An Application of a Bayesian Hierarchical Model for Item Family Calibration

Sandip Sinharay
Matthew S. Johnson
David M. Williamson
An Application of a Bayesian Hierarchical Model for Item Family Calibration

Sandip Sinharay
Educational Testing Service, Princeton, NJ

Matthew S. Johnson
Baruch College, New York, NY

David M. Williamson
Educational Testing Service, Princeton, NJ

February 2003
Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 10-R
Educational Testing Service
Princeton, NJ 08541
Abstract

Item families, which are groups of items related to each other in some way, are increasingly used in complex educational assessments. For example, in automatic item generation (AIG) systems, a test may consist of multiple items generated from a number of item models. Item calibration or scoring for such an assessment requires fitting models that can 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. We fit that hierarchical model using the Markov chain Monte Carlo (MCMC) algorithm. Formulating the MCMC algorithm provides little additional difficulty (compared to fitting a simple item response theory model) even with the additional complexity in the form of hierarchy in the model. We show that the model can take into account the dependence structure inherent among the items and hence is an improvement over the models currently used in similar situations. We introduce the notion of the family expected response function (FERF) as a way to summarize the probability of a correct response to an item randomly generated from an item family and suggest a way to estimate the FERFs. Our work is a step towards creating a tool that can save significant amount of resources for tests with item families, because calibrating only the item families might be enough rather than calibrating each item in the families separately.

Key words: Bayesian methods, hierarchical model, Markov chain Monte Carlo, item model, automatic item generation, expected response function
Acknowledgements

The authors thank Isaac Bejar and Randy Bennett for their helpful comments about the work and Kim Fryer for her help with proofreading. The authors are also thankful to the three referees, Alina von Davier, Paul Holland, and Hariharan Swaminathan, for their useful comments about the work.

In addition, the authors gratefully acknowledge the National Center for Education Statistics for allowing access to the data set for Section 6.
1. Introduction

The operation of large scale high stakes testing programs demands a large number of high quality items to populate item pools. Large pools are especially important in adaptive testing programs 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, interest is increasing in using item models to guide production of items with similar conceptual and statistical properties (see Irvine & Kylonen, 2002, for a survey of some current areas of investigation for item modeling and generation). Whether items result from automatic item generation (AIG) systems or from rigorous manual procedures, items produced from a single item model are related to one another through the common generating model and therefore constitute an item family.

Items from the same well-defined item family should be related to one another. Therefore, 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 that is more general. Our work is aimed at examining the latter model. The model assumes that the item parameters of a three-parameter logistic model (3PL) (Lord, 1980) are normally distributed with a mean vector and a variance matrix that depend on the item family from which the item is generated.

The hierarchical model described here has some similarity to the testlet model (Bradlow, Wainer, & Wang, 1999) in the sense that it describes an extra level of dependence between the item responses of individuals. However, the testlet model describes the extra “local” dependence between a single examinee’s item responses within a testlet, whereas the model we describe here explains the dependence between all examinees’ responses to the same single member from an item family.
We show that it is possible to fit the model easily using the Markov chain Monte Carlo (MCMC) algorithm (Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996). The hierarchical model implies for each item family a family expected response function (FERF) that gives the probability of a correct response to an item randomly generated from the item family for a given examinee ability. We suggest a way to compute estimates of the FERF and an approximate confidence bound for it using the Monte Carlo method and the output of the MCMC algorithm. We study how the results obtained using a hierarchical model compare to those obtained by other existing methods for tackling similar problems. The work is a continuation of Johnson and Sinharay (2002) with extensions to real data examples taken from Williamson, Johnson, Sinharay, and Bejar (2002a; 2002b).

The next section introduces the hierarchical model. Section 3 talks about estimation of the model parameters and the family expected response functions using the MCMC algorithm. Section 4 discusses results from a simulation study. In Section 5, we apply the hierarchical model to an operational data set from a high stakes assessment. Section 6 discusses the application of the model to a data set from the National Assessment of Educational Progress (NAEP). Finally, we discuss our conclusions and future directions in Section 7.

2. The Hierarchical Model

Suppose there are $J$ items denoted by $j = 1, 2, \ldots, J$ in a test. For this work, we assume that all items are dichotomously scored. Consider that the test is given to examinees $i = 1, 2, \ldots, N$. An item family is defined as a group of items that are believed to be related to one another in some way. For example, the items may be generated from the same item model (parent). Let $\mathcal{I}(j)$ be the item family from which the item $j$ is generated. Items $j$ and $k$ are siblings if they come from the same item family, i.e., if $\mathcal{I}(j) = \mathcal{I}(k)$.

There are three approaches for modeling data involving item families. We discuss the models in the context of the 3PL (Lord, 1980) models, but one can think of the models in the context of other types of item response functions (e.g., 2PL model) as well.
Identical Siblings Model

One simple way to model item families is to assume the same item response function for all items in the same family (Hombo & Dresher, 2001). This means modeling the response function for the $i$-th person to the $j$-th item as

$$P_j(\theta_i) = c_{x(j)} + \frac{1 - c_{x(j)}}{1 + \exp\{a_{x(j)}(\beta_{x(j)} - \theta_i)\}}.$$

We call this approach the "Identical Siblings Model" (ISM). While this model can be fit by standard software (like PARSCALE or BILOG), 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.

Unrelated Siblings Model

"Unrelated Siblings Model" (USM) assumes a separate, unrelated (independent) item response function for all items, regardless of their family membership. Mathematically, the model is represented as

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp\{a_j(\beta_j - \theta_i)\}}.$$

While the USM can be fit by standard software, this has the disadvantage that each item has to be individually calibrated. Also, this approach ignores the relationship between siblings in an item family and hence will provide standard errors of item parameters that are larger than the truth.

