Research Report

Analysis of Data From an Admissions Test With Item Models

Sandip Sinharay

Matthew Johnson

Research & Development

April 2005
RR-05-06

Analysis of Data From an Admissions Test With Item Models

Sandip Sinharay
ETS, Princeton, NJ

Matthew Johnson
Baruch College, NY

April 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Item models (LaDuca, Staples, Templeton, & Holzman, 1986) are classes from which it is possible to generate/produce items that are equivalent/isomorphic to other items from the same model (e.g., Bejar, 1996; Bejar, 2002). They have the potential to produce large number of high-quality items at reduced cost. This paper introduces data from the first known application of items automatically generated from item models in a large-scale assessment and deals with several research questions associated with the data. We begin by reviewing calibration techniques for the analysis of data involving item models; one method assumes that the items are isomorphic, while the other treats items generated from the same item model as distinct, but related. A major question for these type of data is whether these items are isomorphic, that is, if they behave the same psychometrically. This paper describes a number of rough diagnostic measures and a rigorous statistical diagnostic to assess the extent of isomorphicity in the items generated from an item model. Finally, this paper discusses the issue of scoring, an area that needs more research, with data involving item models.

Key words: Bayesian hierarchical model, expected response function, item family, item model

Acknowledgments

The authors thank Shelby Haberman, Manfred Steffen, Aurora Graf, Isaac Bejar, Dan Eignor, Robert Mislevy, and Randy Bennett for useful advice, Rene Lawless for her help with the data, and Loriann Fell and Kim Fryer for help with proofreading.

1. Introduction

A large pool of high-quality items is necessary for the smooth operation of any large-scale testing program, especially those with flexible administration times, to address concerns regarding item exposure and potential disclosure. In an attempt to produce quality items at reduced expenses, there is an increasing interest in generating items automatically. In an ongoing project at ETS, items are automatically generated using *item models* (a term borrowed from LaDuca, Staples, Templeton, & Holzman, 1986), classes from which it is possible to generate/produce items that are equivalent/isomorphic to other items from the same model (e.g., Bejar, 2002).

As a simple example (borrowed from Thissen-Roe & Hunt, 2004), an item model on assessing mathematical aspects of graduated rates, calculations with percents, and algebraic manipulation may look like the following:

> In a certain state, for taxable incomes over y, income taxes are calculated as r percent of the first y of taxable income plus t percent of the amount greater than y. If the taxes calculated for a certain taxable income were w, what was the taxable income?

Several instances are then generated from this model by replacing the variables y, r, t, and w with appropriate numbers and choosing appropriate distractors (if this is a multiple choice item).

Items from (or belonging to) a single item model, whether produced by automatic item generation (AIG) systems (Irvine & Kyllonen, 2002) or rigorous manual procedures, are related to one another through the common generating model and therefore constitute a *family* of related items. The items in the same family are called *siblings;* they assess the same subject matter content and are interchangeable psychometrically (Bejar et al., 2003). They have similar conceptual and psychometric/statistical properties. Naturally, use of such items calls for use of calibration models that can account for the dependence structure among the siblings. Glas and van der Linden (2001, 2003) suggested one such model for dichotomous items. The model assumes a three-parameter (3PL) logistic model (Birnbaum, 1968) to start with and then assumes a normal prior distribution on the item

parameters, the mean vector, and the variance matrix of the normal prior depending on the item family to which the item belongs. Johnson and Sinharay (in press) generalized Glas and van der Linden's model to take into account item families with polytomous items; this paper suggests useful graphical summaries, *family expected response function* (FERF), and *family score function*, for item families, as did Sinharay, Johnson, and Williamson (2003).

A recent large-scale operational test, the Graduate Record Examination® (GRE®), pretested items automatically generated from item models in its quantitative section. There are a number of research questions associated with the study, mostly regarding calibration of the items or item models and scoring examinees; one may also want to know about the success of the item modeling process. This paper attempts to answer these questions from a psychometrician's point of view.

There is no unanimous nomenclature in automatic item generation literature. What we call *item model* is also called *item form* (Hively, Patterson, & Page, 1968), *item template* (e.g., Deane & Sheehan, 2003), *schema* (Singley & Bennett, 2002), *item shell* (Glas & van der Linden, 2003), and so on, by other researchers. Similarly, what we call a *sibling* may be referred to as a *clone* (Glas & van der Linden, 2001, 2003) or an *instance* or an *isomorph* or a *variant*, and so on, elsewhere. We refer to all the siblings generated from an item model as an *item family* (a term also used in Glas & van der Linden, 2001, 2003), for the simple reason that the siblings are related to each other; the term item family is not meant to imply that there is a *parent* item within each family.

2. The Data

Items generated from item models were pretested in a recent administration of the GRE. Test developers and researchers created item models for the GRE quantitative section after studying the features of items from previous operational item pools and choosing a number of these items as the basis of item models (Steffen, Graf, & Levin, 2004). There was one item model for each of four main content areas: remainders (MRE), linear inequality (MLI), quadrilateral perimeter (MQP), and probability (MPR). For each of these four areas, one submodel, each corresponding to Difficulty Levels I (very easy), II (moderately easy), III (moderately hard), and IV (very hard), was produced to cover a wide range of difficulty

for each content area and to ensure homogeneity within siblings. Therefore, there are 16 submodels involved in all, one for each combination of content area and difficulty. We denote them as MRE-I (MRE, very easy); MRE-II, MRE-III, and MRE-IV (MRE, very hard); and MLI-I, and so on. According to our terminology introduced in Section 1, we treat items generated from each of these submodels as part of an item family. Henceforth, Item Families 1, 2, 3, and 4 refer to submodels MRE I, II, III, and IV, respectively; Families 5-8 refer to MLI I-IV; Families 9-12 refer to MQP I-IV; and Families 13-16 refer to MPR I-IV. Ten items (siblings) from each submodel, intended to be isomorphic, and created using the Math Test Creation Assistant (TCA) software developed at ETS (Singley & Bennett, 2002), were administered as part of the pretest section of the GRE; they did not count towards the operational score of the examinees. All items were multiple choice with five options.

As an example, an item generated from a MLI-IV (linear inequality very hard) item model is the following:

The statement "$t - 3 \leq -1$ or $3 - t \geq 13$" is equivalent to which of the following?

A. $t \leq 2$

B. $t \leq -10$

C. $-2 \leq t \leq 13$

D. $-9 \leq t \leq 3$

E. $-10 \leq t \leq 2$

Option A is the correct answer for the item.

The data were collected during operational computer-based testing for the GRE in January and February 2003. The operational computer-based GRE has four sections: analytical writing, quantitative, verbal, and a variable section (quantitative or verbal and used for collecting pretest data on new items). The model-based items were embedded within the 28-item quantitative variable sections. Because the examinees did not know that these were pretest items, it can be assumed that the examinees were motivated while responding to these items. Each of the 32,921 examinees received only four model-based items, a single sibling each from four submodels, and one item for each difficulty level. To

avoid potential speededness and context effects, the within-section item positions of these model-based items were controlled—any of the 40 items for any given content area appeared in the same position. Table 1 gives the composition and sequencing of the model-based items in four configurations.

Table 1.
Composition and Sequencing of the Model-Based Items for the GRE Data

Content	Item position	Configuration 1	Configuration 2	Configuration 3	Configuration 4
Remainder	04	Very easy	Very hard	Mod. hard	Mod. easy
Lin. equality	06	Mod. easy	Very easy	Very hard	Mod. hard
Quad. perimeter	12	Mod. hard	Mod. easy	Very easy	Very hard
Probability	19	Very hard	Mod. hard	Mod. easy	Very easy

Each examinee received items from only one of the four configurations given in the last four columns in the table. The number of examinees receiving any of these items varied from 663 to 1,016, with the average being 821.

2.1 The Research Questions

This study is the first known application of automatically generated items or model-based items in a large-scale assessment used in making high-stake decisions. Of the two other sets of studies involving automatically generated items, the British Army Recruitment Battery (BARB), described in Wright (2002), had 1,273 examinees only, while the one in Hornke (2002) was on intelligence test used in making low-stake decisions. Therefore it is important to study the GRE data carefully. This paper first performs simple analyses of the data.

The next question is how to perform calibration for these data. To answer that, this paper analyzes the data set using the Bayesian hierarchical model of Glas and van der Linden (2001) and simpler alternatives and then discusses the findings.

Although research has concentrated on calibrating the item families, there has not been much work on the scoring of the examinees in these situations when the calibration of the families already has been done. This paper examines the issue of scoring under the Bayesian hierarchical model in detail. If a researcher has used the hierarchical model to calibrate the item families, the FERFs (Sinharay et al., 2003) provide a straightforward

way to score individuals incorporating the variability in the family parameters, as this paper shows. This idea of scoring has some similarities with the idea of expected response function of Lewis (1985, 2001). A comparison is also provided to the situation where one did not use the hierarchical model to score individuals, but rather applied a simple 3PL model treating the siblings the same (as suggested by Hombo & Dresher, 2001).

