Calibrating Item Families and Summarizing the Results Using Family Expected Response Functions
Sandip Sinharay, Matthew S. Johnson and David M. Williamson

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2003 28: 295
DOI: 10.3102/1076998602800295

The online version of this article can be found at:
http://jeb.sagepub.com/content/28/4/295

Published on behalf of

American Educational Research Association
and

http://www.sagepublications.com

Additional services and information for Journal of Educational and Behavioral Statistics can be found at:

Email Alerts: http://jebs.aera.net/alerts
Subscriptions: http://jebs.aera.net/subscriptions
Reprints: http://www.aera.net/reprints
Permissions: http://www.aera.net/permissions
Citations: http://jeb.sagepub.com/content/28/4/295.refs.html

Version of Record - Jan 1, 2003
What is This?
Item families, which are groups of related items, are becoming increasingly popular in complex educational assessments. For example, in automatic item generation (AIG) systems, a test may consist of multiple items generated from each of 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 with multiple-choice items. We fit the model using the Markov Chain Monte Carlo (MCMC) algorithm, introduce 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. This work is thus a step towards creating a tool that can save significant amount of resources in educational testing, by allowing proper analysis and summarization of data from tests involving item families.

Keywords: automatic item generation, Bayesian methods, family expected response function, item model, Markov Chain Monte Carlo

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 & Kyllonen, 2002, for a survey of current areas of research in 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 constitute an item family as they are related to one another through the common generating model. Therefore, operational use of item models in real tests will require statistical models that can account for the dependence structure among the items from the same item family. Glas and van der Linden (2001) suggest one such model that is more general. Our work is aimed at examining the latter model, which assumes the item parameters of a three-parameter logistic (3PL) model (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 the former 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 examinee’s responses to the same single member from an item family. Glas and van der Linden’s model is also different from the multilevel item response theory models in Janssen, Tuerlinckx, Meulders, and De Boeck (2000) and Wright (2002).

This work shows how to fit the hierarchical model 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 prediction limit for members of the item family using Monte Carlo integration and the output from the MCMC algorithm. Finally, two real data examples demonstrate the application of the suggested ideas. Results obtained using a hierarchical model are compared to those obtained by other existing methods for tackling similar problems.

The next section reviews three models used for the analysis of item families; the last one of them is the hierarchical model. Section 3 discusses the Bayesian estimation procedure for the model parameters and the family expected response function. Section 4 applies the hierarchical model to an operational data set from a high-stakes assessment. Section 5 discusses the application of the model to a data set from the National Assessment of Educational Progress (NAEP). Section 6 suggests future directions of research.
2. Existing Models for Analyzing Data Involving Item Families

Suppose a test consists of $J$ dichotomously scored items indexed by $j = 1, 2, \ldots, J$. Each of the $J$ items are classified into one of $K$ item families. An item family is defined as a group of items that are related to one another in some way. For example, the items may be generated (either manually or automatically) from the same item model. Let $I(j)$ be the item family from which Item $j$ is generated. Items $j$ and $k$ are siblings if they belong to the same item family, i.e., if $I(j) = I(k)$. Each of the $N$ examinees (who are indexed by $i = 1, 2, \ldots, N$) respond to a single item from each of the $K$ item families. Let $P_j(\theta_i)$ denote the probability of a correct response by Examinee $i$ on Item $j$.

There are three approaches for modeling data involving item families. We discuss the models in the context of the 3PL (Lord, 1980) model, but it is possible to think of the models in the context of other types of dichotomous item response functions (e.g., 2PL model) as well.

Unrelated Siblings Model

The “Unrelated Siblings Model” (USM) assumes a separate, unrelated (independent) item response function for all items, regardless of their family membership. Mathematically, the model assumes

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

where $\logit^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$, $a_j$, $\beta_j$, and $c_j$ are the slope, difficulty and guessing parameters respectively of item $j$. While standard software (e.g., PARSCALE, Muraki & Bock, 1991) can fit the USM, the model has the disadvantage that each item has to be individually calibrated; and hence the USM requires large sample sizes for precise calibration. Also, this model ignores the relationship between siblings in an item family, and hence will provide standard errors of item parameters that are too large.

