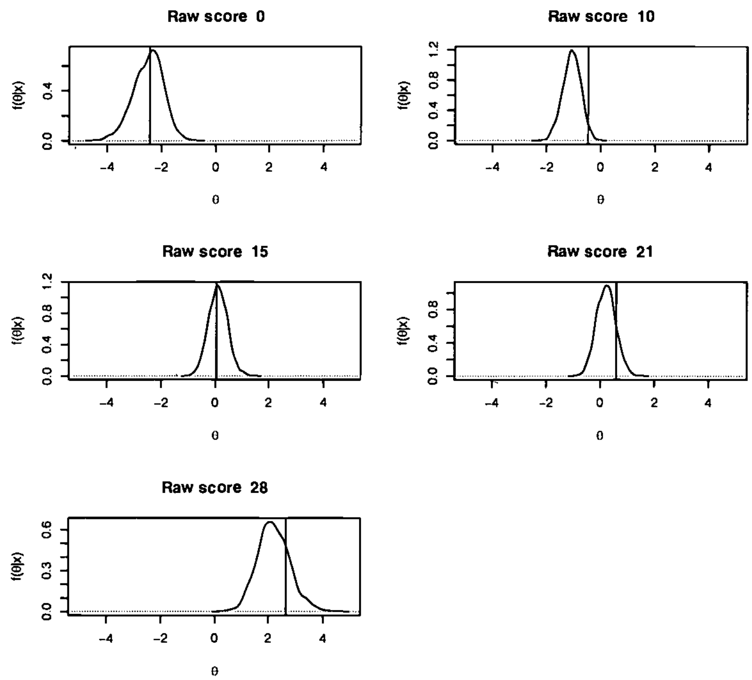


***Figure* 6: Estimated posterior density functions of τ_b, the within family variance component of the difficulties, for the simulation study**



In addition to estimating family parameters, the MCMC algorithm also provides us with sampled values from the approximate posterior distribution of each examinee ability θ_i. Figure 7 contains the approximated posterior densities of five individual θ_i's—an individual with lowest raw score (0), an individual with the 25th percentile raw score (10), an individual with the median raw score (21), and an individual with highest raw score (28). The true values of the θ_i's are also shown using a vertical line. The 95% posterior credible interval contains the true simulating value of θ in all the five cases. Although the 95% credible interval $(-1.76, \ -0.44)$ for the individual with a raw score of ten barely contains the simulating value of $\theta = -0.46$, the posterior probability that the individuals θ is greater than -0.46 is only $Pr\{\theta > -0.46 \mid \boldsymbol{x}_i\} = 0.03$.

Figure 7: **Markov chain Monte Carlo approximated posterior densities for five of the simulated examinees.**



6. Example 2: Analysis of the NAEP MathOnline Data

The National Assessment of Educational Progress (NAEP) is an ongoing educational survey administered by the National Center for Education Statistics (NCES). NAEP regularly reports on the progress of students in fourth, eighth, and twelfth grade on a number of educational subjects (e.g. mathematics, reading).

The Technology-Based Assessment (TBA) project is a NAEP special study sponsored by NCES. The project is designed to explore the use of technology, especially computers, as a tool to improve the quality and efficiency of assessments (NCES, 2002). One of the studies included in the TBA project is the Mathematics Online (MOL) special study (Sandene, Bennett, Braswell, & Oranje, in press). The MOL study translates existing NAEP math questions into a computer delivery system to be used for the assessment of forth and eighth grade students. The main goals of the MOL study were to: (a) determine how computer delivery affects the assessment of the examinees, (b) evaluate the abilities of fourth and eighth grade students to use a computer based assessment, (c) investigate the ability to create alternate versions of the assessment with the use of automatic item generation (AIG) (at grade 8 only). In the following pages we focus on the eighth grade MOL study.

Data

This paper looks at responses of 3793 examinees in grade 8, distributed among four test forms (denoted M2-M5). Each test form has a block of common items (denoted MP) and an additional 26 items that varies, either in content or delivery method, across the four forms. Of the 26 items that varies across the forms, sixteen are multiple-choice items and ten are constructed response items. Of the ten constructed response items, five items have two response categories each, three have three response categories each and two have four response categories each.