Bejar et al. (2003) commented that the feasibility of item modeling rests in part on whether the siblings are sufficiently isomorphic (i.e., there is no difference in item characteristics between siblings within a family). Therefore, one objective of our analysis will be to study to what extent the siblings are isomorphic for these data.

2.2 Structure of the Remainder of the Paper

Section 3 describes in detail Glas and van der Linden's (2001, 2003) hierarchical model and simpler alternatives. Section 4 discusses results from simple statistical analyses of the data and those from the application of the models described in the previous section. Section 5 suggests a formal statistical test of isomorphicity. Section 6 discusses scoring in detail. Finally, the paper concludes with a summary of the findings and thoughts on possible future directions.

3. The Statistical Models for Analyzing Data Involving Item Families

One way to analyze data involving item families is to apply the identical siblings model (ISM) suggested by Hombo and Dresher (2001), which assumes a single response function (for example, the 3PL response function for multiple choice tests) for all items in an item family. Effectively, given a response data matrix with rows for examinees and columns for items, fitting an ISM is equivalent to fitting an IRT model after pooling the data columns for all items in each item family into one column. This model, though easy to fit, is restrictive as it ignores variation within an item family. By ignoring the within-family variation, the ISM analysis incorrectly treats siblings within a family as if they were interchangeable, that is, it does not matter which sibling an individual receives, which implies that responses of two individuals to the same sibling are independent given the family item characteristic curve (ICC). The slope (α) of the resulting item response function will likely be too large,

suggesting that there is more information available from responses to siblings drawn from that family, and the standard errors of the item parameters will likely be too small.

Another approach is to use the unrelated siblings model (USM) that assumes a separate, unrelated item response function for all siblings, ignoring the family membership. While the USM can be fitted by standard software, this has the disadvantage that each item has to be individually calibrated before operational use, making the USM unusable in most practical applications of model-based items (see the discussion in Johnson & Sinharay, in press).

A more formal way to analyze data involving item families is to apply the related siblings model (RSM), a hierarchical model (Glas & van der Linden, 2001, 2003) whose first component is a simple IRT model, such as the 3PL model:

$$\Pr(Y_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j)\ \text{logit}^{-1}\left(a_j(\theta - b_j)\right), \tag{1}$$

where Y_j is the score of an examinee with ability θ on Item j and $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$. After making the transformations $\alpha_j \equiv \log(a_j)$, $\beta_j \equiv b_j$, and $\gamma_j \equiv \text{logit}(c_j)$, Glas and van der Linden use a normal distribution to relate the item parameters of items within the same item family as

$$(\alpha_j, \beta_j, \gamma_j)' | \boldsymbol{\lambda}_{\mathcal{I}(j)}, \boldsymbol{T}_{\mathcal{I}(j)} \sim \mathcal{N}_3(\boldsymbol{\lambda}_{\mathcal{I}(j)}, \boldsymbol{T}_{\mathcal{I}(j)}), \tag{2}$$

where $\mathcal{I}(j)$ is the item family to which Item j belongs. The population distribution for the latent abilities is usually assumed to be $\mathcal{N}(0, 1)$. The family mean vector $\boldsymbol{\lambda}_{\mathcal{I}(j)}$ can be partitioned as $\boldsymbol{\lambda}_{\mathcal{I}(j)} = (\boldsymbol{\lambda}_{\alpha_{\mathcal{I}(j)}}, \boldsymbol{\lambda}_{\beta_{\mathcal{I}(j)}}, \boldsymbol{\lambda}_{\gamma_{\mathcal{I}(j)}})'$, and the diagonal elements of the family variance $\boldsymbol{T}_{\mathcal{I}(j)}$ will be referred to as $\tau^2_{\alpha_{\mathcal{I}(j)}}$, $\tau^2_{\beta_{\mathcal{I}(j)}}$, and $\tau^2_{\gamma_{\mathcal{I}(j)}}$ respectively.

This paper further assumes

$$\boldsymbol{\lambda}_{\mathcal{I}(j)} \sim \mathcal{N}\left(\boldsymbol{\mu}_\lambda, \boldsymbol{V}_\lambda\right), \quad \boldsymbol{\mu}_\lambda = (0, 0, \text{logit}(0.2))', \quad \boldsymbol{V}_\lambda = 100\boldsymbol{I}, \tag{3}$$

where $\boldsymbol{I}$ is an identity matrix, and assumes independent inverse-Wishart prior distributions (e.g., Gelman, Carlin, Stern, & Rubin, 2003, pp. 574-575) on the family variances:

$$\boldsymbol{T}^{-1}_{\mathcal{I}(j)} \sim \text{Wishart}_\nu(\mathbf{S}). \tag{4}$$

The prior in (4) implies that the prior mean of $\boldsymbol{T}^{-1}_{\mathcal{I}(j)}$ is $\nu\mathbf{S}$, and that *a priori* there is information that is equivalent to ν observations of the item parameter vector $\boldsymbol{\eta}_j$. We use

$\nu = 4$, the smallest number to ensure that the density is finite everywhere, and assume $\boldsymbol{S} = \frac{10}{\nu} I_3$.

Glas and van der Linden (2003) discussed how to choose items from item families using a Bayesian item selection algorithm in a computer adaptive test. They used a criterion that required the item family selected at any point to have the minimum expected posterior variance. They commented that the smaller the variation within a family, the better the test adapted to the examinee ability.

3.1 Family Expected Response Function

To graphically summarize the output from an RSM for an item family, Sinharay et al. (2003) suggested the FERF that described the probability that an examinee with ability θ correctly responded to a randomly selected item from the item family. The FERF for item family k is obtained as

$$P(\theta|k) \equiv \int_{\boldsymbol{\lambda}_k, \boldsymbol{T}_k} \int_{\boldsymbol{\eta}} P(\theta|\alpha, \beta, \gamma) \phi_3(\boldsymbol{\eta}|\boldsymbol{\lambda}_k, \boldsymbol{T}_k) d\boldsymbol{\eta} f(\boldsymbol{\lambda}_k, \boldsymbol{T}_k|\boldsymbol{X}) d\boldsymbol{\lambda}_k d\boldsymbol{T}_k, \tag{5}$$

where $P(\theta|\alpha, \beta, \gamma) \equiv \text{logit}^{-1}\gamma + (1 - \text{logit}^{-1}\gamma)\ \text{logit}^{-1}\left(e^{\alpha}(\theta - \beta)\right)$, $\boldsymbol{\eta} = (\alpha, \beta, \gamma)'$, $\phi_3(\boldsymbol{\eta}|\boldsymbol{\lambda}_k, \boldsymbol{T}_k)$ is the density function of the multivariate normal prior distribution on $\boldsymbol{\eta}$ and $f(\boldsymbol{\lambda}_k, \boldsymbol{T}_k \mid \boldsymbol{X})$ is the joint posterior distribution of $\boldsymbol{\lambda}_k$ and $\boldsymbol{T}_k$ given the response matrix $\boldsymbol{X}$.

Sinharay et al. (2003) and Johnson and Sinharay (in press) demonstrated that a plot showing the estimate of the FERF of the family along with the estimates of the item response functions of the items provided useful information that included some idea of isomorphicity of the items within each family.

3.2 Estimating the Model and the Family Expected Response Function

Glas and van der Linden (2001) and Sinharay et al. (2003) described Markov chain Monte Carlo (MCMC) algorithms to fit the RSM. The latter paper suggested using Monte Carlo integration to estimate the FERF defined in (5) and discussed how to attach a 95% prediction interval with the estimate.

Step 1 required in the estimation process for the k-th item family consists of the following two substeps: (a) generate a sample of size M from $f(\boldsymbol{\lambda}_k, \boldsymbol{T}_k \mid \boldsymbol{X})$; and (b) for

each of the above M draws of $(\boldsymbol{\lambda}_k, \boldsymbol{T}_k)$, generate m values of the item parameter vector $\boldsymbol{\eta}_j$ from the multivariate normal distribution $\Phi_3(\boldsymbol{\eta}_j|\boldsymbol{\lambda}_k, \boldsymbol{T}_k)$.

Step 2, which uses the sampled item parameters from Step 1, repeats the following substeps for a number of values of θ: (a) for each of the Mm draws of $\boldsymbol{\eta}_j$ obtained in Step 1, compute $P(\theta|\alpha, \beta, \gamma)$; (b) take the mean/median of the above probabilities as an estimate of $P(\theta|k)$; and (c) 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 θ in the interval (-4,4) in Step 2 to estimate the FERF and uses $M = 1,000$, $m = 10$.