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

$$P_j(\theta_i) = c_{I(j)} + (1 - c_{I(j)}) \logit^{-1}(a_{I(j)}(\theta_i - \beta_{I(j)})).$$

We call the above model the “Identical Siblings Model” (ISM). While standard software can fit this model, ISM 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.

Related Siblings Model

One way to overcome the limitations of the above mentioned models 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). This first component of the model is a 3PL model

\[ P_j(\theta) = c_j + (1 - c_j) \logit^{-1}[a_j(\beta_j - \theta)], \]

The population distribution for the latent abilities is normal with mean \( \mu \) and variance \( \sigma^2 \). After making the transformations \( \alpha_j \equiv \log(a_j) \) and \( \gamma_j \equiv \logit(c_j) \), Glas and van der Linden use a normal distribution to relate the item parameters of items within the same item family as

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

where the family mean vector \( \lambda_{I(j)} \) can be partitioned as \( \lambda_{I(j)} = (\lambda_{\alpha I(j)}, \lambda_{\beta I(j)}, \lambda_{\gamma I(j)})' \), and the diagonal elements of the family variance \( T_{I(j)} \) will be referred to as \( \tau^2_{\alpha I(j)}, \tau^2_{\beta I(j)}, \tau^2_{\gamma I(j)} \) respectively.

To fix the origin and scale of the ability parameters \( (\theta_i) \), the model sets the parameters of their prior distribution at \( \mu = 0, \sigma^2 = 1 \).

By integrating the individual item parameters \( (\alpha_j, \beta_j, \gamma_j)' \), 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 and \( \lambda_{I(j)} \) and \( T_{I(j)} \).

The ISM and USM are limiting cases of the RSM. If the variance parameters in \( T_{I(j)} \) become very small in an RSM, there is no variation among the items in the same family and the RSM becomes almost identical to the ISM. On the other hand, if the variances grow 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 the RSM becomes indistinguishable from the USM.

**Family Expected Response Function (FERF)**

Parameters are difficult to interpret in many models and even more so in higher levels of an hierarchical model. One of the goals of this work was providing an effective graphical summary of the item families. For the RSM we examine a response function that is “typical” for each item family. The family expected response function (FERF) describes the probability that an examinee with ability \( \theta \) correctly responds to a randomly selected item from the item family. The hierarchical structure of the RSM suggests that obtaining the response function requires taking the average of

\[ P_j(\theta|a_j, b_j, c_j) = c_j + (1 - c_j) \logit^{-1}[a_j(\theta - \beta_j)], \]

over all possible values of the item parameters that one may observe for that particular item family. We suggest averaging out the item parameters with respect to
their prior distributions and then averaging out the item family parameters with respect to their posterior distribution, i.e.,

\[ P[\theta | I(j)] = \int_{\lambda_{(j)}, \theta_{(j)}} \int_{\eta_{j}} P[\theta | \eta_{j}] \phi_{3}(\eta_{j} | \lambda_{(j)}, T_{(j)}) d\eta_{j} f(\lambda_{(j)}, T_{(j)} | X) d\lambda_{(j)} dT_{(j)} \]  

(3)

to obtain the FERF for item family \( I(j) \), where \( \eta_{j} = (\alpha_{j}, \beta_{j}, \gamma_{j})' \), \( \phi_{3}(\eta_{j} | \lambda_{(j)}, T_{(j)}) \) is the density function of the multivariate normal prior distribution on \( \eta_{j} \) and \( f(\lambda_{(j)}, T_{(j)} | X) \) is the joint posterior distribution of \( \lambda_{(j)} \) and \( T_{(j)} \) given the response matrix \( X \).

The FERFs provide the basis of a nice graphical summary of the results from the hierarchical model. A plot showing the estimate of the FERF of the family along with the estimates of the item response functions of the items in the same item family will be very informative. One such plot for each item family will not only show how the response functions behave in the families, but will also provide some information about the success of the item generation process used. Substantial variation among the response functions will indicate that the item that the item generation process was not successful enough in creating similar items and needs improvement. Our examples later will demonstrate the usefulness of the response function plots more clearly.