The items on form M2 are the "parent" items. These are all written by human, and representative of the NAEP mathematics item pool. Form M2 is a paper & pencil assessment much like the standard NAEP assessments (only shorter) with calculators provided for the students.

The content of form M3 is identical to form M2. However, form M3 is a computer-based assessment form and students must use an online calculator.

Forms M4 and M5 contain eleven items (six CR and five MC) that are identical to items on forms M2 and M3. The remaining fifteen items on forms M4 and M5 are automatically generated (Singley & Bennett, 2002) from an item model based on the items on form M2. However these fifteen items are not the same on forms M4 and M5. So when compared to one another, forms M4 and M5 have eleven identical items and fifteen items that vary across the two forms. Forms M4 and M5 are paper & pencil forms with calculators provided where necessary like form M2.

Table 1 summarizes the 26 item families; it provides the reader with the item types (e.g. 3PL, 3 category etc.) in each family and whether or not the items in M4 and M5 were automatically generated.

Sinharay, Johnson, and Williamson (2003) analyzed the multiple choice items from this data set with Glas and van der Linden's (1999) RSM. The later part of this section analyses this data set using the RSM introduced in Section 3 to demonstrate the practicality of the model. There are twenty-six item families, one corresponding to each item on form M2. Although some of the items are identical across the forms, we treat the items on each form as distinct items.

Table 1: **The item-types and indicators of whether not the items are automatically generated for the item families for the MOL data set**

Family	Type	AIG items (Y/N)
1	3PL	N
2	2PL	Y
3	3 Category	Y
4	3PL	N
5	3PL	Y
6	3PL	Y
7	2PL	Y
8	3PL	N
9	3PL	Y
10	4 Category	N
11	3PL	Y
12	3PL	Y
13	3 Category	N

Family	Type	AIG items (Y/N)
14	3PL	Y
15	2PL	N
16	2PL	N
17	2PL	N
18	3PL	N
19	3 Category	Y
20	3PL	Y
21	3PL	Y
22	3PL	N
23	3PL	Y
24	3PL	Y
25	3PL	Y
26	4 Category	N

The eleven item families for items that do not vary across the forms give us some idea about the amount of tolerable variation (Rizavi, Way, Davey, & Herbert, 2002), as the variation is simply a combination of sampling and administration variation. The analysis ignores two pieces of information that should be utilized in a full analysis of this data. The first is the common block of items in each form (MP). The second is whether the examinee completed the assessment online or with paper & pencil.

Analysis

We use the same Wishart prior distribution for $\boldsymbol{T}_{\mathcal{I}}^{-1}$ as we did for the simulated data set. We approximate the posterior distribution of the model parameters using 100,000 iterations from an MCMC algorithm; the first 10,000 iterations treated are discarded, and the remaining 90,000 iterations are thinned by selecting every 9th iteration for inclusion in the final sample of data from the approximated posterior distribution. Convergence diagnostics are applied, as in the simulation study, to make sure that the MCMC algorithm converges.

Figures 8 and 9 contain the estimated family score functions (along with the 95% prediction intervals) and the item score functions for each family; the former shows the families without any automatically generated items and the latter shows the families with automatically generated items.

The item families without AIG items generally have a set of ISFs that are closer to the FSFs than families with AIG items. This is, of course, not surprising considering the fact a family without AIG items contains the same item appearing in different forms. Despite the generally close ISFs for item families without AIG items, there is some variation evident among the ISFs for these families. The greatest observed variation occurs in Families 13 and 26, where the item on Form M3 behaves different than the other three items, and in Family 10, where the ISFs appear to be quite different at the high end of the scale.