4. The Analysis of the GRE Data

4.1 Preliminary Analysis

Figure 1 shows the proportion correct scores of the items for the different item families. For each content area (MRE, MLI, MQP, or MPR—shown along the x-axis), there are four

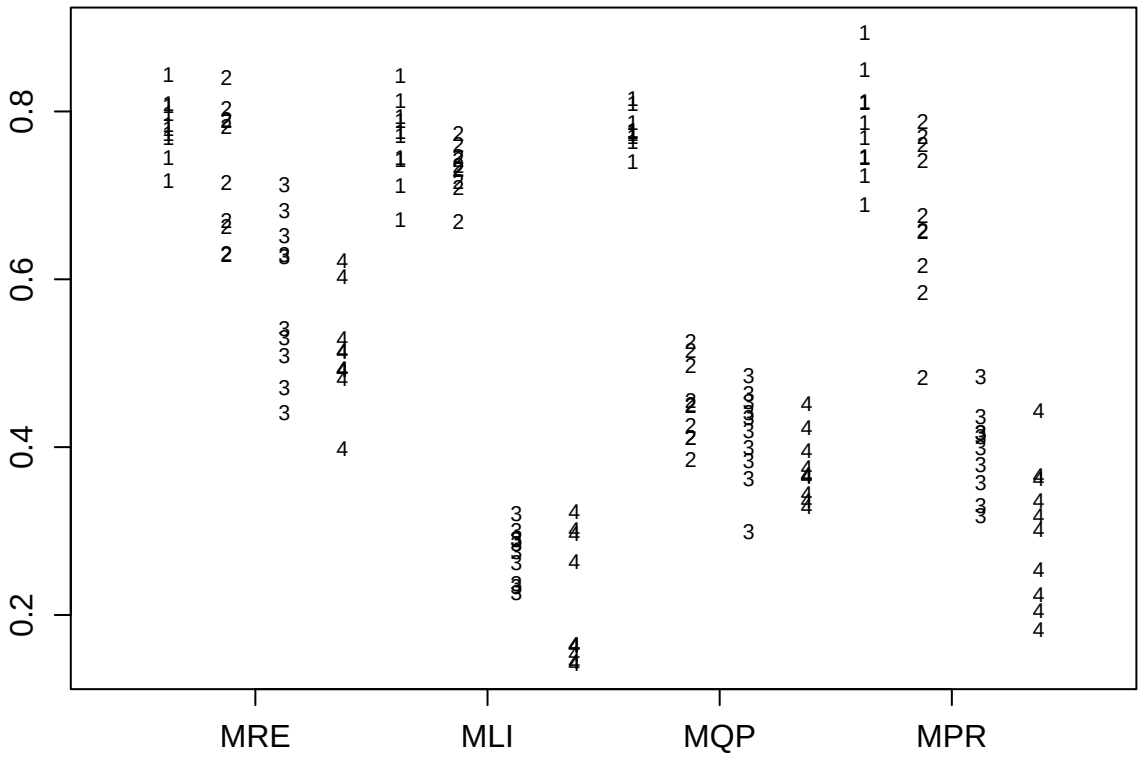

***Figure* 1. Proportion corrects for different item families for the GRE Data.**

vertical lines, one each for the four difficulty levels that define the submodels, along which are plotted the proportion correct for the items in that submodel. A symbol 1 corresponds to an item of Difficulty Level I (very easy), 2 is for Difficulty Level II (moderately easy), and so on.

Each family has its own pattern; the only common pattern being that the items in Difficulty Level 1 are easier on average than those in Level 2, which are in turn easier than those in Level 3, and so on. The difference between the difficulty levels vary over the families. For MRE, there is a substantial overlap of any two successive difficulty levels. The same is true for MPR with the exception of Levels 2 and 3. For MLI, there is a big difference between the first two levels and the last two. For MQP, Level 1 is much easier than the other three, the latter three being close together. There is some variation of proportion correct scores within any family; we will examine later if the magnitude of the variation is significant statistically or practically.

There are a few more interesting patterns. For example, within MLI, Difficulty Level 4, the items seem be divided into two different clusters—one cluster has proportion corrects all close to 0.15 and another cluster has proportion corrects close to 0.3. This family will be discussed later in Section 4.3.

It is possible to perform a rough check by testing the hypothesis of equality of the 10 proportion corrects within any submodel (because under isomorphicity, the items within an item family have the same true proportion corrects) using the χ^2 test statistic (e.g., Rohatgi, 1976)

$$\sum_{i=1}^{10} \frac{n_i(\hat{p}_i - \hat{p})^2}{\hat{p}(1-\hat{p})}, \quad \hat{p} = \frac{\sum_i n_i \hat{p}_i}{\sum_i n_i},$$

with symbols having the usual meaning and with an asymptotic χ^2_9 distribution. This test results in values of the observed χ^2 statistic as shown in Table 2. Note that the 95th percentile of the reference distribution is 16.9 and the 99th percentile is 21.7—so all of the entries in the table are statistically significant at 5% level and 15 out of 16 at 1% level. The table summarizes the level of variation within each item family as a number.

Thus the preliminary analysis of the data shows that although the siblings do not appear to be very different from each other, the item families are not isomorphic statistically.

Table 2.
Values of the χ^2 Statistic for a Rough Test of Isomorphicity

	Difficulty level			
Model	1	2	3	4
Remainder	64.0	206.0	240.9	118.3
Lin. equality	94.5	37.6	32.8	246.1
Quad. perimeter	17.7	67.1	110.9	42.0
Probability	144.2	314.9	76.2	276.1

4.2 Analysis Under USM and RSM Assumptions

Here, the USM takes the form of a simple 3PL model that treats the siblings as different items. Using the notation in Section 3., the noninformative prior distributions assumed (as in, e.g., Sinharay et al., 2003) are:

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

A problem here is that each examinee answers only four model-based items. Therefore, we use the posterior means and standard deviations (SDs) from the operational quantitative section (from a separate analysis performed during operational scoring of examinees) as the means and SDs, respectively, of the normal prior distributions on the proficiency parameters. It is possible to calibrate in a more rigorous manner (e.g., using all the responses in the operational test), but this paper does not investigate that. Figure 2 shows the estimated ICCs for the different item families.

Next, the RSM is fitted to the data. As with the USM analysis, we use the posterior means and SDs from the operational quantitative section as the means and SDs, respectively, of the normal prior distributions on the proficiency parameters. Figure 3 shows the estimated FERFs along with the estimated ICCs. Table 3 shows the posterior median of the variance of the item parameters.

Figure 4, similar to one used in Sinharay et al. (2003), summarizes the sampled values of the mean difficulty λ_{β_k} (along the x-axis) and the within-family variance $\tau^2_{\beta_k}$ (along the y-axis) for all item families. For each family, a point (denoted by the family name) shows the posterior median of λ_{β_k} versus the posterior median of $\tau^2_{\beta_k}$. A horizontal line around a point denotes an approximate 95% equal-tailed credible interval for λ_{β_k}; a vertical line