3. Bayesian Estimation for the Related Siblings Model

While the advantage of the hierarchical model 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 prefer Bayesian estimation for the RSM. For one, maximum likelihood estimation for this model is extremely difficult. Secondly, a Bayesian analysis allows examining any posterior summaries of any functions of the model parameters. The family expected response function is one such function, for example.

In this work we use a Markov chain Monte Carlo (MCMC) algorithm to perform Bayesian estimation of the RSM. The goal of the MCMC methods is to create a Markov chain in the parameter space whose distribution converges to the joint posterior distribution of the model parameters. This work uses what Patz & Junker (1999) call the “Metropolis within Gibbs” Markov chain, which can be shown to converge to the joint posterior distribution. The algorithm requires model parameters to be drawn from their conditional posterior distribution, where the distribution is conditional on all other model parameters and the observed data.

Prior Distributions on Family Means and Variances

Bayesian estimation of the RSM requires assuming some prior distributions on the family means \( \lambda_{(j)} \) and family variances \( T_{(j)} \). This work assumes independent multivariate normal prior distributions for \( \lambda_{(j)} \),

\[ \lambda_{(j)} \sim N_{3}(\mu_{\lambda} = (0, 0, \mu_{\lambda}), V_{\lambda} = \text{Diag}(100^2, 100^2, \sigma_{\lambda}^2)) \]  

(4)
and independent inverse Wishart prior distributions on the $T_{i(j)}$s,

$$T_{i(j)}^{-1} \sim \text{Wishart}(W_1, W_2).$$

The noninformative prior distributions on the first two components (those corresponding to the discrimination parameter $\alpha_j$ and difficulty parameter $\beta_j$) of $\lambda_{i(j)}$ in Equation 4 allow the data to provide most of the information about the posterior distribution. However, the prior distribution on the third component (that corresponding to the guessing parameter $\gamma_j$) is taken to be somewhat informative to indicate the information we have about multiple choice items; it is possible to extract some information about the guessing parameter from the test type and format. For example, in a test with multiple choice items with five choices and good distractors for the other four choices, an examinee will have a probability of around 0.2 on average to get any item correct by random guessing; this results in taking

$$\mu_{\lambda_j} = \logit(0.2) = -1.39,$$

and a value of $\sigma^2_{\lambda_j}$ that is not too large. When the number of choices is different, or some prior information about the quality of the distractors is available, the prior distribution may be changed accordingly.

Equation 5 implies an a priori mean of $W_1W_2$ for $T_{i(j)}^{-1}$ and that a priori there is information that is equivalent to $W_1$ observations on $(\alpha_j, \beta_j, \gamma_j)$. The investigator should set these parameters depending on the knowledge of the problem.

**Brief Details of the MCMC Algorithm**

Item parameters $\alpha$, $\beta$, and $\gamma$ and the latent abilities $\theta$ have the same form of posterior distribution as what would have been obtained under a simple 3PL model. These parameters are drawn from their respective conditional distributions using Metropolis steps described in Patz and Junker (1999).

Let $\lambda_k$ and $T_k$ denote the item family mean vector and covariance matrix of the $k$-th item family respectively. Conditional on the item parameters $\alpha$, $\beta$, and $\gamma$, $\lambda_k$ and $T_k$, are independent of $\theta$ and the observed data $X$. The full conditional distributions of the $\lambda_k$s and $T_k$s are trivariate normal and Inverse Wishart respectively (Sinharay, Johnson, & Williamson, 2003; Gelman, Carlin, Stern, & Rubin, 1995, Section 3.6).

Hence, the addition of the hierarchical component in the model amounts to additional sampling from multivariate 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.

**Estimating the Family Expected Response Function**

We use Monte Carlo integration to estimate the FERF defined in Equation 3. We also discuss how to attach a 95% prediction interval with the estimate. The first
step required in the estimation process for the $k$th item family consists of the following two substeps:

1. Generate a sample of size $M$ from $f(\lambda_k, T_k | X)$. This can be obtained, without any additional simulation, as a subsample from the posterior sample generated by the MCMC algorithm used to fit the RSM.