Examination of the families that contain AIG items reveals a couple of clearly visible deviations. Most obvious is the fact that the entire family of items for family 9 is flat, suggesting students have the same random chance to correctly answer that question, regardless of their ability level. Since this is true for both the human generated item (appearing on form M2 and M3) and the AIG items (appearing on form M4 and M5), it appears that this is the result of a characteristic of the item type or content rather than the result of anything inherent in automatic item generation. In fact, in the operational analysis of this data set, this item has been dropped from the analysis (Sandene, Bennett, Braswell, & Oranje, in press).

Family 5, also an AIG family, contains one ISF that is quite different from the other three. In this family, the manually generated items in M2 and M3 and the AIG item appearing in M5 all have very similar ISFs while the AIG item from block M4 deviates dramatically from the other three item ISFs in the family. The extent of the deviation appears to impact the response function for the family as a whole.

Figure 10 contains the estimated posterior densities for five examinees from the Math Online study. Four of these examinees have varying raw scores (from very low to very high)— the estimated posterior distributions reflect that. The fifth examinee did not respond to any item, and hence the posterior we observe is simply the $N(0, 1)$ prior distribution.

Figure 8: Estimated FSFs, corresponding 95% prediction intervals (bold solid lines), and ISFs (lighter curves: small dashed lines for M2, dotted lines for M3, dots and dashes for M4, and long dashes for M5) for the 11 families that have no AIG items.