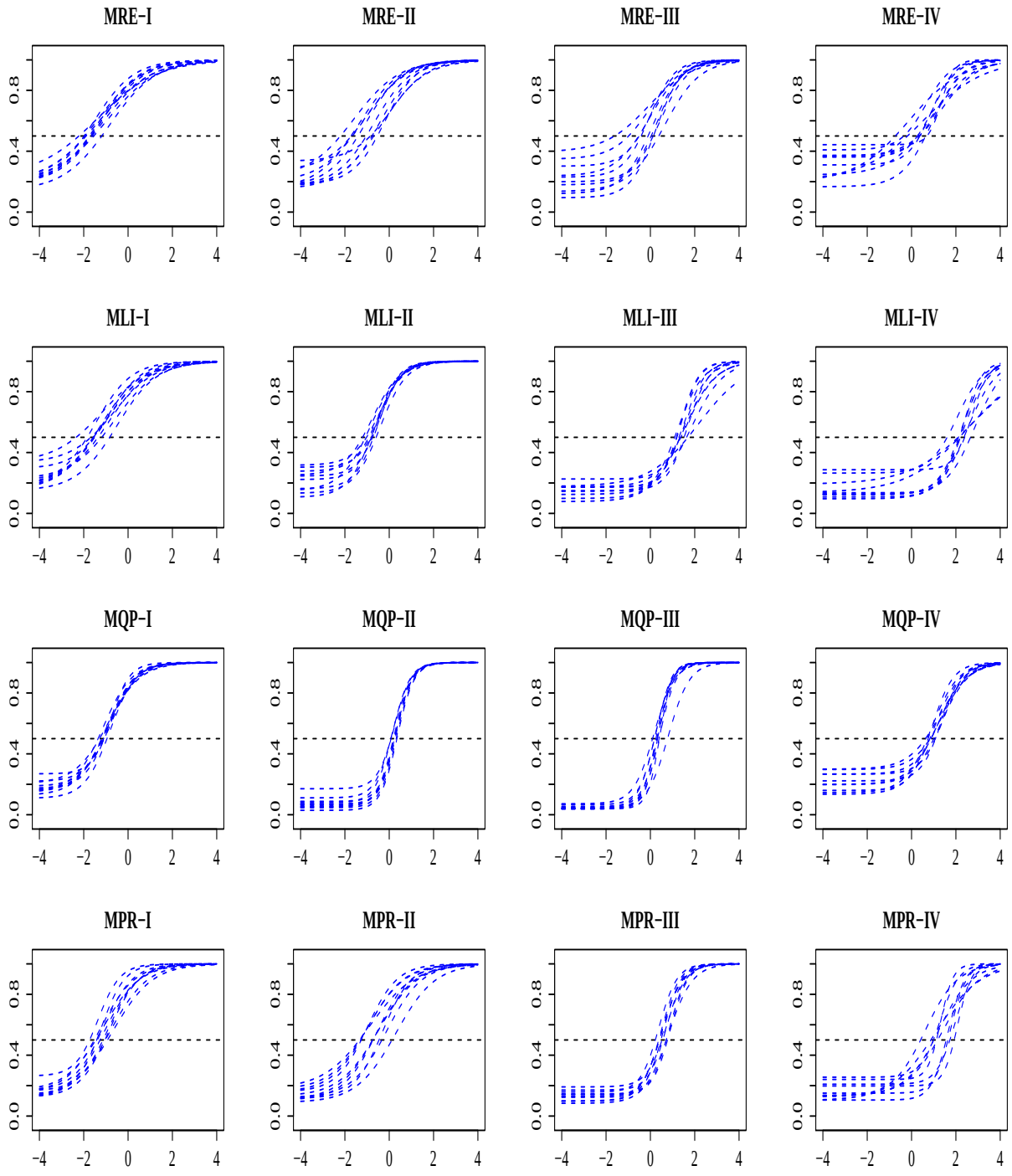

***Figure* 2. ICCs from USM analysis.**

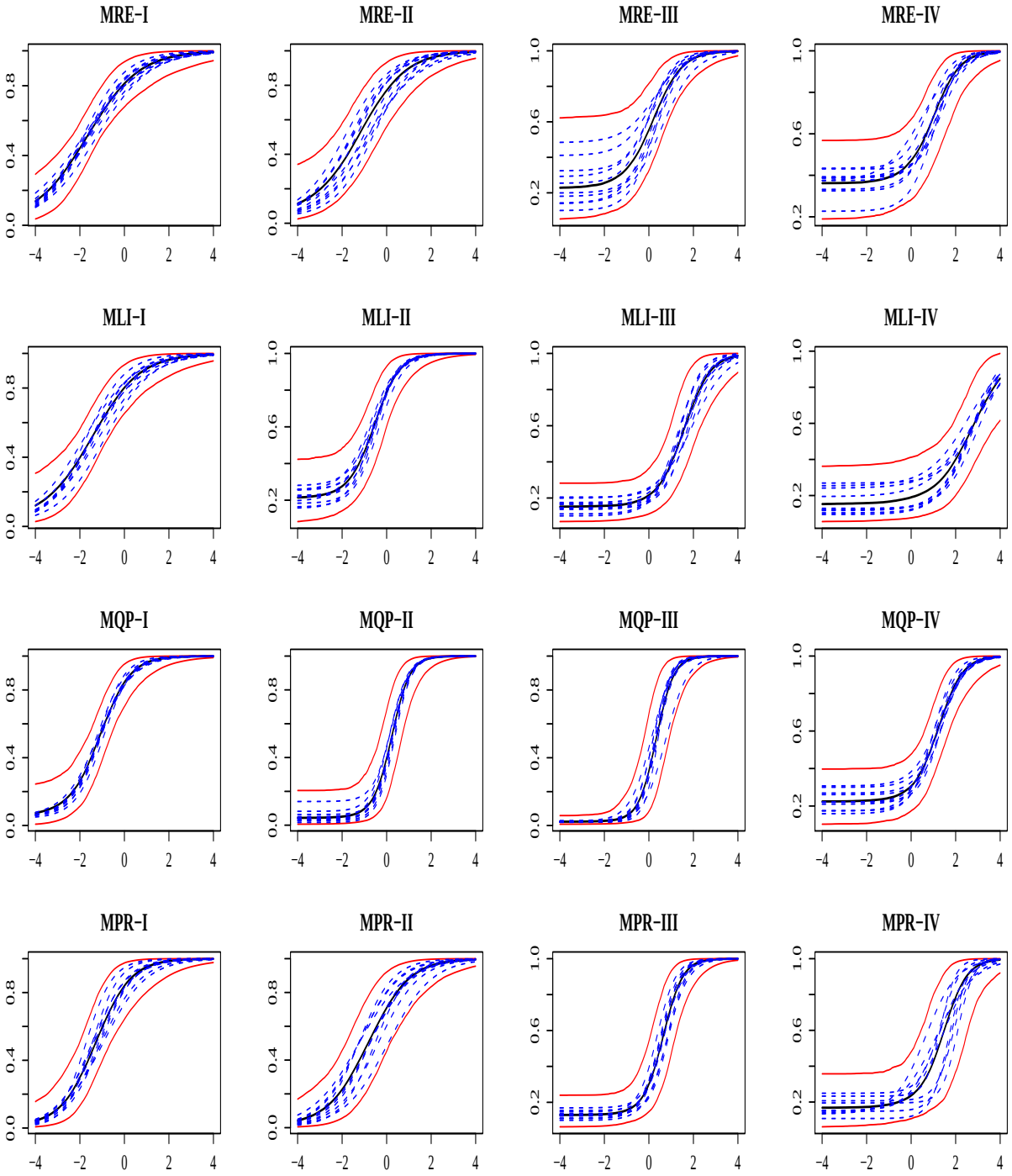

***Figure* 3. FERFs from RSM analysis.**

Table 3.
Posterior Median of the Variance of the Item Parameters for the RSM Analysis

Item family	Variance of log-slope ($\tau^2_{\alpha_{\mathcal{I}(j)}}$)	Variance of difficulty ($\tau^2_{\beta_{\mathcal{I}(j)}}$)	Variance of logit-guessing ($\tau^2_{\gamma_{\mathcal{I}(j)}}$)
MRE-I	0.05	0.08	0.18
MRE-II	0.05	0.19	0.36
MRE-III	0.06	0.08	0.53
MRE-IV	0.08	0.11	0.14
MLI-I	0.05	0.10	0.23
MLI-II	0.05	0.06	0.17
MLI-III	0.07	0.09	0.14
MLI-IV	0.08	0.11	0.27
MQP-I	0.05	0.05	0.16
MQP-II	0.05	0.05	0.58
MQP-III	0.06	0.07	0.14
MQP-IV	0.06	0.06	0.15
MPR-I	0.06	0.09	0.22
MPR-II	0.05	0.19	0.37
MPR-III	0.05	0.07	0.10
MPR-IV	0.11	0.22	0.20

around a point denotes a similar credible interval for $\tau^2_{\beta_k}$. Different line types are used for different content areas.

4.3 Discussion of the Results

The content area MQP has item families that are on an average most isomorphic, the ICCs for the siblings within each family being close to each other. MRE and MPR seem to have the least isomorphic families. The families at Difficulty Level I are most isomorphic, and Difficulty Level IV is probably the least isomorphic. For a number of families (MRE-III, MRE-IV, MLI-II, MLI-IV, MQP-IV), the lower asymptotes of the siblings differ substantially.

The lower asymptotes for a number of the item response functions differ across the two analyses. Take, for example, the MQP-I family. Under the USM analysis, the response functions all have an asymptote in the neighborhood of 0.2, the mean of the prior