2. For each of the above $M$ draws of $(\lambda_k, T_k)$, generate $m$ values of the item parameter vector $\eta_j$ from the multivariate normal distribution $\phi_3(\eta_j | \lambda_k, T_k)$.

The second step, which uses the sampled item parameters from Step 1, repeats the following substeps for a number of values of $\theta$:

1. For each of the $Mm$ draws of $\eta_j$ obtained in Step 1, compute $P_j(\theta | \eta)$ using Equation 2

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

3. The 2.5th and 97.5th percentiles of the $Mm$ probabilities above form an approximate 95% prediction interval to attach with the estimate obtained above.

This work uses 100 equidistant values of $\theta$ in the interval ($-4, 4$) in Step 2 to estimate the FERF, and uses $M = 1000$, $m = 10$. This description shows that estimation of the FERF is quite straightforward given the output from the MCMC algorithm and takes little additional time.

Sinharay, Johnson, and Williamson (2003) use these methodologies in a simulation study to fit the RSM to data generated from the RSM, USM, and ISM. Their work shows that the RSM explains data generated from the models quite well, estimating the generating parameters accurately. The following two sections discuss results of fitting the RSM to two real data sets, one each from a high-stakes and low-stakes assessment.

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

The Assessment and the Data Set

Operational data from a high stakes assessment (Williamson, Johnson, Sinharay, & Bejar, 2002) consists of a number of complex constructed response items, each of which are scored on a 3-point polytomous scale. Each administration of the assessment consists of six items; one from each of six distinct domains. Each domain consists of a family of four to six items that are constructed to be isomorphically equivalent (i.e., use identical features in scoring, measure the same knowledge and skills, but appear to be substantially different items by virtue of substantial changes to surface features). For any given examinee, a single item is drawn at random from each of the six item families to construct the examinee’s assessment. A total of 579 examinees took the assessment.

RSM Analysis

Although the items are scored on a three-point scale, one of the score categories is an “undecided” category used by the raters when they cannot determine whether
the response is correct or incorrect. In this article we are concerned with dichoto-
mous responses, and therefore we absorb this “undecided” category into the “incor-
rect” score category. Because the test consists of constructed response items
(where the chance of getting an item correct by guessing is almost zero), the 2PL
model should explain the data adequately. The model assumes the probability of
a correct response to be:

\[ P_j(\theta_j | a_j, \beta_j) = \text{logit}^{-1}[a_j(\theta_j - \beta_j)]. \]

After applying the transformation \( \alpha_j = \log(a_j) \), the prior distribution on the item
parameters is assumed to be:

\[ (\alpha_j, \beta_j) ^ T \mathbf{\pi}_{(j)}, \mathbf{T}_{I(j)} \sim N_2(\mathbf{\alpha}_{(j)}, \mathbf{T}_{I(j)}). \]

The prior distributions for \( \mathbf{\alpha}_{(j)} \) and \( \mathbf{T}_{I(j)} \) are given by the relevant components of
Equations 4 and 5. We assume, implying that \( W = 3 \), implying that \textit{a priori}, we
have information equivalent to that of three items in each item family. The num-
ber may seem high, keeping in mind that we have only four to six items per fam-
ily in the data set. However, three is the smallest value that ensures that the
expectation of the prior distribution on \( \mathbf{T}_{I(j)} \) is finite. We take \( W_2 \) to be a diagonal
matrix with elements \( \frac{10}{3} \). The prior mean of the precision matrix \( \mathbf{T}_{I(j)}^{-1} \) implied by
these assumptions is \( 10 \times I_3 \).

The MCMC estimation procedure is conducted through five chains of 20,000
iterations each. Time-series plots and Gelman-Rubin convergence diagnostics
(Gelman & Rubin, 1992) suggest that the MCMC algorithm converges. The first
4,000 iterations in each chain are treated as burn-in and the remaining 16,000 iter-
ations in each chain are thinned by selecting every 10th iteration. Finally, the five
chains are pooled together, resulting in a final sample of size 8,000 from the pos-
terior distribution of each parameter. The estimates of the family expected
response functions are obtained as described in Section 3; the estimated item
response functions (IRF) are produced using the posterior median for each item
parameter.