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 & van der Linden, 2001). Mathematically, this model assumes

\[
P_j(\theta_i|a_j, \beta_j, c_j) \equiv Pr\{X_{ij} = 1 \mid \theta_i, a_j, \beta_j, c_j\} = c_j + \frac{1-c_j}{1+\exp[a_j(\beta_j-\theta_i)]}
\]

\[
\theta_i \sim \mathcal{N}(\mu, \sigma^2)
\]

\[
\alpha_j \equiv \log\{a_j\}
\]

\[
\gamma_j \equiv \log\left\{\frac{c_j}{1-c_j}\right\}
\]

\[
(\alpha_j, \beta_j, \gamma_j)^t \mid \lambda_{\mathcal{I}(j)}, T_{\mathcal{I}(j)} \sim \mathcal{N}_3(\lambda_{\mathcal{I}(j)}, T_{\mathcal{I}(j)})
\]

(1)

It should be noted that expectations are not invariant under transformations, e.g.,

\[E[a_j] \neq E[e^{\alpha_j}].\] However, because the transformation is monotone increasing, the quantiles (including the median) are invariant under the transformation. To fix the origin and scale of the ability parameters (\(\theta_i\)s), we set the parameters of their prior distribution at \(\mu = 0, \sigma^2 = 1\) as is done with standard IRT models. To complete the model specification, we assume independent multivariate normal prior distributions for \(\lambda_{\mathcal{I}(j)}\)s,

\[
\lambda_{\mathcal{I}(j)} \sim \mathcal{N}_3\left(\mu_\lambda \equiv \begin{pmatrix} 0 \\ 0 \\ \mu_{\lambda_\beta} \end{pmatrix}, V_\lambda \equiv \begin{bmatrix} 100^2 & 0 & 0 \\ 0 & 100^2 & 0 \\ 0 & 0 & \sigma_{\lambda_\gamma}^2 \end{bmatrix}\right),
\]

(2)

and independent inverse Wishart prior distributions on the \(T_{\mathcal{I}(j)}\)s,

\[
T_{\mathcal{I}(j)}^{-1} \sim \text{Wishart}(W_1, W_2),
\]

(3)

where, \(M \sim \text{Wishart}(W_1, W_2)\), \(M\) having dimension \(k \times k\), implies that the density of \(M\) is proportional to

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

Note that by integrating the individual item parameters \((\alpha_j, \beta_j, \gamma_j)^t\) out of the likelihood, the resulting model correctly accounts for the fact that responses by two individuals to the same item are correlated even when conditioning on the family parameter \(\lambda_{\mathcal{I}(j)}\) and \(T_{\mathcal{I}(j)}\).

Notice also in (2) that we assume noninformative prior distributions on the first two components (those corresponding to the discrimination parameter \(\alpha_j\) and difficulty parameter \(\beta_j\) of \(\lambda_{\mathcal{I}(j)}\), to allow the data to provide all the information about the posterior
distribution. However, the prior distribution on the third component (that corresponding to the guessing parameter $\gamma_j$) is usually taken to be somewhat informative, mainly because a noninformative prior on the guessing parameter results in a very slow convergence of the MCMC algorithm used to fit the model. Also, it is possible to extract some information about the guessing parameter from the test type and format. For example, if we are dealing with multiple choice items with five choices and good distractors for the other four choices, we can assume that, on average, an examinee will have a probability of around 0.2 to get any item correct by random guessing; this results in taking

$$
\mu_{\lambda_3} = \log \left( \frac{0.2}{1 - 0.2} \right) = -1.39
$$

and a value of $\sigma_{\lambda_3}^2$ that is not too large. When the number of choices is different, or some prior knowledge about the quality of the distractors is known, the prior distribution is changed accordingly.

Figure 1 shows a plot of the probability density function of the prior distribution with $\mu_{\lambda_3} = -1.39$ and $\sigma_{\lambda_3}^2 = 0.01$. Note that the variable whose density is plotted is a transformation of the third component of $\lambda_{\pi(j)}$ to make it in the same scale as the guessing parameters $c_j$s.

The assumption (3) implies a prior mean of $W_1W_2$ for $T_{\pi(j)}^{-1}$ and that a priori we have information that is equivalent to $W_1$ observations on $(\alpha_j, \beta_j, \gamma_j)$. We can set these parameters depending on our knowledge of the problem.
One important thing to notice here is that the ISM and USM are limiting cases of the RSM. If the variance parameters in $T_{I(j)}$ become very small in the RSM, there is no variation among the items in the same family and we get the ISM. On the other hand, if the variances become very large (much larger than the variance between the $\lambda_{I(j)}$s), the family means overlap so that there is no way to distinguish between families and we get the USM.

**Family Expected Response Function (FERF)**

For the RSM, it may be interesting to look at a response function typical to an item family or the probability of a correct response for an item family as a function of the examinee ability $\theta$. Note that the hierarchical structure of the RSM suggests that to obtain the response function, one has to take the average of

$$P_j(\theta|a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp\{a_j(\beta_j - \theta)\}}$$

over all possible values of the item parameters that one may observe for that particular item family. We suggest averaging out the item parameters as shown below

$$P(\theta|I(j)) = \int_{\lambda_{I(j)}, T_{I(j)}} \int_{\eta_j} P_j(\theta|\eta_j) f(\eta_j|\lambda_{I(j)}, T_{I(j)}) d\eta_j f(\lambda_{I(j)}, T_{I(j)}|X) d\lambda_{I(j)} dT_{I(j)}$$

(5)

to obtain the FERF for item family $I(j)$ that gives the probability of a correct response to an item randomly generated/selected from that particular item family for a given examinee ability, where $\eta_j = (\alpha_j, \beta_j, \gamma_j)^t$, $f(\eta_j|\lambda_{I(j)}, T_{I(j)})$ is the density function of the multivariate normal prior distribution on $\eta_j$ and $f(\lambda_{I(j)}, T_{I(j)}|X)$ is the joint posterior distribution of $\lambda_{I(j)}$ and $T_{I(j)}$ given the response matrix $X$.

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 model in our work.

3. Bayesian Estimation for the RSM

We use Bayesian estimation in our work. The MCMC algorithm, specifically the *Gibbs sampler* (Geman & Geman, 1984), helps us to generate a sample from the joint posterior distribution of the parameters of the model by drawing iteratively from the
conditional posterior distribution of each model parameter. Item parameters \( \alpha, \beta, \text{ and } \gamma \) and the ability latent variables \( \theta \) have the same form of posterior distribution as what would have been obtained under a simple 3PL model (Lord, 1980). These parameters are drawn from their respective conditional distributions using Metropolis steps as described in Patz and Junker (1999). Denote the item family mean vector and covariance matrix of the \( k \)-th item family as \( \lambda_k \) and \( T_k \) respectively. Conditional on the item parameters \( \alpha, \beta, \text{ and } \gamma \), \( \lambda_k \) and \( T_k \) are independent of \( \theta \) and the observed data \( X \). The conditional distributions of the \( \lambda_k \)'s, which are trivariate normal and independent over the families (i.e., over \( k \)), are given by

\[
\lambda_k \mid \alpha, \beta, \gamma, T_k \sim \mathcal{N}_3 \left( V_k \left\{ V^{-1}_\lambda \mu_\lambda + J_k T_k^{-1} \left( \begin{array}{c} \bar{\alpha}_k \\ \bar{\beta}_k \\ \bar{\gamma}_k \\ \end{array} \right) \right\}, V_k \right),
\]

where

\[
V_k = (J_k T_k^{-1} + V^{-1}_\lambda)^{-1},
\]