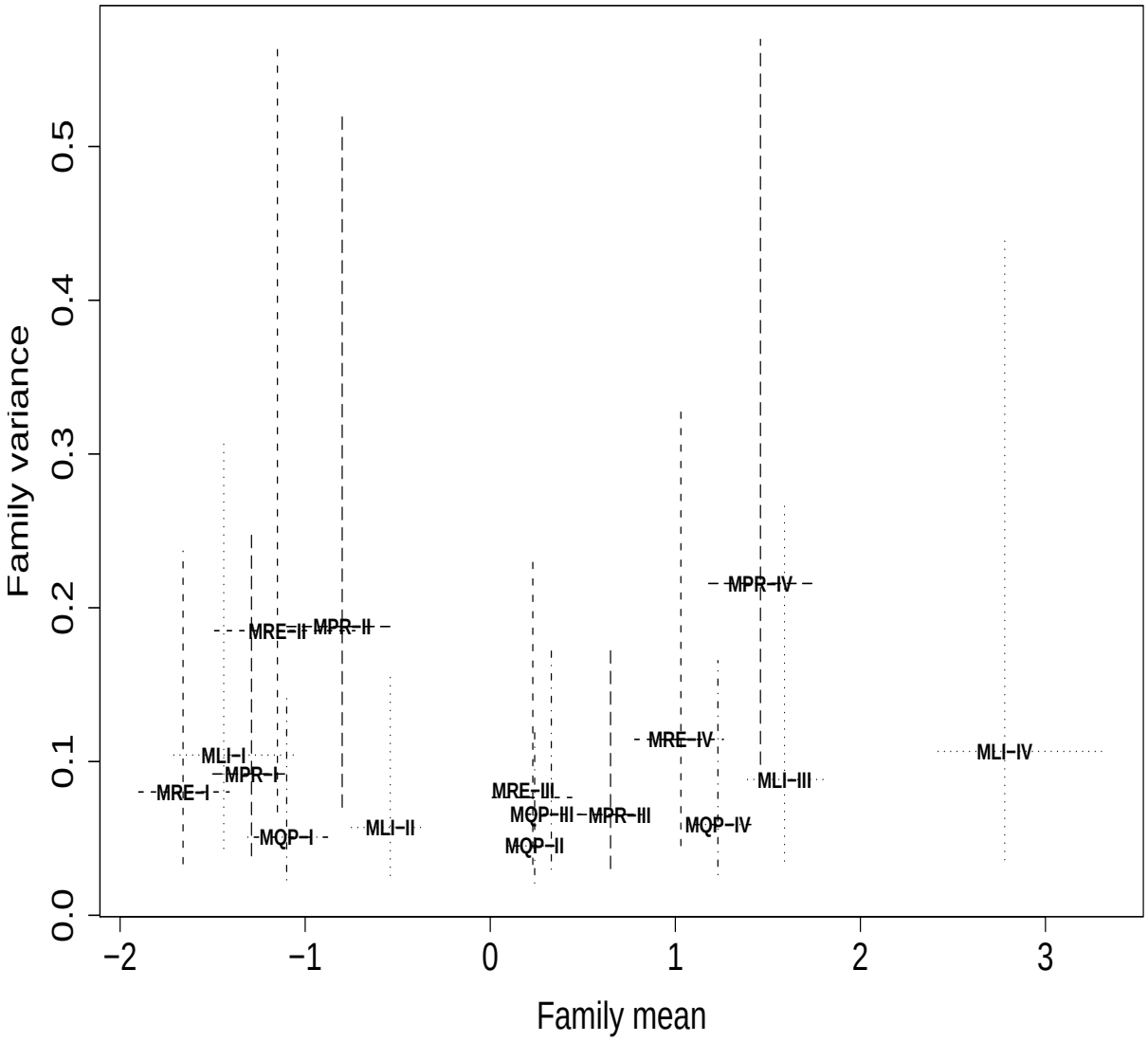

***Figure* 4. Posterior summary of the mean and variance of the difficulty parameters for all the item families.**

distribution on the asymptote. Under the RSM, these same items and their corresponding FERF have an asymptote that is substantially lower. This phenomenon occurs because of the prior structures placed on the lower asymptotes under the two models. The USM assumes a common prior for all items, regardless of family measurement; these priors are centered at 0.2. The items under the USM analysis do not have the power individually

to pull their asymptotes away from 0.2; they are all shrunk towards that prior mean. The RSM analysis, on the other hand, assumes a prior that forces the asymptotes for all items within a family to be centered on *some* value, not necessarily 0.2. Under the RSM, information is borrowed across the items within a family and the family mean asymptote is allowed to differ from 0.2.

To interpret the values in Table 3, it will be helpful to remember that the between-family variances of the mean family parameters ($\boldsymbol{\lambda}_{\mathcal{I}(j)}$) are 0.12, 1.70, and 2.48 for the slope, difficulty, and guessing, showing substantial between-family variation for the difficulty and guessing parameters, respectively. (For MRE, the between-family variances are 0.12, 1.51, and 3.45; for MLI, they are 0.09, 3.76, and 1.92; for MQP, they are 0.07, 0.91, and 1.39; for MPR, they are 0.12, 1.62, and 4.32.)

The estimated variances in Table 3 are the least for the slope parameters and the largest for the guessing parameters and support the earlier finding that there is a substantial variation of the guessing parameters within families. The variances correspond well to the plot of the estimated ICCs from the USM (Figure 2) and RSM (Figure 3). The more isomorphic a family is, the less variance for the difficulty and guessing parameters (the variance for the slope parameters do not vary much). The table, as in the above mentioned figures, shows that all submodels for content area MQP have low estimated variance of difficulty parameters; the same variance is highest on an average for content area MPR.

Figure 4 shows that the item families are well spread out, especially with respect to mean difficulty. As was found in Figure 1, MLI-IV is the most difficult family by a clear margin. The content area MRE seems to have the easiest families. The estimated mean difficulty for MRE-I is less than that of MLI-I, MQP-I, and MPR-I; the same is true for the other difficulty levels. None of the credible intervals for the within-family variance parameters contain the between-family variance of 1.70, providing enough statistical evidence to suggest that the difficulty of items within each family is less variable than that for items across families.

The family MLI-IV shows an interesting pattern in Figures 1, 2, and 3—the siblings seem to belong to two different clusters. Graf, Steffen, and Lawless (2004) described this set of 10 items in detail. Six of these siblings contain a type of distractor that corresponds to

a popular misconception and is an *attractive distractor*. Consider the example item shown in Section 2., which is one of these six siblings. The statement "$t - 3 \leq -1$ or $3 - t \geq 13$" is equivalent to $t \leq 2$ or $t \leq -10$ (i.e., $t \leq 2$) so that the answer is A. However, a popular misconception is to reverse the inequality and end up with the answer $t \leq 2$ and $t \geq -10$ (i.e., $-10 \leq t \leq 2$), which corresponds to the distractor E. About 40-50% of examinees are found to select this distractor, bringing down the proportion corrects for these siblings to about 15%, even lower than the chance level of 20%. On the other hand, four other siblings use a different distractor type (not corresponding to a common mistake) instead of the attractive distractor; as a result, about 5% examinees choose this distractor and the proportion correct for these four siblings is about 30%. The outcome of this clustering of the siblings (into two clusters) is the large value of the χ^2-type statistic in Table 2 and large estimated variances in Table 3.

The family MPR-II is another interesting family. Figures 1, 2, and 3 also show substantial variation of the siblings within this family. In Table 2, the χ^2 statistic is largest for this family; Table 3 shows large variances for this family. These items asked the probability, given that a number x belongs to a set I of integers, of the event E that $y = ax - b$ is positive/negative, where a and b are integers and I consists of consecutive one-digit positive integers. The standard way of solving these items involves computing for each possible value of x the value of y and then checking if the event E occurs. Graf (2003) found that the larger the probability of E in an item of this type, the larger the number of computations required, and the less the proportion correct for the item; the correlation coefficient between the probability of E and the proportion correct is -0.91. The variance between the difficulty of the items within this family is then caused by a varying number of computations required for the items.

As Graf et al. (2004) commented, in order to make the item families more isomorphic, one has to revise the item models by removing the above mentioned factors (distractors for MLI-IV and number of computations for MLI-IV) and other similar ones. Of course, this means substantially more work for the item writers.

5. A Posterior Predictive Test of Isomorphicity

A question of substantial interest to test administrators in situations like this is when the siblings from an item family can be considered isomorphic. The answer to the question might determine the calibration and scoring procedure to be employed. If a family is found to be nearly isomorphic, then the simple ISM approach (Hombo & Dresher, 2001) might be enough; if not, one might need the more complicated hierarchical model. One may have a rough answer by looking at Figures 1 to 4 and Tables 2 and 3, but the following discussion suggests a formal statistical test to answer the question.

The ISM was fitted to the data, using the posterior means and SDs of the proficiencies from the operational quantitative section as the means and SDs, respectively, of the normal prior distributions on the proficiency parameters. The prior distributions on the item parameters are

$$\alpha_j \sim N(0, 10^2), \quad \beta_j \sim N(0, 10^2) \text{ and } \gamma_j \sim N(-1.39, 10),$$

chosen to make the results comparable to the results from RSM. To assess if the model fits the aspect of the data that is of primary interest here, we first compute the proportion correct scores for all the 160 siblings. Then we compute the standard deviation (SD) of the proportion corrects of the 10 siblings for each item family, giving a total of 16 within-family SDs. If the SDs predicted under the ISM are close to the observed SDs, then we can be confident that scoring using ISM in a future test with items from these item families will not be unfair to the students as the variation in difficulty of the items generated from an item family is exactly as predicted by the ISM. A number of posterior predictive data sets (Guttman, 1967; Rubin, 1984; Gelman et al., 2003), which are natural predicted data sets from a Bayesian point of view, were generated and the predicted SDs computed.