Figure 1 summarizes the sampled values of the mean difficulty (\( \lambda_{h_{k}} \)) and the
within family standard deviation (\( \tau_{h_{k}} \)) for all item families. The horizontal axis of the plot is for the mean and the vertical axis for the variance. For each fam-
ily, a point (denoted by the family name) shows the posterior median of \( \lambda_{h_{k}} \) ver-
sus the posterior median of \( \tau_{h_{k}} \). A horizontal line around the point denotes an
approximate 95% equal-tailed interval for \( \lambda_{h_{k}} \); a vertical line around the point
denotes an approximate 95% equal-tailed interval for \( \tau_{h_{k}} \). The item families all
tend to be relatively easy, signified by the negative sign on five out of the six
families.

The between family standard deviation of the mean family difficulties is 0.62,
which is a moderately low variation.
The posterior median of the standard deviation ($\tau_{\beta_{k}}$) in each of the six families is less than the between-family standard deviation of 0.62. However, each of the 95% posterior credible intervals contains the between-family standard deviation of 0.62, which seems to indicate that there is not enough statistical evidence to suggest that the difficulty of items within each family is less variable than that for items across families. This lack of evidence is probably due to the fact that the number of examinees taking each item ranges between 78 to 133, quite small numbers.

Figure 2 contains the estimated item response function (in dashed lines) and family-expected response functions along with approximate 95% prediction intervals (in solid lines) for the item families. In general, the estimated item response functions appear to concentrate around the corresponding estimated family-expected response function for each item family, clearly demonstrating the interdependence of the siblings. Family B1 has the greatest degree of variation among the estimated item response functions, while Family B6 shows the least variation. The item response functions for Family B4 are similar with one notable exception.

These results suggest that the efforts put forth in this assessment to create item families with similar statistical properties vary somewhat in the degree of success.
Some families contain items that are highly concentrated around the family expected response function (e.g., Family B6); other families have items that are more variable (e.g., Family B1). Family B4 demonstrates another type of family that one might encounter when attempting to create isomorphic families; most of the items are concentrated near the family expected response function, but there is a single ill-behaved item that is relatively different from its siblings.

The point estimates of the within family variance found in this analysis are similar to the between-administration variance reported by Rizavi, Way, Davey, & Herbert (2002) for a high-stakes admissions test where the same subset of items were re-estimated over eight administrations of the test. If variations in the item response functions for items from the same family is similar to variations obtained from recalibration of an identical item, the goal of creating isomorphic items with similar statistical performance has been achieved.

**USM Analysis**

When the number of examinees that respond to each item becomes large, the RSM should become indistinguishable from a USM, which ignores the relationship among siblings. This is because the effect of the prior distributions on the item parameters (in RSM) becomes negligible with a large number of examinees tak-
Calibrating Item Families and Summarizing the Results

ing each item. In fact, if the number of examinees becomes very large, it may be more efficient to simply use the USM for analyzing the data, assuming that the interest is only in calibrating the items once, as is with this example. If the interest were in using the calibrated item models for automatic item generation in the future, RSM would be the more appropriate model. However, this data set does not have a very large number of examinees and therefore the results from a model (RSM) that relates items within the same family may not be that similar to the results found by ignoring the relationship (as in USM). Therefore, we examine the results of the USM analysis of this data set.

We fit the unrelated siblings version of the 2-parameter logistic model to the data set using the MCMC algorithm. The prior distributions on the item parameters are

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

Because we are ignoring valuable information in the data set we find that the width of the 95% equal-tailed credible intervals for the item parameters under the USM are, on average, 1.5 times as much as that for those found with the RSM analysis. Figure 3 contains the estimated item response functions (in dashed lines) from the USM analysis. For comparison purposes, the figure also shows the estimated

![Figure 3](http://example.com/figure3.png)

**FIGURE 3.** Estimated item response functions (dashed lines) from the unrelated siblings model analysis for the high-stakes assessment example.
family expected response functions along with the 95% prediction bounds obtained using the RSM analysis (solid lines). The estimated item response functions in this plot differ considerably from those in Figure 2. In most families, one or two items look different from the others in the same family. However, all of the estimated IRFs 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% equal-tailed credible intervals for the item parameters are, on an average, 0.60 times as much as that for the RSM analysis. This is because application of ISM implies assumption of larger number of examinees (than with RSM) taking the items. There are very slight differences between the estimated IRFs and FERFs (obtained earlier from the RSM analysis).