\( \mu_\lambda \) and \( V_\lambda \) are the prior mean and variances of \( \lambda_k \)'s respectively,

\[
\bar{\alpha}_k = \frac{1}{J_k} \sum_{j : X(j)=k} \alpha_j, \quad \bar{\beta}_k = \frac{1}{J_k} \sum_{j : X(j)=k} \beta_j, \quad \bar{\gamma}_k = \frac{1}{J_k} \sum_{j : X(j)=k} \gamma_j,
\]

and \( J_k \) is the number of members in item family \( k \).

The conditional distributions of the \( T_k \)'s, which are independent over the families (i.e., over \( k \)), are given by

\[
T_k \mid \eta, \lambda_k \sim \text{Inv-Wishart} \left( J_k + W_1, \left\{ \sum_{j : X(j)=k} (\eta_j - \lambda_k)(\eta_j - \lambda_k)^t + W_2^{-1} \right\}^{-1} \right),
\]

where

\[
\eta_j = (\alpha_j, \beta_j, \gamma_j)^t.
\]

Hence, the addition of the hierarchical component in the model amounts to additional sampling from normal and inverse Wishart distributions, which are both easy to achieve. So the hierarchy of the model does not pose significant difficulties for the Bayesian estimation procedure. Note that the MCMC algorithm described here (and referred to as the “Gibbs
sampler” here) is also known as the “Metropolis-Hastings within Gibbs” algorithm (Patz & Junker, 1999) because the algorithm is basically a Gibbs sampler with a switch to Metropolis steps for generating \( \alpha, \beta, \gamma, \) and \( \theta. \)

**Estimating the Family Expected Response Function**

We use Monte Carlo integration to estimate the FERF defined in (5). We also discuss how to attach a 95% prediction interval with the estimate. The steps required in the estimation process for the \( k \)-th family are shown below:

i) Generate a sample of size \( M \) from \( f(\lambda_k, T_k | X) \), which can be obtained as a by-product of the MCMC algorithm used to fit the hierarchical model.

ii) For each of the above \( M \) values, draw \( m \) values of the item parameter vector \( \eta_j \) from \( \Phi_3(\eta_j | \lambda_k, T_k). \)

iii) Set the ability at \( \theta. \)

iv) For each of the \( Mm \) draws of \( \eta_j \) obtained in Step iii), compute \( P_j(\theta|\eta) \) using (4).

v) Take the mean of the above probabilities as an estimate of \( P(\theta|k) \), the probability of correct response to an item randomly drawn from item family \( k. \)

vi) The 2.5th and 97.5th percentiles of the \( Mm \) probabilities above form an approximate 95% prediction interval to attach with the estimate obtained in the previous step. estimated \( P_k(\theta|k). \)

Steps iii) to vi) are repeated for a number of values of \( \theta \) to obtain the estimated FERF for family \( k. \) We use 100 equidistant values of \( \theta \) in the interval (-4,4) to estimate the function. We use \( M = 1000 \) and \( m = 10. \) Note that we already have a posterior sample that we obtained using the MCMC algorithm as described earlier in this section. To obtain a sample of size \( M \) from \( f(\lambda_k, T_k | X) \) in Step i) above, we draw a subsample of size \( M \) from the posterior sample. Step ii) is simple because here we only have to sample from a multivariate normal distribution using the draws of \( \lambda_k \) and \( T_k \) from the previous step. So
the estimation of the FERF is quite straightforward given the output from the MCMC and takes little additional time.

4. Results of a Simulation Study

In this section, we first investigate how the RSM performs when fitted to a data set generated so that it has a hierarchical structure. This study should give us some idea about the performance of the RSM. We also examine the performance of the USM and ISM (the existing models for calibrating item families) when the data set has a hierarchical structure. We further generate data sets under the USM and ISM assumption (implying no hierarchical structure in the data sets) and analyze them using the RSM.

Generating the Data Sets

We first describe generating the data set under RSM assumption. We assume that there are 10 item models, 10 items from each item model, and a total of 5,000 examinees. Each examinee receives exactly one item from each item model. The sampling design is balanced so that any item is assigned to exactly 500 examinees. The steps used to generate data sets from the RSM are:

i) Generate \( \theta_i \sim N(0, 1), i = 1, \ldots 5000. \)

ii) Generate

\[
U_{1k} \sim U(0.5, 1.5), k = 1, \ldots 10,
\]
\[
U_{2k} \sim U(-2, 2), k = 1, \ldots 10,
\]
\[
U_{3k} \sim U(0.2, 0.4), k = 1, \ldots 10,
\]

and then define \( \lambda_k = (U_{1k}, U_{2k}, U_{3k})', k = 1, \ldots 10. \) The \( U_{jk}s \) reflect the range of values observed when 3PL models are fitted to real data sets. For example, from our experience, the discrimination parameters in 3PL models usually fall between 0.5 and 1.5 and hence \( U_{jk}s \) are generated from that range. The \( \lambda_k's \) are generated from the \( U_{jk}s \) using the proper transformations (remembering the transformations while in the
RSM section). The between family variance implied by the way we generate the family means is given by

\[
Var(\lambda_k) = \begin{pmatrix}
0.10 & 0 & 0 \\
0 & 1.33 & 0 \\
0 & 0 & 0.08
\end{pmatrix}.
\]

iii) For \(j = 1, 2, \ldots 100\), generate

\[
\begin{pmatrix}
\alpha_j \\
\beta_j \\
\gamma_j
\end{pmatrix} \sim N_3\left(\lambda_{i(j)}, T_{x(j)} \equiv \begin{pmatrix}
0.02 & 0 & 0 \\
0 & 0.13 & 0 \\
0 & 0 & 0.01
\end{pmatrix}\right).
\]

Each variance parameter here is taken as a fraction of the corresponding between family variance. This was done because if the item parameters in this step are generated with a variance as big as the variance between the \(\lambda_k\)s, then (remembering that the USM is a limiting case of the RSM when the variance of the item parameters become very large) we would be generating data from the USM and not the RSM.

iv) Compute \(a_j = \exp(\alpha_j), c_j = \frac{\exp(\gamma_j)}{1 + \exp(\gamma_j)}, j = 1, 2, \ldots 100\).

v) Generate \(X_{ij} \sim Bern(P_j(\theta_i), P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp(\alpha_j(\beta_j - \theta_i))}\).