The left graph Figure 5 shows a plot of the observed and predicted SDs for the ISM. The dots denote the observed values while the boxplots denote the empirical distribution of the predicted SDs. The ISM severely underpredicts the within-family SDs—the assumption of the same item response function for all the items within a family seems to be too restrictive to reproduce the SDs accurately. The medians of the predicted SDs are all between 0.013 and 0.016. The 95th percentiles of the predicted SDs are all less than 0.023, while the

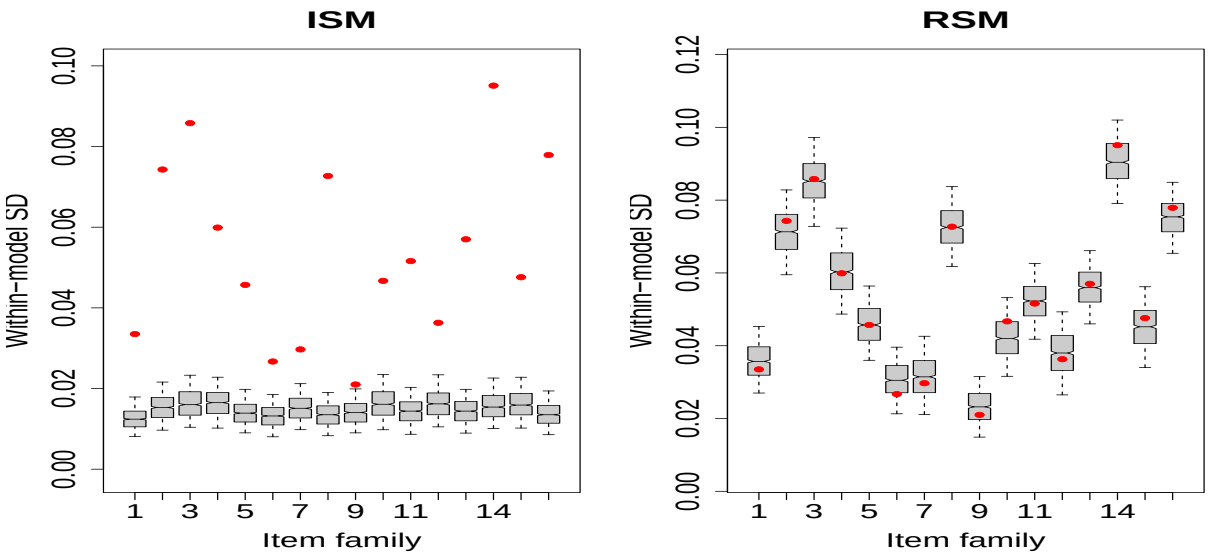

***Figure* 5. The observed and predicted within-family standard deviation of proportion corrects for the ISM (left graph) and RSM (right graph).**

observed SDs are all larger than 0.023, a number of them being much larger than this value.

The right graph in Figure 5 shows a plot of the observed and predicted SDs for the RSM. The model seems to explain the within-family SDs satisfactorily, with the posterior predictive intervals containing the observed SD for all the item families.

These results show that the RSM performs better than the ISM in explaining the aspect of the data that is of practical interest here. However, the more important question is what the practical consequences are if we use the simple ISM instead of the better fitting but more complicated RSM. The later part of the following section deals with this question.

6. Scoring in Future Tests

One major goal of the AIG initiatives is to be able to calibrate an item family (not the individual items from it) once, and then use the items generated from it to score examinees on future tests without the need to calibrate the items individually. Although research has concentrated on calibrating the item families, there has not been much work on the scoring of the examinees in these situations when the calibration of the families is already done. Glas and van der Linden (2003) considered the issue of choice of optimal items in computerized adaptive tests (CAT), which involved scoring individuals; however, they did

not incorporate the variability in the family parameters in the scoring scheme. This section discusses the issue of scoring with RSMs in detail.

Consider a future CAT involving J items generated from item models. Suppose the observed response vector of an examinee with ability θ in such a test is $(y_1, y_2, \ldots y_J)$. Let $\boldsymbol{B}_j$ denote the item parameter vector $(a_j, b_j, c_j)'$ for Item j. Further, let

$$P(y_j|\theta, \boldsymbol{B}_j) \equiv P(Y_j = 1|\theta, \boldsymbol{B}_j)^{y_j} \left(1 - P(Y_j = 1|\theta, \boldsymbol{B}_j)\right)^{1-y_j},$$

where $P(Y_j = 1|\theta, \boldsymbol{B}_j)$ is given by (1). The conditional posterior distribution given $\widehat{\boldsymbol{B}}_j$ for the examinee under the ISM is proportional to

$$p(\theta) \prod_{j=1}^{J} P(y_j \mid \theta, \widehat{\boldsymbol{B}_j}), \tag{6}$$

where $p(\theta)$ denotes the prior distribution on θ, $\widehat{\boldsymbol{B}}_j$ being the item parameter estimate from calibration under the ISM assumption.

Let us denote

$$P(y_j|\theta, \boldsymbol{\eta}_j) \equiv P(Y_j = 1|\theta, \boldsymbol{\eta}_j)^{y_j} \left(1 - P(y_j = 1|\theta, \boldsymbol{\eta}_j)\right)^{1-y_j},$$

where $\boldsymbol{\eta}_j = (\alpha_j, \beta_j, \gamma_j)'$ is the transformed item parameter vector for Item j. To score individuals while using the RSM, Glas and van der Linden (2003) suggested obtaining the posterior distribution of θ as proportional to

$$p(\theta) \prod_{j=1}^{J} \int_{\boldsymbol{\eta}_j} P(y_j \mid \theta, \boldsymbol{\eta}_j) \mathcal{N}_3(\boldsymbol{\eta}_j | \widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}(j)}) d\boldsymbol{\eta}_j, \tag{7}$$

where $\widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}(j)}$ were the estimated mean and variance (from calibration under the hierarchical model assumption) of the item parameters for the item family $\mathcal{I}(j)$ that contains Item j. For future reference, scoring using (7) will be referred to as scoring under RSM Scheme 1.

However, the above approach ignores the variability of $\boldsymbol{\mu}_{\mathcal{I}(j)}$'s and $\boldsymbol{\Sigma}_{\mathcal{I}(j)}$'s, and fixes them at $\widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}(j)}$. A complete Bayesian approach should take this variability into account and obtain the posterior distribution of θ as proportional to

$$p(\theta) \prod_{j=1}^{J} \int_{\boldsymbol{\eta}_j} P(y_j \mid \theta, \boldsymbol{\eta}_j) \mathcal{N}_3(\boldsymbol{\eta}_j | \boldsymbol{\mu}_{\mathcal{I}(j)}, \boldsymbol{\Sigma}_{\mathcal{I}(j)}) p(\boldsymbol{\mu}_{\mathcal{I}(j)}, \boldsymbol{\Sigma}_{\mathcal{I}(j)} | \boldsymbol{X}) d\boldsymbol{\eta}_j d\boldsymbol{\mu}_{\mathcal{I}(j)} d\boldsymbol{\Sigma}_{\mathcal{I}(j)}, \tag{8}$$

where $p(\boldsymbol{\mu}_{\mathcal{I}(j)}, \boldsymbol{\Sigma}_{\mathcal{I}(j)}|\boldsymbol{X})$ is the posterior distribution of the family parameters given the data $\boldsymbol{X}$ from the calibration stage. For future reference, scoring using (8) will be referred to as scoring under RSM Scheme 2.

Comparison of (6), (7), and (8) makes it obvious that scoring with the hierarchical model is more complicated than with the ISM. To approximate the integrals in (7) and (8), Monte Carlo integration (e.g., Gelman et al., 2003) is used. For (7), a random sample from $\mathcal{N}_3(\boldsymbol{\eta}_j|\widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}(j)})$ is generated; the average of the quantity $P(y_j \mid \theta, \boldsymbol{\eta}_j)$ computed over these sampled values provides an estimate of $\int_{\boldsymbol{\eta}_j} P(y_j \mid \theta, \boldsymbol{\eta}_j)\mathcal{N}_3(\boldsymbol{\eta}_j|\widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}(j)})d\boldsymbol{\eta}_j$. For (8), for each draw of a posterior sample of $\boldsymbol{\mu}_{\mathcal{I}(j)}$ and $\boldsymbol{\Sigma}_{\mathcal{I}(j)}$, a number of $\boldsymbol{\eta}_j$s are generated from the prior distribution $\mathcal{N}_3(\boldsymbol{\eta}_j|\boldsymbol{\mu}_{\mathcal{I}(j)}, \boldsymbol{\Sigma}_{\mathcal{I}(j)})$, and finally the average of the quantity $P(y_j \mid \theta, \boldsymbol{\eta}_j)$ is computed over all the sampled values of $\boldsymbol{\eta}_j$. The standard errors for the Monte Carlo integration were found to be too small in magnitude to affect the posterior moments substantially. There is a close relation between the averaging in (8) and the estimation of FERF in (5). For example, for any j, the integral in (8) is $P(\theta|\mathcal{I}(j))$ if y_j is 1, and $1 - P(\theta|\mathcal{I}(j))$ if y_j is 0. Therefore, once the RSM is fitted and the FERFs estimated, the results can be used for scoring individuals under Scheme 2 without much additional computation.

The approach of integrating out the item parameters in (7) and (8) is similar to the approach of computing the *expected response functions* (ERF) suggested by Lewis (1985, 2001); Mislevy, Sheehan, and Wingersky (1993); Mislevy, Wingersky, and Sheehan (1994); and, to some extent, Tsutakawa and Johnson (1990). The ERF approach suggests that scoring individuals (when calibration has been performed) should be based on the following posterior distribution of proficiency

$$p(\theta)\prod_{j=1}^{J}\int_{\boldsymbol{B}_j} P(y_j \mid \theta, \boldsymbol{B}_j)p(\boldsymbol{B}_j)d\boldsymbol{B}_j, \tag{9}$$

instead of the distribution $p(\boldsymbol{B}_j)$ quantifying the information on $\boldsymbol{B}_j$ from the calibration process as defined in (6). For any item j, the integrand in (9) is defined as the expected response function for that item. This work also examines scoring for ISM using the ERF approach, but the results are very similar to those using (6) (because the ERF results in little gain for moderate to large sample size, a phenomenon reported in Lewis, 2001, and in

Tsutakawa & Johnson, 1990), and so we do not report the results here.