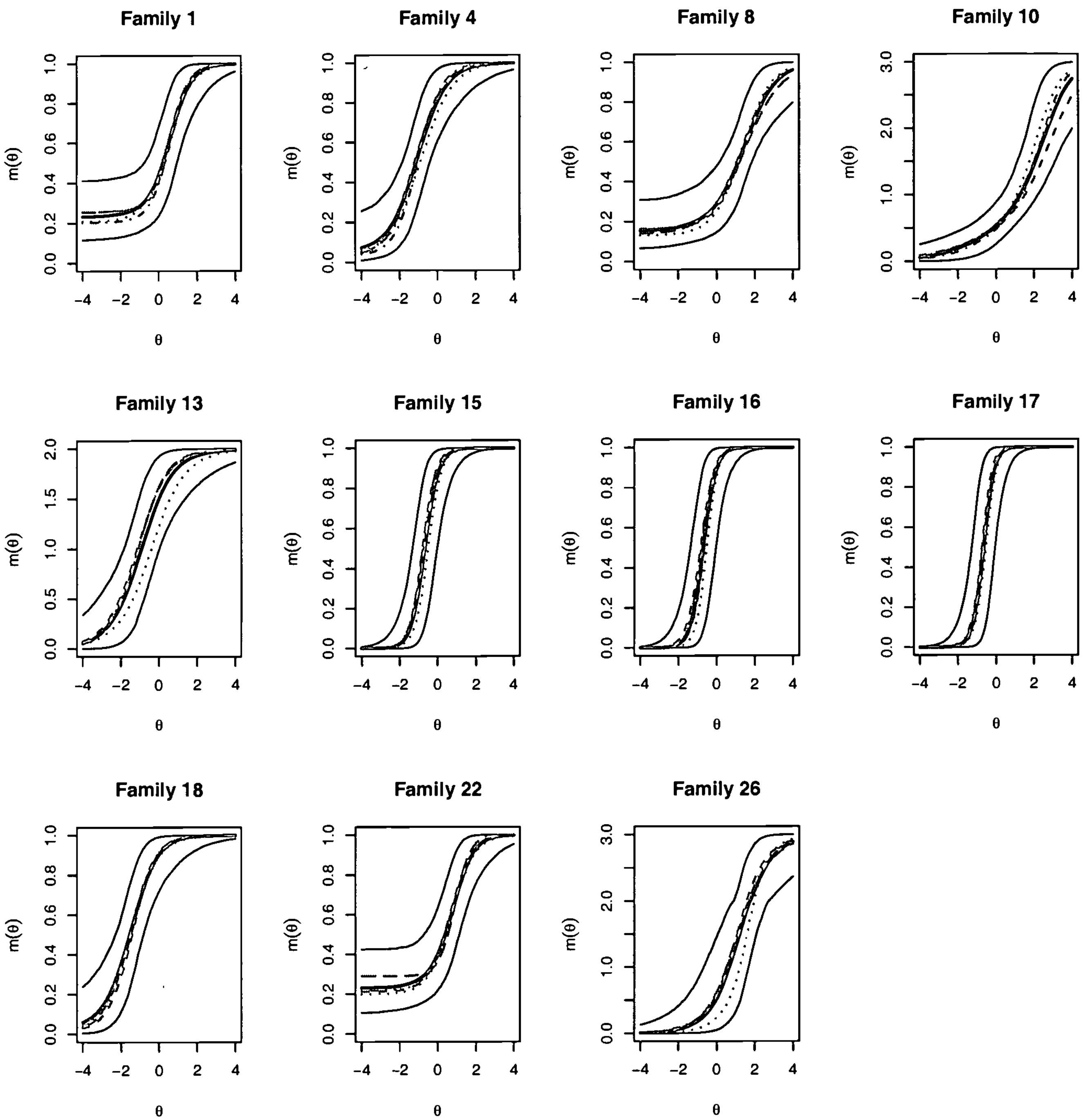


Figure 9: **Estimated FSFs, corresponding 95% prediction intervals (bold solid lines), and ISFs (lighter curves: small dashed lines for M2, dotted lines for M3, dots and dashes for M4, and long dashes for M5) for the 15 families with AIG items.**