Data sets from the USM and ISM are generated similarly and are not described here. We set the variation of the item parameters and the examinee ability parameters in the data sets generated from the USM and ISM about the same as that in the RSM.

**Analysis**

While analyzing data sets (generated from any of the three models) under the RSM assumption, the model defined by (1) is used. We use prior distributions for the item family parameters that are given by (2) and (3). We take \(\sigma^2_{\lambda} = 10^2, W_1 = 5\) and take \(W_2\) to be a diagonal matrix with 2 as its elements. This implies that a priori we have information equivalent to that of five items in each item family. The prior mean of the precision matrix \(T_{x(j)}^{-1}\) implied by this assumption is \(10 \times I_3\).
While analyzing under the USM and ISM, the prior distributions assumed on the parameters, using the same type of parameterization as in (1), are:

\[
\theta_i \sim N(0, 1), \quad \alpha_j \sim N(0, 10^2), \quad \beta_j \sim N(0, 10^2) \text{ and } \gamma_j \sim N(-1.39, 0.5).
\]

These prior distributions are used for all the USM and ISM analyses in this work and hence are not repeated further.

The MCMC algorithm, specifically the Gibbs sampler, is used to fit all the models to the data set. We run five chains of 10,000 iterations each. To monitor convergence of the algorithm, we examine the Gelman-Rubin convergence measure (Gelman & Rubin, 1992) and the time-series plot of a carefully chosen set of parameters (including all of the variance parameters, some mean parameters, and few item parameters and ability parameters). These measures indicate that the sampler converges after 2,000 iterations for the RSM and about 1,000 iterations under the USM and ISM. We retain the final 8,000 iterations from each of the five chains in the posterior sample, giving us a final sample of size 40,000. We compute 95% posterior intervals for the model parameters and compare the intervals against the corresponding true values. We also examine histograms of the sampled values of the parameters. Finally, for the analysis under the RSM, we look at the estimated FERF for each item family along with the predicted item characteristic curves for the individual items.

The following are the results obtained under the different situations arising out of the “data model”-“analysis model” combinations.

**Data From an RSM, Analyzed Using an RSM**

Suppose \( \lambda_k = (\lambda_{a_k}, \lambda_{b_k}, \lambda_{g_k}) \). Then \( \lambda_{a_k} \)s are family means corresponding to the slope parameters, \( \lambda_{b_k} \)s family means of the difficulty parameters, and \( \lambda_{g_k} \)s family means corresponding to the guessing parameters. Figures 2, 3, and 4 show plots of the histograms of the sampled values of \( \lambda_{a_k} \), \( \lambda_{b_k} \), and \( \lambda_{g_k} \) for all the item models. In all these plots, the item models should be counted across rows. On each histogram, a vertical line is drawn at the true value of the corresponding parameter. The figures show that the 95% posterior intervals almost always cover the true values of the \( \lambda \)s with the only exception of Item Model 6. Looking at Figure 4, we see that the true value of \( \lambda_{g_k} \) for the Item Model 6 is
Figure 2: Histograms of the sampled values of λₐₙₛ for all the item models (ordered across rows) for the RSM-RSM case.

significantly larger than the other λᵦₙₛ and hence the model cannot explain that large a value.

Figure 3 tells us that the RSM can very well differentiate between the item models so far as the difficulty parameters are concerned. The confidence intervals for the λₐₛ are clearly distinct and mostly nonoverlapping.

For the item parameters, the coverage of the 95% posterior intervals are 100% for the αₛ, 95% for the βₛ, and 95% for the γₛ. Interestingly, the only item model for which the intervals don’t cover the true item parameters is Item Model 6, which was found to be
Figure 3: Histograms of the sampled values of $\lambda_{b_k}$s for all the item models (ordered across rows) for the RSM-RSM case.

Figure 5 shows a histogram of $\tau_{b_k}$s, the sampled values of the within family variance of the difficulty parameters. As easily seen from (1), $\tau_{b_k}$ is the second diagonal element of $T_k$. The true variance parameters are always covered in the 95% posterior intervals. Also, remembering that the true variances were taken to be the same, the model clearly does a good job.

Since we have the sampled values of the examinee ability parameters $\theta$s for all the iterations of the MCMC, we look at them to examine how the RSM performs so far.
Figure 4: Histograms of the sampled values of $\lambda_{gs}$ for all the item models (ordered across rows) for the RSM-RSM case.

As person fit is concerned. Figure 6 shows histograms of the sampled values of $\theta_i$s for five typical individuals—those for which the true $\theta_i$s are the minimum, 25th percentile, median, 75% percentile, and the maximum among all the individuals. Figure 6 shows a histogram of the sampled values of those $\theta_i$s. The true values of the $\theta_i$s are shown on the histograms with a vertical line. We find an interesting pattern here. Because each examinee is only responding to 10 items, there is substantial shrinkage to the prior mean. The expected a posteriori (EAP) estimates are shrunk towards zero, and the 95% posterior intervals do not cover the true $\theta_i$s for the extreme $\theta_i$s, i.e., those corresponding to the true minimum and
Figure 5: Histograms of the sampled values of the within family variance \( \tau_{j} \) of the difficulty parameters for all the item models (ordered across rows) for the RSM-RSM case.

maximum. For the intermediate \( \theta \)'s, the intervals cover the true values, but there is still evidence of shrinkage when the EAPs are examined.

We also plot the estimated item characteristic curves (ICCs) for all items for four item families (1, 2, 6, and 9) in Figure 7. These plots are created using the item response function described in the RSM section and the posterior median of the item parameters, \( \alpha_j \), \( \beta_j \), and \( \gamma_j \). For each item family, we also draw the estimated FERF (estimated using the technique suggested in Section 3) and a 95\% prediction interval in a bold line on top of the 10 ICC estimates for that item family. The estimated ICCs flock together
Figure 6: Histograms of the sampled values of five typical $\theta$s for the RSM-RSM case.

around the corresponding estimated FERF for each family clearly demonstrating their interdependence. Also, all the estimated ICCs lie within the 95% prediction interval.