The findings from the above mentioned works on ERF provide us with rough guidelines about the difference between the scores provided by ISM and RSM. One has to keep in mind, though, that our context is slightly different from that with ERF; the ERF works concentrate on the difference between (6) and (9) whereas we concentrate on the difference between (6), (7), and (8). Lewis (2001) reported for Rasch models (using simulations), that ignoring the uncertainty in item parameters (and fixing them at point estimates) resulted in underestimation for high-ability students and overestimation for low-ability students, as well as a decrease in the posterior SDs. The effect is mild for low to moderate uncertainty in item parameters and becomes more severe as the amount of uncertainty in item parameters increases. Tsutakawa and Johnson (1990) reported the same effect as in Lewis (2001) using a calibration sample of 400 examinees, but also reported little effect for a calibration sample of 1,000 examinees. Mislevy et al. (1994), using a calibration sample of 100 examinees, found for the 3PL model that ignoring the variance in item parameters did not affect the posterior means, but it considerably underestimated posterior standard deviations.

A slight complication for the GRE data is that for some item families the ISM point estimates and the RSM estimates $\widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}$ are slightly different, which is another source of difference between the ability estimates provided by ISM and RSM.

6.1 Scoring for Four Items

Suppose the (future) CAT starts with four items and chooses Item 5 onwards based on the posterior mean, under a $\mathcal{N}(0,1)$ prior, obtained from the responses to the first four items. We simulated responses of 250 examinees to four items each (one from each content area) under the assumption that the parameters of an item from family $\mathcal{I}$ has the distribution $\mathcal{N}_3(\widehat{\boldsymbol{\mu}}_{\mathcal{I}}, \widehat{\boldsymbol{\Sigma}}_{\mathcal{I}})$. Because this is a short test, to maximize information on θ, low-ability students (true $\theta \leq -0.67 \equiv$ the 25th percentile of the standard normal distribution) are assumed to receive items from very easy submodels, that is, they receive one item each from families MRE-I, MLI-I, MQP-I, and MPR-I, and so on. The true abilities were generated from a $\mathcal{N}(0,1)$ distribution. Where required, the posterior median of a parameter plays the role of a point estimate of the parameter.

Figure 6 compares posterior means and posterior SDs of examinee abilities for scoring under the ISM, RSM Scheme 1, and RSM Scheme 2. This paper obtains these values by

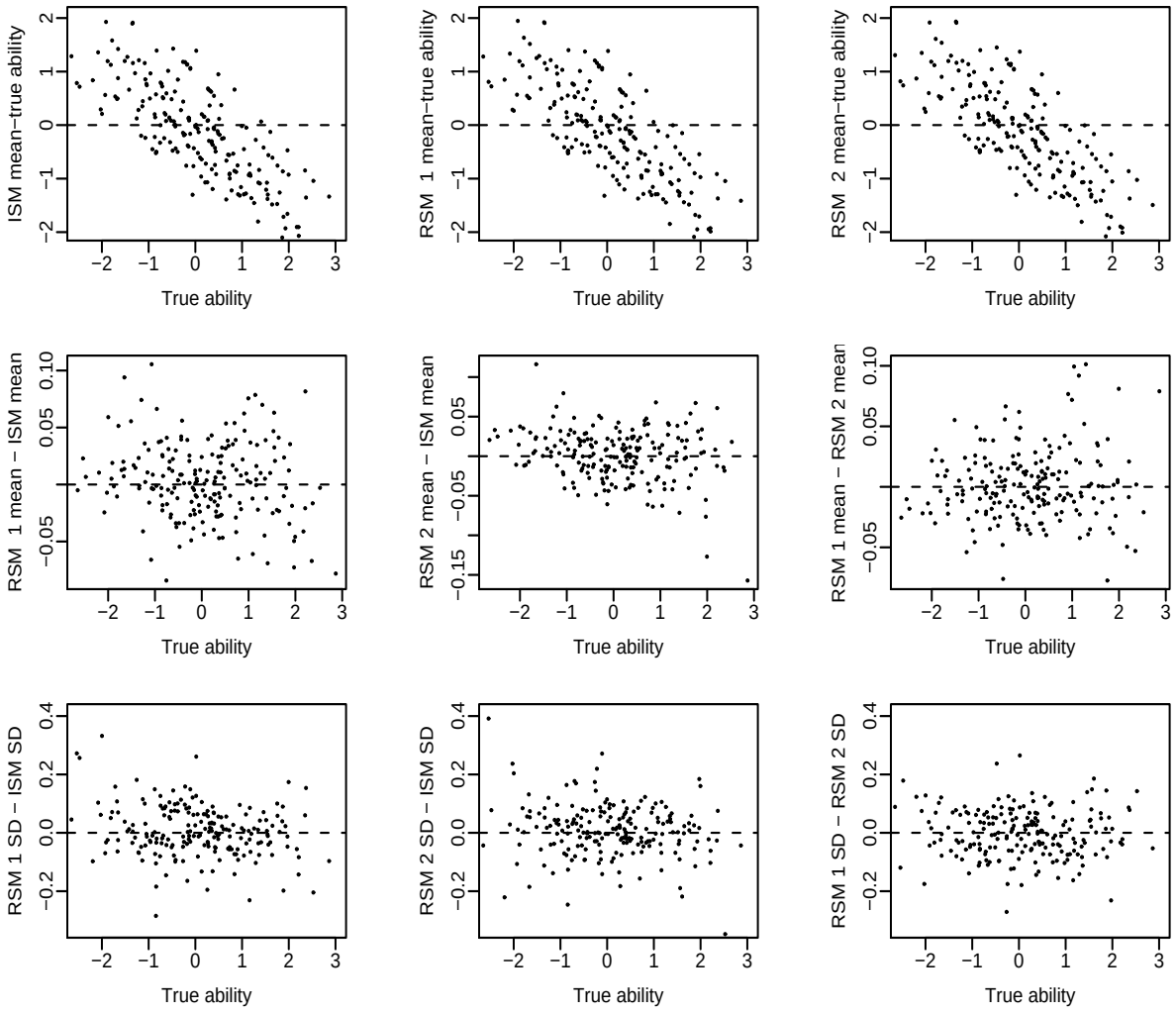

***Figure* 6. The comparison of posterior means and SDs of examinee abilities under the ISM and the hierarchical model for four-item test for $\mathcal{N}(0,1)$ population distribution.**

running an MCMC algorithm using the posterior distributions given in (6), (7), and (8). Table 4 shows the RMSE's, that is, the quantity

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{generating } \theta_i - \text{posterior mean of } \theta_i)^2},$$

and the average posterior SDs over all the examinees for this situation.

Table 4.
RSMEs for the Different Scoring Techniques

No. of items	Prior variance	RMSE			Av. posterior SD		
		ISM	RSM Sch. 1	RSM Sch. 2	ISM	RSM Sch. 1	RSM Sch. 2
4	1	0.88	0.88	0.88	0.75	0.76	0.76
4	25	2.73	2.70	2.75	2.32	2.35	2.40
16	1	0.49	0.49	0.49	0.48	0.48	0.49
16	25	1.15	1.17	1.21	0.80	0.83	0.88

The posterior means and SDs differ slightly over the three techniques, evident both from the plots and the table. For the top row of plots, a negative correlation is the outcome of shrinkage under Bayesian estimation with $\mathcal{N}(0, 1)$ prior. Because there are only four items, there is little information about θ in the data, and hence the posterior mean is pooled towards 0 for most examinees—this results in decrease of the difference (of the posterior mean and the true value) as true ability increases. There is no visible pattern in the plots in the second or third row. There is hardly any difference of RMSE or average posterior SD for the three approaches to scoring.

Figure 7 compares posterior means and posterior SDs of examinee abilities under the ISM, RSM Scheme 1, and RSM Scheme 2 when the prior variance is taken as 25 (i.e., the prior is more diffuse, making the posterior mean become close to the MLE). The corresponding RMSEs and average posterior SDs are given in Table 4. The difference between ISM and RSM is much more here than in Figure 6 for both mean and SDs. The RSM produces slightly larger posterior SDs on average than the ISM. RSM Scheme 1 and RSM Scheme 2 perform almost equally well. The RSM has a tendency to overestimate the high posterior means and underestimate the low posterior means. We do not observe the patterns as in Lewis (2001), Mislevy et al. (1994), or Tsutakawa and Johnson (1990) here.

The GRE will become a linear test (from adaptive) in the near future, and the above analysis may not be the most appropriate to the GRE then. However, other CATs might find the above results useful; also, even for the GRE, the above analysis (along with the following) provides some idea about the practical impact of the lack of isomorphicity of the item families.