**Estimation of the Ability Parameters**

Figure 4 (above) contains the approximated posterior means and the corresponding 95% equal-tailed credible intervals for the ability parameters obtained from application of the three models for seven individuals whose number-correct scores range from 0 to 6.

The figure shows that there is hardly any difference in the results produced by the three models, implying 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 standard normal 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.

**5. Example 2: Analysis of Math Online Data**

**The Assessment and the Data Set**

This study analyzes data from the National Assessment of Educational Progress (NAEP) Math OnLine (MOL) special study (Sandene, Bennett, Braswell, & Oranje, in press). The sample considered consists of 3,793 examinees in Grade 8,
distributed among four test forms. Each of the four forms has a block of common items (denoted MP) and an additional 26 items (denoted M2, M3, M4, and M5 in the four forms), consisting of 16 multiple-choice and 10 constructed-response items. The number of items of each type appearing in the four forms M2-M5 are presented in Table 1, as are the number of examinees taking them. There were no overlapping students in this design; that is, no one took more than one of the forms.

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 and M5 were constructed to be parallel to Form M2. Of the 26 items in each of these forms, 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 automatically generated items on Form M5 are different from those 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, Forms M4 and M5 were administered via paper and pencil with a calculator provided for the calculator-active block.

For this analysis, the MP block is not considered and only the 16 multiple choice items out of the 26 items in the other forms are analyzed. The five dichotomous item families that have no automatically generated items are families 1, 2, 5, 10, and 13.

**RSM Analysis**

The RSM, defined in Section 2, is fit to the data. We use prior distributions for the item family parameters that are given by Equations 4 and 5. Making these prior distributions noninformative results in problems with the convergence of the MCMC (probably an outcome of each item family consisting of only four items), and hence we make the prior distributions informative by taking $\sigma^2_{\lambda} = 1, W_1 = 5$, 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 $T_{n0}^{-1}$ 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 Gelman-Rubin convergence diagnostics of the parameters, we ensure that the MCMC algorithm converges. The first 2,000 iterations in each chain are treated as burn-in and the remaining 18,000 iterations in each chain are thinned by selecting every 9th iteration. Finally, the five chains are pooled together, resulting in a final posterior sample of size 10,000.

Figure 5 shows the estimated item and family expected response functions along with approximate 95% prediction intervals for the families. The form (M2-M5) on which each item appeared is shown using different line types (as mentioned in the caption of the figure). Item families without automatically generated items (1, 2, 5, 10, and 13) generally have item response functions that are more concentrated around the family expected response functions than those families containing automatically generated items. In fact the family with the most similar set of IRFs is Family 1, which is a family without any automatically generated items.
items. This is not surprising, considering the fact that families without automatically generated items are presenting a series of IRFs all on the same item appearing in different forms.

Despite the relative similarity of the item parameter and item response function estimates of the items from families without automatically generated items, some variation is evident. Family 5 exhibits higher within-family variation than does Family 1. In Family 10, there is one item whose guessing parameter noticeably deviates from that of the others and from the mean guessing parameter of the family.

Examination of the families containing automatically generated items reveals a number of obvious deviations. Most noticeable is the fact that all the items in Family 6 have almost horizontal response functions, which indicates that all students have a random chance of getting the item correct, regardless of their ability level. However, this is true for both the human generated items (appearing in Forms M2 and M3) and the automatically generated items (appearing in Forms M4 and M5). Hence, this phenomenon is the result of a characteristic of the item type or content rather than the result of anything inherent in automatic item generation systems. This family was later removed from the operational analysis of this data set (Sandene, Bennett, Braswell, & Oranje, in press).

Another obvious variation in item response functions occurs in Family 3. The manually generated item in forms M2 and M3 perform similar to the automatically generated item on M5, while the automatically generated item appearing on form 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 a whole.