Data From an RSM, Analyzed Using an USM

We fit the USM to the simulated data set because the USM is the gold standard of calibration if one has enough data. The coverage of the 95% posterior intervals in this analysis are 97%, 96%, and 97% respectively for $\alpha$s, $\beta$s, and $\gamma$s. However, the confidence intervals here are much wider than the RSM-RSM case; on average, the width of the confidence intervals are 1.5 times as much as that for the RSM-RSM case. This is expected because in the RSM-RSM analysis, information about an item is available from other items generated from the same item model, thus making the estimation process more precise.

The histograms of the sampled values for the five typical $\theta$s look very similar to the RSM-RSM case and hence are not provided here. This suggests that so far as scoring
Figure 7: Estimated ICCs and FERFs for four item families for the RSM-RSM case.

is concerned, if a hierarchical structure exists in the data, there is little difference between analysis using the USM and analysis using the RSM. However, if the former is applied, one has to calibrate each item separately, while in the latter case, it is enough to calibrate the item families alone, thereby saving considerable resources.

We show the estimated ICCs along with the estimated FERFs and the corresponding 95% prediction bounds from the RSM analysis in bold lines for four item models (1, 2, 6, and 9) in Figure 8. Each estimated FERF lies in the middle of the ICCs for the same family, but the ICCs are much more widely spread than in Figure 7. Also, a few of the estimated ICCs lie outside the prediction bounds as a result of our assumption that the items are independent.

Data From an RSM, Analyzed Using an ISM

The coverage of the 95% posterior intervals for the item parameters are 91%, 80%, and 96% respectively. The confidence intervals are slightly thinner than the RSM-RSM case — on an average, the width of the confidence intervals are 0.85 times as much as that
Figure 8: Estimated ICCs and FERFs for four item families for the RSM-USM case.

for the RSM-RSM case. This is also expected because the number of observations per item is larger here than in the RSM-RSM analysis, making the confidence intervals more precise.

The histograms of the sampled values for the five typical \( \theta \)s look very similar to the RSM-RSM case and hence are not provided here. This similarity implies that so far as scoring is concerned, if a hierarchical structure exists in the data, there is not much of a difference between analyzing using the ISM and analyzing using the relatively more complicated RSM. However, we believe this is because we kept the within family variance quite low, and the ISM assumption may not give such good results when the within family variance is larger. It is also possible that we have only 10 items given to each individual, making it difficult to get a true picture about the ability estimation. We plan to explore this issue further in future.

We plot the estimated ICCs along with the estimated FERFs for the RSM-RSM case for four item models (1, 2, 6, and 9) in Figure 9. The dashed line represents the ICCs, and the dotted line is the estimated FERF for the RSM-RSM case. Note that the estimated curves are very close for Families 1, 6, and 9.
Figure 9: Estimated ICCs and FERFs for four item families for the RSM-ISM case.

Data From an USM, Analyzed Using an RSM

Here we provide the results obtained when a data set generated from the USM is analyzed using the RSM – these results should help us understand how RSM performs when it is not the true model. The coverage of the 95% posterior intervals for the item parameters are 92% for the \( \alpha \)s, 92% for the \( \beta \)s, and 96% for the \( \gamma \)s.

Figure 10 shows a histogram of the sampled values of \( \tau_b \)'s for this analysis. It is clear that the histograms are mostly wider than the earlier situation (data from an RSM and analyzed using an RSM, referred to as the RSM-RSM case later). This is exactly what is to be expected because it was discussed earlier that the USM is a limiting case of the RSM when the \( \tau \)s become very large. We plot the estimated ICCs along with the estimated FERFs and the prediction bounds for four item models (1, 2, 6, and 9) in Figure 11. Each FERF lies in the middle of the ICCs for the same family, but the ICCs are much more widely spread than in Figure 7.

Histograms of the sampled values for the same five \( \theta \)s appear very similar to the
Figure 10: Histograms of the sampled values of $\tau_{bh}$s for all the item models (ordered across rows) for the USM-RSM case.

earlier analyses.

Data From an ISM, Analyzed Using an RSM

We analyze a data set generated from the ISM using the RSM here to study the performance of the latter when it is not the true model. The coverage of the 95% posterior intervals for the item parameters were 97% for the $\alpha$s, 90% for the $\beta$s, and 91% for the $\gamma$s. Interestingly, of the 22 posterior intervals that don’t contain the true values, 20 belong to Item Model 6, proving again that the RSM does not fit the observations for Item Model 6.

Figure 12 shows a histogram of the sampled values of $\tau_{bh}$s for this analysis. It is
Figure 11: Estimated ICCs and FERFs for four item families for the USM-RSM case.

It is obvious that the histograms are mostly thinner than the RSM-RSM case. This is exactly what is to be expected because it was discussed earlier that the ISM is a special case of the RSM when the $\tau$s become very small.

We plot the estimated ICCs along with the estimated FERFs and the prediction bounds for four item models (1, 2, 6, and 9) in Figure 13. We see that the estimated ICCs are very close together for all families, which is expected because we are dealing with the exact same items answered by different individuals in each family. The estimated FERFs are in the middle of the estimated ICCs and the prediction bands are thin. Histograms of the sampled values for the same five $\theta$s that we looked at earlier appear very similar to the earlier analysis.

The results from the USM-RSM case and ISM-RSM case show that the RSM does quite well even when the data does not have any hierarchical structure.
Figure 12: Histograms of the sampled values of $\tau_{jk}$s for all the item models (ordered across rows) for the ISM-RSM case.

5. Example 1: Analysis of a High Stakes Assessment Data Set

The Data Set

This study analyzes operational data from a high stakes assessment (Williamson, Johnson, Sinharay, & Bejar, 2002a) consisting of a number of complex constructed response tasks, each of which are scored on a 3-point polytomous scale. Each administration of the assessment consists of six tasks; one from each of six distinct task domains. Within each domain is a family of tasks constructed to be isomorphic equivalent tasks (i.e., demand performance of the same domain tasks, use identical features in scoring, have highly similar
**Figure 13:** Estimated ICCs and FERFs for four item families for the ISM-RSM case.

statistical performance, and measure the same knowledge and skills, but by virtue of substantial changes to surface features, appear to be substantially different items). For any particular administration, the task is drawn at random from the task family for the domain in question. That is, there are six pools of isomorphic tasks, and for any given examinee, one task is drawn at random from each of the six pools to construct the examinee’s assessment. A breakdown of the sample size by isomorphic task pool is provided in Table 1.