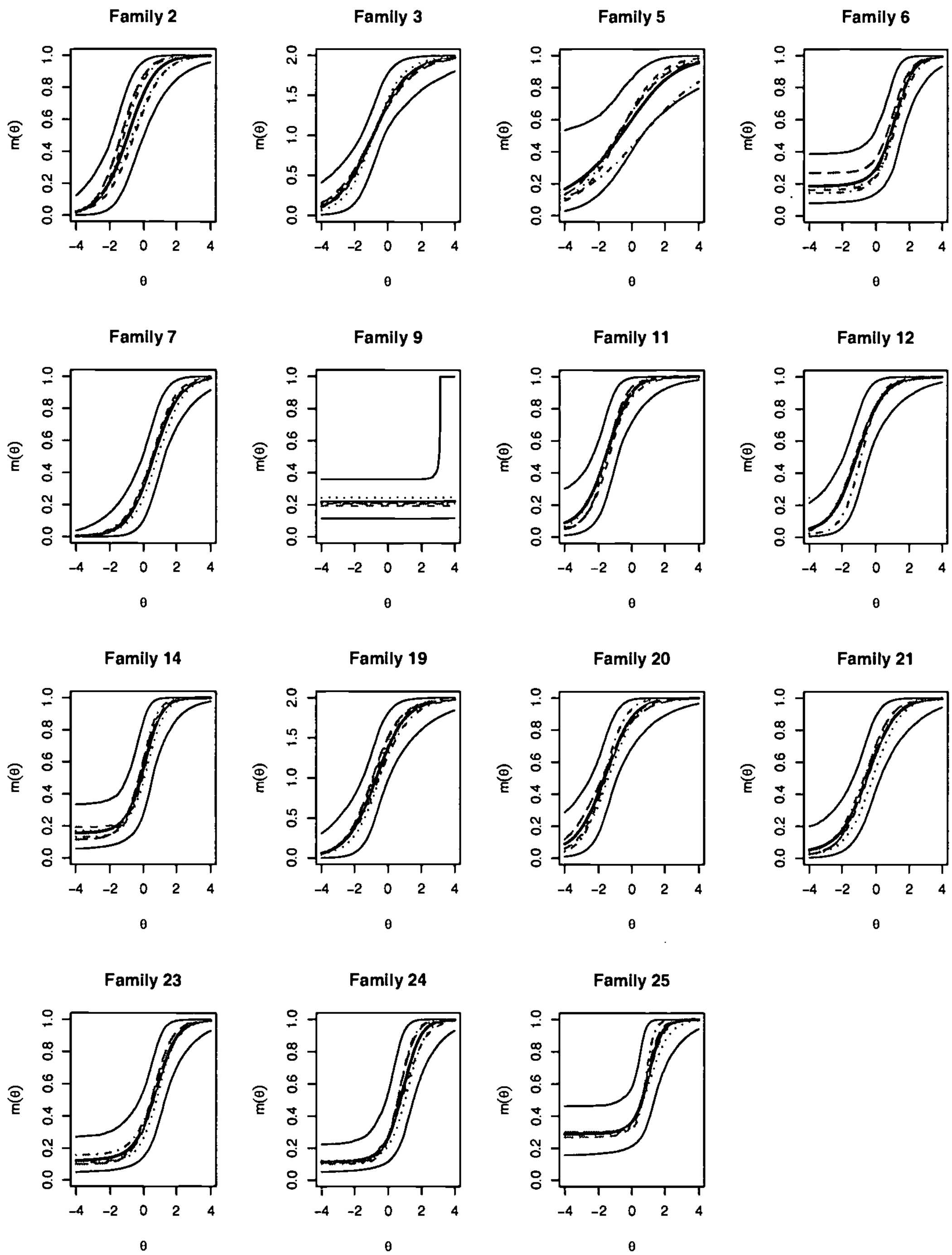
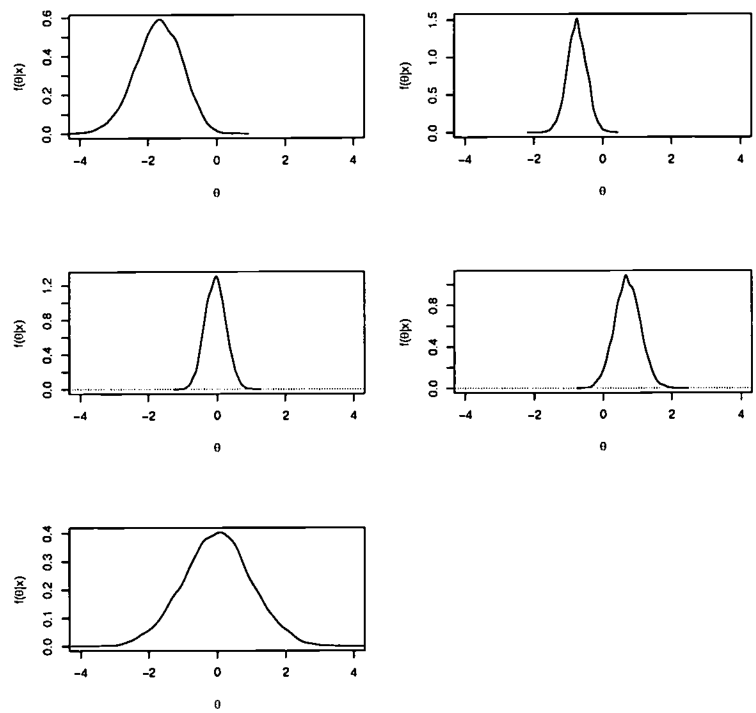


Figure 10: **The posterior densities of five examinees from the NAEP Math Online special study.**



7. Conclusions and Future Work

Our work shows that when a test consists of item families, the RSM can take into account the dependence among the items belonging to the same item family. The MCMC algorithm for Bayesian model fitting allows us to include the additional parameters in the hierarchical model without much additional difficulty. This work also suggests a useful way to summarize the results for an item family using the FSF. We believe that this work is an important step in creating a statistical tool that can be used to analyze tests involving item families. Such tests require calibration of an item family only once; the items belonging to the same family may be used in future tests without going into the trouble of calibrating those items. This will be very useful in automatic item generation systems where items are automatically generated from item models. However, a lot of additional research is required prior to such operational applications.

First among them is to find out the sample size required to achieve a pre-specified accuracy. It is clear that the model proposed is more complicated that a simple IRT model (USM) and hence would require a larger sample size that what is required for an USM— it will be helpful to be able to provide some guidance as to "how large" the sample size should be. Also, given a specific number of examinees who will take a test involving item families, we will like to determine the optimum values of the number of siblings per family and the number of examinees per sibling.