The four items in Family 8 are closely related; however there is an obvious difference in the guessing parameter of the items and the mean guessing parameter \( \lambda_k \) for the family. The difference is an outcome of the strong effect of the prior distribution selected for the mean guessing parameters (which results from each item family consisting of a few items). Family 7 and 12 shows similar characteristics.

Families 4, 9, 11, and 15 all have minor deviations between the IRFs for items within those families.

Figure 5 also suggests that a number of families with automatically generated items appear to have items that behave similarly for both the human generated item and the automatically generated items. These include Families 7 and 8. Some others, like Family 12, have IRFs for the automatically generated items that are closer to the IRF for one administration of the human generated item than the IRF for the other administration of the same human generated item.

The 95% prediction bounds are very wide for all the item families, providing virtually no information about where the next item from an item family might lie. This is probably due to the fact that the basic model for this example is a 3PL model and there are only four items in each item family, which provide little information about the variability of the response function for the family.
USM Analysis

We also analyzed the Math Online data by ignoring the family structure inherent in this data. The estimated item response functions from the USM analysis and the estimated family expected response functions obtained from the RSM analysis are provided in Figure 6.

The different item forms are denoted using the same line types as in Figure 5. The item response functions found under this USM analysis are almost indistinguishable from the estimates obtained using the RSM (see figure 5). The one exception to this statement is item Family 6. The fact that the USM and RSM analyses provide very similar estimates at the item level seems to suggest that the number of examinees in this data set is large enough to overcome the influence of the prior distribution on the item parameters. However, the RSM does still have advantages. Because the RSM calibrates the item model, it provides the only way if the interest were in using the calibrated item models for automatic item generation in the future.

ISM Analysis

Finally, we analyzed the Math Online data using the ISM, which assumes that each item within a family is identical. The single estimated response function for each item family found under the ISM and the estimated family expected response

FIGURE 6. Estimated family expected response functions from the RSM analysis (solid bold lines) and estimated item response functions from the USM analysis for the Math Online data.
function obtained under the RSM analysis appear in Figure 7 in dashed lines and solid bold lines respectively. Interestingly, Figures 5, 6, and 7 suggest that for each family except Family 6, the response function obtained using the ISM analysis follows the individual item response functions (obtained by the RSM, or, even the USM) more closely than does the estimated family expected response function obtained from RSM. This is probably an outcome of the fact that each item family consists of too few items, making it difficult for the RSM to explain the data adequately. Also, even for families with automatically generated items, two of the four items are actually the same item repeated, making the ISM a good candidate to fit the 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.

Estimation of the Ability Parameters

Just like in the previous example, estimates of abilities from the three models hardly differ (results not included here), showing that the RSM does an equally good job of scoring as does the USM. We need to explore this issue further to find out under what conditions do the ability estimates obtained under the three model assumptions differ.

FIGURE 7. Estimated item response functions (in dashed lines) from the ISM analysis for the Math Online data. The estimated family expected response functions from the RSM analysis (in solid bold lines) are also provided for comparison.
6. 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 take into account constructed response items. Johnson and Sinharay (2003) describe initial findings regarding one such extension.

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 item family variances 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

Calibrating Item Families and Summarizing the Results


Authors

SANDIP SINHARAY is Associate Research Scientist, Educational Testing Service, Rosedale Road, Princeton, NJ 08541; ssinharay@ets.org. His areas of specialization are Bayesian statistics, model checking and model selection, educational statistics, and statistical computing.

MATTHEW S. JOHNSON is Assistant Professor, Department of Statistics and Computer Information Systems, Zicklin School of Business at Baruch College, City University of New York, One Bernard Baruch Way, New York, NY 10010; Matthew_Johnson@baruch.cuny.edu. His areas of specialization are educational and behavioral statistics, categorical data analysis, and applied Bayesian statistics.

DAVID M. WILLIAMSON is Research Scientist, Educational Testing Service, Rosedale Road, Princeton, NJ 08541; dmwilliamson@ets.org. His areas of specialization are psychometrics, automated scoring of computerized simulations, and evidence centered design.

Manuscript received February 2003
Revision received May 2003
Accepted June 2003