**RSM Analysis**

The current version of the software is designed for dichotomous cases only, with the extension to polytomous cases under current development. Therefore, the polytomous data are transformed into dichotomous data by collapsing the two lowest score categories into a single response category. Because the data consists of constructed response tasks (where the chance of getting an item correct by guessing is almost zero), we assume a 2PL model
as the basic item response model, i.e., we assume,

\[ P_j(\theta_i|a_j, \beta_j) = \frac{1}{1 + \exp\{a_j(\beta_j - \theta_i)\}}. \]

After the transformations on \( \theta_i \) and \( a_j \) as in (1), the prior distribution on the item parameters are taken as

\[ (\alpha_j, \beta_j)^t | \lambda_{\mathcal{I}(j)}, T_{\mathcal{I}(j)} \sim \mathcal{N}_2(\lambda_{\mathcal{I}(j)}, T_{\mathcal{I}(j)}). \]

The prior distributions assumed for \( \lambda_{\mathcal{I}(j)} \)'s and \( T_{\mathcal{I}(j)} \)'s are given by the relevant components of (2) and (3). We assume \( W_1 = 3 \), implying that a priori, we have information equivalent to that of three items in each item family. The number may seem high, keeping in mind that we have only four to six items per family in the data set. However, three is the least value to make sure that the prior density is finite everywhere. We take \( W_2 \) to be a diagonal matrix with elements \( \frac{10}{3} \). The prior mean of the precision matrix \( T_{\mathcal{I}(j)} \) implied by these assumptions is \( 10 \times I_2 \).

The MCMC estimation procedure is conducted through five chains of 20,000 iterations each. Looking at the time-series plots and autocorrelation function plots of a carefully chosen set of parameters (including all of the variance parameters, some mean parameters, and few item parameters and ability parameters), we find out that the MCMC algorithm converges within 4,000 iterations. Hence, the first 4,000 iterations in each chain are treated as burn-in and therefore not included in the determination of the posterior

<table>
<thead>
<tr>
<th>Task set</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>572</td>
</tr>
<tr>
<td>B2</td>
<td>575</td>
</tr>
<tr>
<td>B3</td>
<td>571</td>
</tr>
<tr>
<td>B4</td>
<td>572</td>
</tr>
<tr>
<td>B5</td>
<td>518</td>
</tr>
<tr>
<td>B6</td>
<td>511</td>
</tr>
</tbody>
</table>
Figure 14: Histograms of the sampled values of $\lambda_{b_k}$s for the high stakes assessment data.

distributions of the parameters. The remaining 16,000 iterations in each chain are thinned by selecting every 10th iteration for inclusion in the final data set, determining the posterior distribution of the parameters. Finally, the five chains are pooled together, resulting in a final sample of 8,000 draws from the posterior distribution of each parameter. The estimated ICCs are produced using the median value of the distribution for each parameter.

Figure 14 shows a plot of the histograms of the sampled values of the mean difficulty, i.e., $\lambda_{b_k}$s for all the item families. The plot shows significant between family variation in average difficulty of items.

Figure 15 shows a plot of the histograms of the sampled values of the variance of the difficulty parameter, i.e., $\tau_{b_k}$s for all the item families. The plot shows that the
Figure 15: Histograms of the sampled values of $\tau_{b_k}$s for the high stakes assessment data.

estimated variance looks very similar across families and that it is not very high, with the median around 0.2 for most of the families.

The estimated ICCs and FERFs (along with an approximate 95% prediction interval) for the six isomorphic families are provided in Figure 16.

The greatest degree of variation in the estimated task ICCs from the estimated FERF is in Family B1 while the least variation is observed in Family B6. The ICCs for Family B4 are similar with one notable exception. In general, the families of isomorphic tasks showed considerable similarity in the item response functions for their respective members as well as for the family response function for the isomorphic set.

These results suggest that efforts to construct complex constructed response tasks
Figure 16: The estimated ICCs and FERFs from RSM analysis for all the item families for the high stakes assessment example.

that are isomorphic equivalent tasks can range somewhat in their degree of success, with some being consistently equivalent (e.g. Family B6), some being more variable (e.g. Family B1), and others being largely consistent but with notable deviations (e.g. Family B4).

The variation computed from these isomorphic items is similar to those obtained from a study (Rizavi, Way, Davey, & Herbert, 2002) in which the same subset of items from verbal and quantitative sections of a high stakes admissions test were recalibrated through eight administrations and the variation in item parameters evaluated. If variations in ICCs for isomorphic constructed response tasks are consistently similar to variations obtained
from recalibration of an identical multiple choice item, then the goal of creating isomorphic
crafted response tasks with highly similar statistical performance has been largely met.

**USM Analysis**

It will be of considerable interest to see what would happen if we ignore the
relationship between the items (tasks) within a family. If we have enough data, this is the
gold standard for item calibration. Hence, we fit the usual 2-parameter logistic model (the
USM according to our earlier definition) to the data set using the MCMC algorithm. The
width of the 95% confidence intervals for the item parameters are, on an average, 1.5 times
as much as that for the RSM analysis. Figure 17 shows the estimated ICCs (in dashed lines)
and the estimated FERFs with the two prediction bounds (in solid bold lines) obtained
earlier using the RSM analysis. The estimated ICCs in this plot differ considerably from
those in Figure 16. In most families, one or two items look different from the others in
the same family. However, all of the estimated ICCs fall within the 95% prediction band
obtained from the RSM analysis.

**ISM Analysis**

Finally, we fitted an ISM to the data set. The width of the 95% confidence intervals
for the item parameters are, on an average, 0.60 times as much as that for the RSM
analysis. The estimated ICCs (in dashed lines) and the estimated FERFs (in solid bold
lines) obtained earlier using the RSM analysis are provided in Figure 18. There are only
slight differences between the estimated ICCs and the FERFs for each family.

**Estimation of the Ability Parameters**

Table 2 shows the posterior means and standard deviations of the ability parameters
obtained by the three different models for seven individuals whose number-correct scores
range from 0 to 6.
Figure 17: Estimated ICCs and FERFs from the USM analysis for all the item families for the high stakes assessment data.

Looking at the table, we hardly find any differences in the results obtained by the three models. Like our simulated data example, this implies that the RSM does an equally good job of scoring as the USM, the gold standard. However, part of the reason may be that the number of items for each examinee is so few that the prior distribution on the abilities has a big effect on the posterior distribution, and hence there is no difference among the results obtained by the different methods.
Figure 18: Estimated ICCs and FERFs from the ISM analysis for all the item families for the high stakes assessment data.