We would also like to study if the results of the analysis are sensitive to the prior distributions on the model parameters. Our analyses so far indicate that they are, especially the prior distributions on the hyperparameters corresponding to the within family variance when each item family consists of a few items. This is especially true with the MOL data set where there are only four items in each item family.

Finally, it will be quite helpful to be able to include covariates in the model, either task feature variables or demographic variables. For example, our analysis of MOL data in Section 6 ignored the facts that some examinees completed the assessment online while some others did it with paper and pencil and that some item families have AIG items while some do not. A model taking those facts into account might perform better than the one proposed here.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories for mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian Random Effects Model for Testlets. *Psychometrika,* 64, 153-168.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 6, 721-741.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman and Hall.

Glas, C. A. W. and van der Linden, W. J. (2001). "Modeling Variability in Item Parameters in CAT." Paper presented at the North American Psychometric Society Meeting in King of Prussia, PA.

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

Hombo, C. and Dresher, A. (2001). "A Simulation Study of the Impact of Automatic Item Generation under NAEP-like Data Conditions." Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Irvine, S. H. and Kyllonen, P.C. (eds.) (2002). *Item Generation for Test Development.* Mahwah, NJ: Lawrence Earlbaum Associates.

Janssen, R., Tuerlinckx, F., Meulders., M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*(3), 285-306.

Johnson, M. S. and Sinharay, S. (2002). "A Hierarchical Model for Item Model Calibration." Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). "Equations of state calculations by fast computing machine." *Journal of Chemical Physics,* 21, 1087–1091.

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

Muraki, E. (1992). "A Generalized Partial Credit Model: Application of an E-M Algorithm." *Applied Psychological Measurement,* 16, 159-176.

Muraki, E. and Bock, R.D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data.* Scientific Software International, Chicago, IL. [Computer Program].

NCES (2002). http://nces.ed.gov/nationsreportcard/studies.

Patz, R. and Junker, B. (1999). "Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses." *Journal of Educational and Behavioral Statistics,* 24, 342-366.

Rizavi, S., Way, W. D., Davey, T. and Herbert, E. (2002). "Tolerable variation in item parameter estimation." Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Samejina, F. (1969). "Estimation of latent trait ability using a response pattern of graded scores." *Psychometrika Monograph*, 17.

Sandene, B., Bennett, R. E., Braswell, J., & Oranje, A. (in press). *The Math Online Study: final report.* Washington, DC: National Center for Educational Statistics, Office of Educational Research and Improvement, US Department of Education.

Singley, M. K. and Bennett, R. E. (2002). "Item Generation and Beyond: Applications of Schema Theory to Mathematics Assessment." In *S. Irvine & P. Kyllonen (Eds.), Item Generation for Test Development. Hillsdale, NJ: Erlbaum.*

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). *An application of a Bayesian hierarchical model for item family calibration* (ETS RR-03-04). Princeton, NJ: Educational Testing Service.

Williamson, D. M., Johnson, M. S., Sinharay, S. and Bejar, I. (2002a). "Hierarchical IRT Examination of Isomorphic Equivalence of Complex Constructed Response Tasks." Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Williamson, D. M., Johnson, M. S., Sinharay, S. and Bejar, I. (2002b). "Applying Hierarchical Model Calibration to Automatically Generated Items." Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wright, D. (2002). Scoring tests when items have been generated. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 277-286). Hillsdale, NJ: Erlbaum.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM034943

I. DOCUMENT IDENTIFICATION:

Title:	Calibration of Automatically Generated Items Using Bayesian Hierarchical Modeling
Author(s):	Matthew S. Johnson and Sandip Sinharay
Corporate Source: Educational Testing Service	Publication Date: Unpublished Work

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Sandip Sinharay	Printed Name/Position/Title: Sandip Sinharay/Associate Research Scientist	
Organization/Address: Educational Testing Service, Rosedale Rd., MS 12T, ~~Princeton, NJ 08541~~	Telephone: 609-734-5079	FAX: 609-734-5420
	E-Mail Address: ~~sshinaray@ets.org~~ sshinaray@ets.org	Date: 5/12/03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org