6. Example 2: Analysis of Math Online Data

The Data Set

This study analyzes data from the National Assessment of Educational Progress (NAEP) Math OnLine (MOL) special study (Sandene, Bennett, Braswell, & Oranje, 2003). The sample consists of 3,793 examinees in Grade 8, distributed among four test forms. Each of the four forms had a block of common items (denoted MP) and an additional 26 mathematics items (denoted M2-M5 for the four forms), consisting of 16 multiple choice and 10 open-ended items. The number of items of each type appearing in the four forms
Table 2: *Estimated Abilities and Standard Errors for Seven Individuals Obtained by the RSM, USM, and ISM for the High Stakes Assessment Data*

<table>
<thead>
<tr>
<th>Number-correct score</th>
<th>RSM Mean</th>
<th>RSM SD</th>
<th>USM Mean</th>
<th>USM SD</th>
<th>ISM Mean</th>
<th>ISM SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.71</td>
<td>0.67</td>
<td>-1.66</td>
<td>0.66</td>
<td>-1.67</td>
<td>0.65</td>
</tr>
<tr>
<td>1</td>
<td>-1.34</td>
<td>0.62</td>
<td>-1.31</td>
<td>0.60</td>
<td>-1.25</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>-0.71</td>
<td>0.63</td>
<td>-0.66</td>
<td>0.68</td>
<td>-0.72</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>-0.42</td>
<td>0.64</td>
<td>-0.39</td>
<td>0.64</td>
<td>-0.35</td>
<td>0.60</td>
</tr>
<tr>
<td>4</td>
<td>0.22</td>
<td>0.61</td>
<td>0.37</td>
<td>0.63</td>
<td>0.14</td>
<td>0.63</td>
</tr>
<tr>
<td>5</td>
<td>0.54</td>
<td>0.65</td>
<td>0.48</td>
<td>0.64</td>
<td>0.48</td>
<td>0.66</td>
</tr>
<tr>
<td>6</td>
<td>0.98</td>
<td>0.73</td>
<td>0.94</td>
<td>0.75</td>
<td>1.14</td>
<td>0.72</td>
</tr>
</tbody>
</table>

M2-M5 are presented in Table 3, as are the sample sizes from administration.

Table 3: *Item Type and Generation by Form for the MOL Data*

<table>
<thead>
<tr>
<th>Form</th>
<th>Graded response</th>
<th>Multiple choice</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human</td>
<td>AIG</td>
<td>Human</td>
</tr>
<tr>
<td>M2</td>
<td>10</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>10</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>M5</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

The 26 mathematics items comprising Form M2 were written by human item writers and were assembled to be representative of the item pool, to the extent possible. This Form was administered as a paper and pencil assessment, with one subset of items as a calculator-active block, with calculators provided for the students.

Form M3 is identical to Form M2 and uses the same 26 items. However, this form was administered as a linear computerized assessment with an online calculator provided for the calculator-active block of items.
Form M4 was constructed to be parallel to Form M2. Of the 26 items, 11 were identical to the items appearing on Form M2 while 15 items were automatically generated items (Singley & Bennett, 2002) different from, but intended to be parallel to, the corresponding items on Form M2. Like Form M2, Form M4 was administered via paper and pencil with a calculator provided for the calculator-active block.

Form M5 was constructed to be parallel to Form M2. Of the 26 items, 11 were identical to the items appearing on Form M2, while 15 items were automatically generated items (Singley & Bennett, 2002) different from, but intended to be parallel to, the corresponding items on Form M2. The generated items for Form M5, however, are different items than the generated items appearing on Form M4. For each automatically generated item on Form M4, a corresponding item is generated from the same item model on Form M5. Like Form M2, Form M5 was administered via paper and pencil with a calculator provided for the calculator-active block of items.

For this analysis, the MP block was not considered and only the 16 dichotomously scored (multiple choice) items of the other 26 items in each form were analyzed. In addition, no overlapping students exist in this design; that is, no one took more than one of the forms.

**RSM Analysis**

The RSM, defined by (1), is fit to the data. We use prior distributions for the item family parameters that are given by (2) and (3). Making these prior distributions noninformative results in problems with the convergence of the MCMC (we believe this to be an outcome of only four items being in each item family), and hence we make (2) and (3) informative by taking \( \sigma^2_{\lambda_o} = 1 \), \( W_1 = 5 \), and \( W_2 = A \) diagonal matrix with 2 as the diagonals. This implies that a priori, we have information equivalent to that of five items in each item family. The prior mean of the precision matrix \( T^{-1}_{i(j)} \) implied by this assumption is \( 10 \times I_3 \).

The MCMC estimation procedure is conducted through five chains of 20,000 iterations each. Looking at the time-series plots and autocorrelation function plots of a
carefully chosen set of parameters (including all of the variance parameters, some mean parameters, and few item parameters and ability parameters), we find out that the MCMC algorithm converges within 2,000 iterations. Hence, the first 2,000 iterations in each chain are treated as burn-in and therefore not included in the determination of the posterior distributions of the parameters. The remaining 18,000 iterations in each chain are thinned by selecting every 9th iteration for inclusion in the final data set determining the posterior distribution of the parameters. Finally, the five chains are pooled together, resulting in a final sample of 10,000 draws from the posterior distribution of each parameter.

The estimated ICCs and FERFs (along with an approximate 95% prediction interval) are provided by family in Figure 19. Note that the Families 1, 2, 5, 10, and 13 have no AIG items. The ICCs for items appearing in different forms (M2-M5) are plotted as different type of lines (as mentioned in the caption of the figure). Those item families without AIG items generally have more closely corresponding ICCs than families with AIG items; the most similar set of ICCs are represented in Family 1. This is, of course, not surprising considering the fact that families without AIG items are presenting a series of ICCs all on the same item appearing in different forms. Despite the generally close ICCs for item families without AIG items, some variation is evident in some of the ICCs for these families from the close match of Family 1 to greater variability evident in Family 5, as well as some deviation in the guessing parameter for one item in Family 10.

Examination of the families that contain AIG items reveals a couple of immediately obvious deviations. Most obvious is the fact that the entire family of items for Family 6 is flat at approximately random chance for all levels of ability. However, this is true for both the human generated item (appearing in Forms M2 and M3) and the AIG items (appearing in Forms M4 and M5) and the ICCs are consistent with the classical statistics calculated on the items. Hence, this phenomenon is the result of a characteristic of the item type or content rather than the result of anything inherent in AIG.

Another obvious deviation in ICCs occurs in Family 3. In this instance, the manually generated item and the AIG item appearing in M5 have very similar ICCs, while the AIG item appearing in M4 deviates dramatically from the other items in the family. The extent of the deviation also appears to impact the response function for the family as
Figure 19: Estimated ICCs and FERFs from RSM analysis for all the item families for the MOL data. The bold solid lines represent estimated FERFs and 95% prediction intervals on them. The thin solid lines represent estimated ICCs for items in Form M2; dotted lines for items in Form M3; lines with small dashes for items in Form M4; and lines with longer dashes for items in Form M5. Items in Forms M4 and M5 are AIG items, if at all.

The four ICCs in Family 8 are closely related, however there is an obvious difference in the guessing parameter between the ICCs and the FERF. The difference appears to be an artifact of the prior selected for the family parameters $\lambda_{gn}$s, the mean of the guessing parameters. Families 4, 9, 11, and 15 all have minor deviations between the ICCs for items...
within those families.

A closer look at Figure 19 suggests that a number of the families with AIG items appear to have ICCs that are quite similar for both the human generated item and the AIG items. These include Families 5, 7, 12, 14, and 16. Still others, including Families 12 and 16, have ICCs for the AIG items that are as close or closer to the ICC for one administration of the human generated item than the ICC for the other administration of the same human generated item.

**USM Analysis**

We then fit the USM to the same data set. The estimated ICCs (in dashed lines) and the estimated FERFs (in solid bold lines) obtained earlier (using the RSM analysis) are provided in Figure 20. The ICCs are plotted using line types similar to those in Figure 19. The ICCs in this plot are hardly different from those in Figure 19 except for Item Family 6. This shows that the RSM does a pretty good job of item calibration for this situation and may save the test administrators a lot of resources because one does not gain much at all by calibrating each item separately.

**ISM Analysis**

Finally, we fit the ISM to the same data set. The estimated ICC for the item family (remember that for this analysis, one assumes that the items generated from each item family are identical, and hence we have only one ICC for each family) and the estimated FERF obtained earlier (using the RSM analysis) are provided in Figure 21 in dashed lines and solid bold lines respectively.

Interestingly, Figures 19, 20, and 21 suggest that for each family except for Family 6, the ICC obtained using the ISM analysis follows the individual ICCs (obtained by the RSM, or, even the USM) more closely than does the estimated FERF from RSM. This is probably an outcome of the fact that we have too few items for each item family, making it difficult for the RSM to fit well. Also, even for families with AIG items, two of the four items are actually the same item repeated, making the ISM a good candidate to fit the
Figure 20: Estimated ICCs and FERFs from the USM analysis for all the item families for the MOL data. The bold solid lines represent estimated FERFs and 95% prediction intervals on them. The thin solid lines represent estimated ICCs for items in Form M2; dotted lines for items in Form M3; lines with small dashes for items in Form M4; and lines with longer dashes for items in Form M5. Items in Forms M4 and M5 are AIG items, if at all.

data. A tighter prior distribution for the variance parameters (or the assumption that the variance matrices for the item parameters are same over the families) might result in a better fit of the RSM.
Figure 21: Estimated ICCs from the ISM analysis for all the item families for the MOL data. The dashed lines represent the estimated ICCs for each family. The estimated FERFs from the RSM analysis are also provided for comparison in solid bold lines.

Estimation of the Ability Parameters

Just like in the previous example, we find hardly any differences in the estimates of abilities obtained by the three models and don’t include the results here. Again, this shows that the RSM does an equally good job of scoring as does the USM, the gold standard. We need to explore this issue further to find out when and how the ability estimates obtained under the three model assumptions differ.
7. Conclusions and Future Work

Our work shows that when a test consists of item families, the RSM can take into account the dependency 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. Hence, this work is an important step in showing that it may be enough to calibrate the item family 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.

Our first priority is to develop diagnostic tools to help researchers decide which of the three models discussed here (RSM, USM, and ISM) would be preferable to analyze a given data set. For example, an RSM may not be a good fit to data sets with a few number of items per family. A part of our research will be to figure out how the ability estimates differ under the different model assumptions. Our next priority is the extension of the hierarchical model to polytomous data to make it more widely applicable. To model the responses of the examinees to polytomous items, we will use the generalized partial credit model (GPCM) (Muraki, 1992), which assumes that the response of student $i$ to item $j$ has a multinomial distribution with probabilities

$$
P_{ijk}(\theta_i|a_j, \beta_j, \delta_{j1}, \delta_{j2}, \ldots, \delta_{jK}) = P(X_{ij} = k|\theta_i, a_j, \beta_j, \delta_{j1}, \delta_{j2}, \ldots, \delta_{jK})$$

$$= \frac{\exp\{k a_j (\theta_i - \beta_j) - a_j \sum_{i=1}^{k} \delta_{ji}\}}{\sum_{k=1}^{K} \exp\{k a_j (\theta_i - \beta_j) - a_j \sum_{i=1}^{k} \delta_{ji}\}}$$

for $k = 1, 2, \ldots, K$, where $a_j$ is the slope parameter for the item (constant over the different categories), $\beta_j$ is the overall difficulty for the item, and $\delta_{jk}$ is the difficulty parameter for the $k$-th category, $k = 1, 2, \ldots, K$. In the next level of the model, we assume that $a_j \equiv \log\{a_j\}$ and that

$$\eta_j \equiv (\alpha_j, \beta_j, \delta_{j1}, \delta_{j2}, \ldots, \delta_{jK})^t \mid \lambda_{I(\cdot)}, T_{I(\cdot)} \sim \mathcal{N}_{K+2}(\lambda_{I(\cdot)}, T_{I(\cdot)}).$$

The prior distributions assumed on $\lambda_k$s and $T_k$s are similar to (2) and (3). The posterior distributions of the parameters are also similar to those discussed in the same section.
Specifically, the posterior distributions of $\lambda_k$s and $T_k$s will be very similar to (6) and (7).

We would also like to find out 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 to the prior distributions on the hyper-parameters $\tau$s when only a few items belong to each item family. For example, results are sensitive to the prior distributions for the MOL data set where each item family consists of only four items.

Finally, we would like to expand the hierarchical model so that it can take into account covariates, either task feature variables or demographic variables.
References