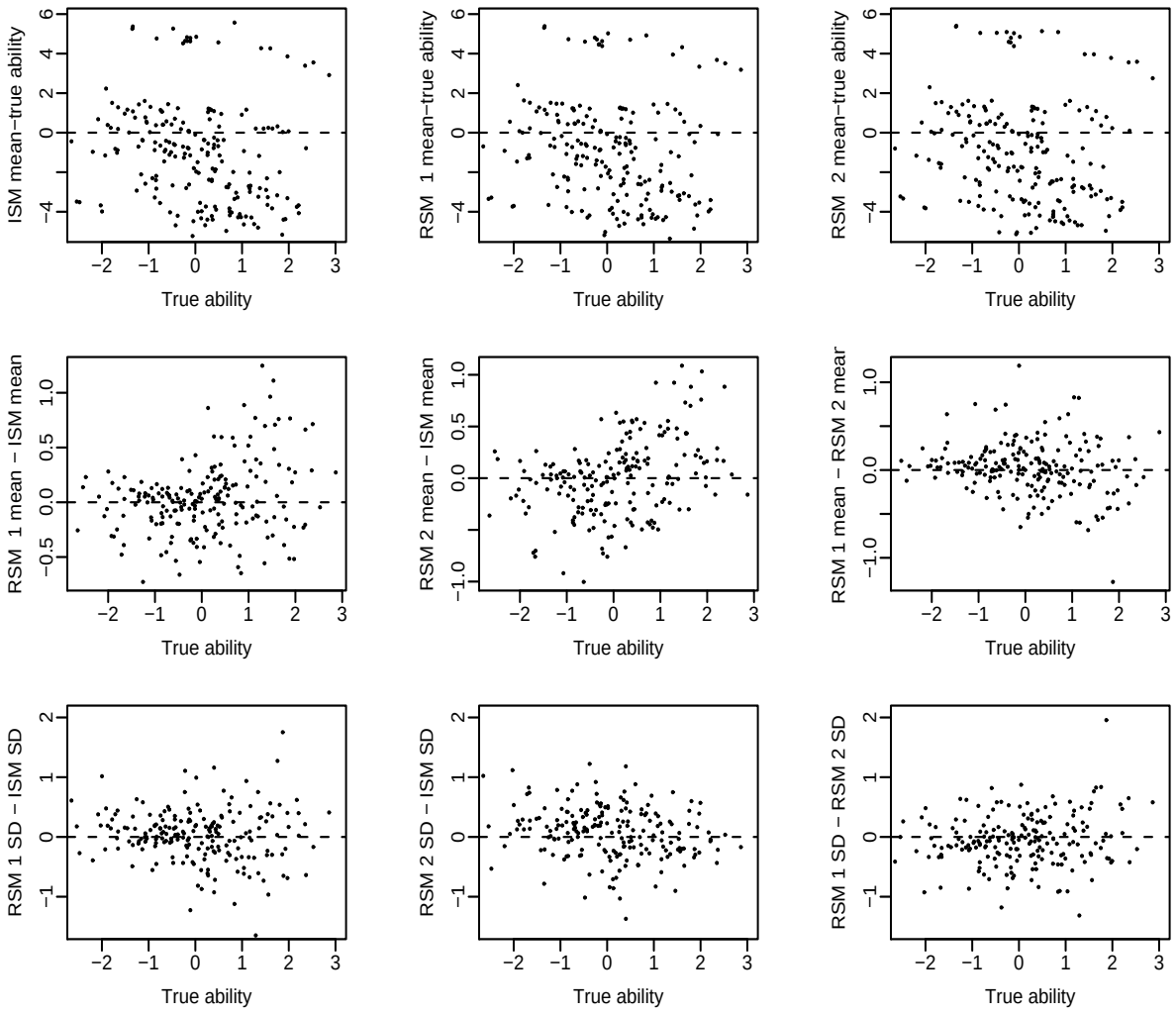

***Figure* 7. The comparison of posterior means and SDs of examinee abilities under the ISM and the RSM for 4-item test for $\mathcal{N}(0, 5^2)$ population distribution.**

6.2 Scoring for 16 Items

Now consider another scenario—one where each examinee answers 16 items, one from each item family that appears in this example. Figure 8 compares posterior means and posterior SDs of the examinee abilities under the ISM, RSM Scheme 1, and RSM Scheme 2 under this scenario when the prior variance is 1. As expected, the errors in ability estimation are smaller here than in Figure 6. There is virtually no difference between ISM and RSM (Scheme 1 or 2).

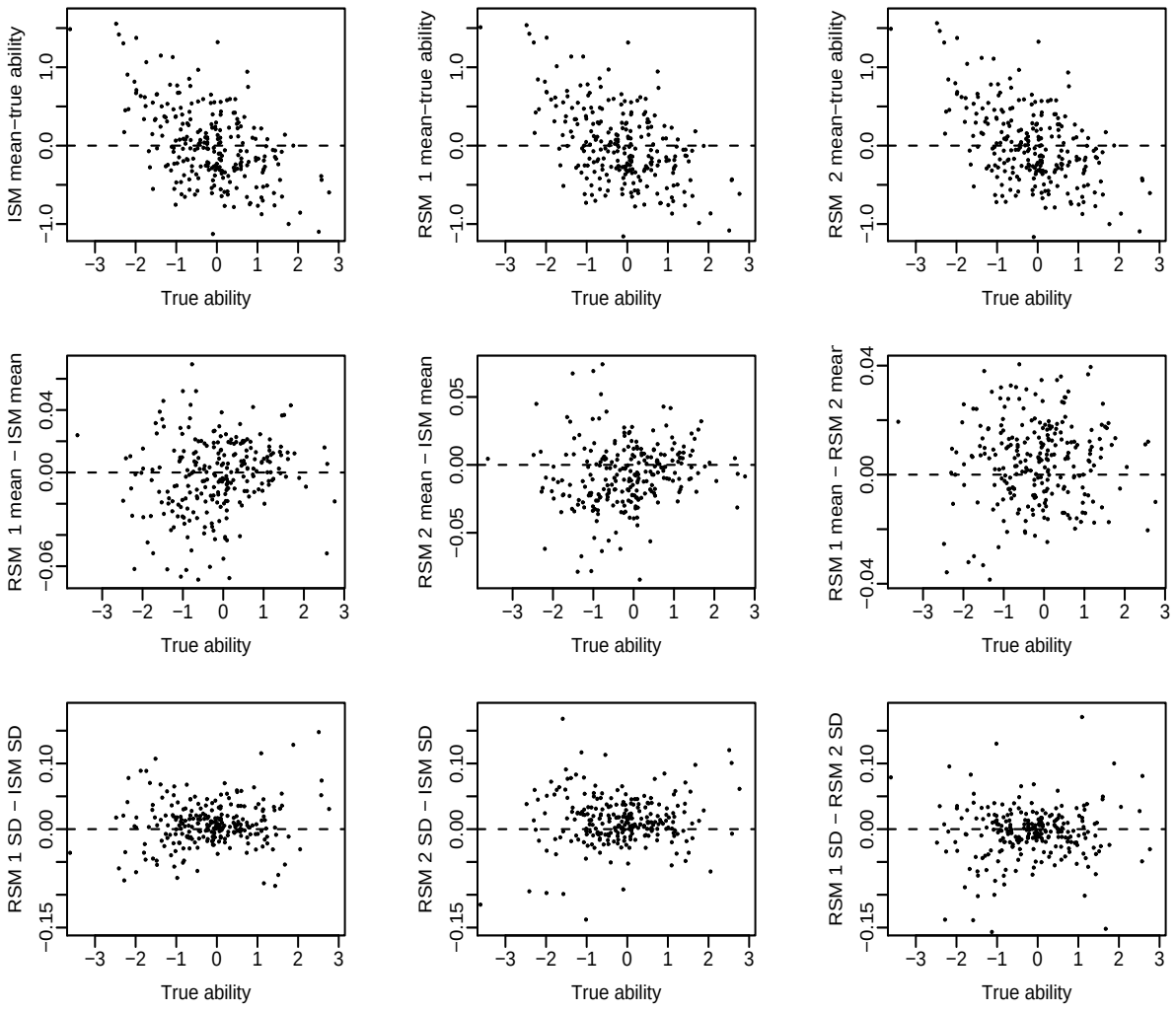

***Figure* 8. The comparison of posterior means and SDs of examinee abilities under the ISM and the RSM for 16-item test for $\mathcal{N}(0,1)$ population distribution.**

Figure 9 compares posterior means and posterior SDs of examinee abilities under the ISM, RSM Scheme 1, and RSM Scheme 2 in this scenario when the prior variance is 25. The amount of shrinkage of posterior means for both of these cases is much less than that for four items, which is expected. The plots, especially those for the SDs, show some outliers. The ISM results in smaller SDs on average compared to RSM Scheme 1 or RSM Scheme 2. We observe the patterns as in Lewis (2001), Mislevy et al. (1994), or Tsutakawa and Johnson (1990) to some extent here.

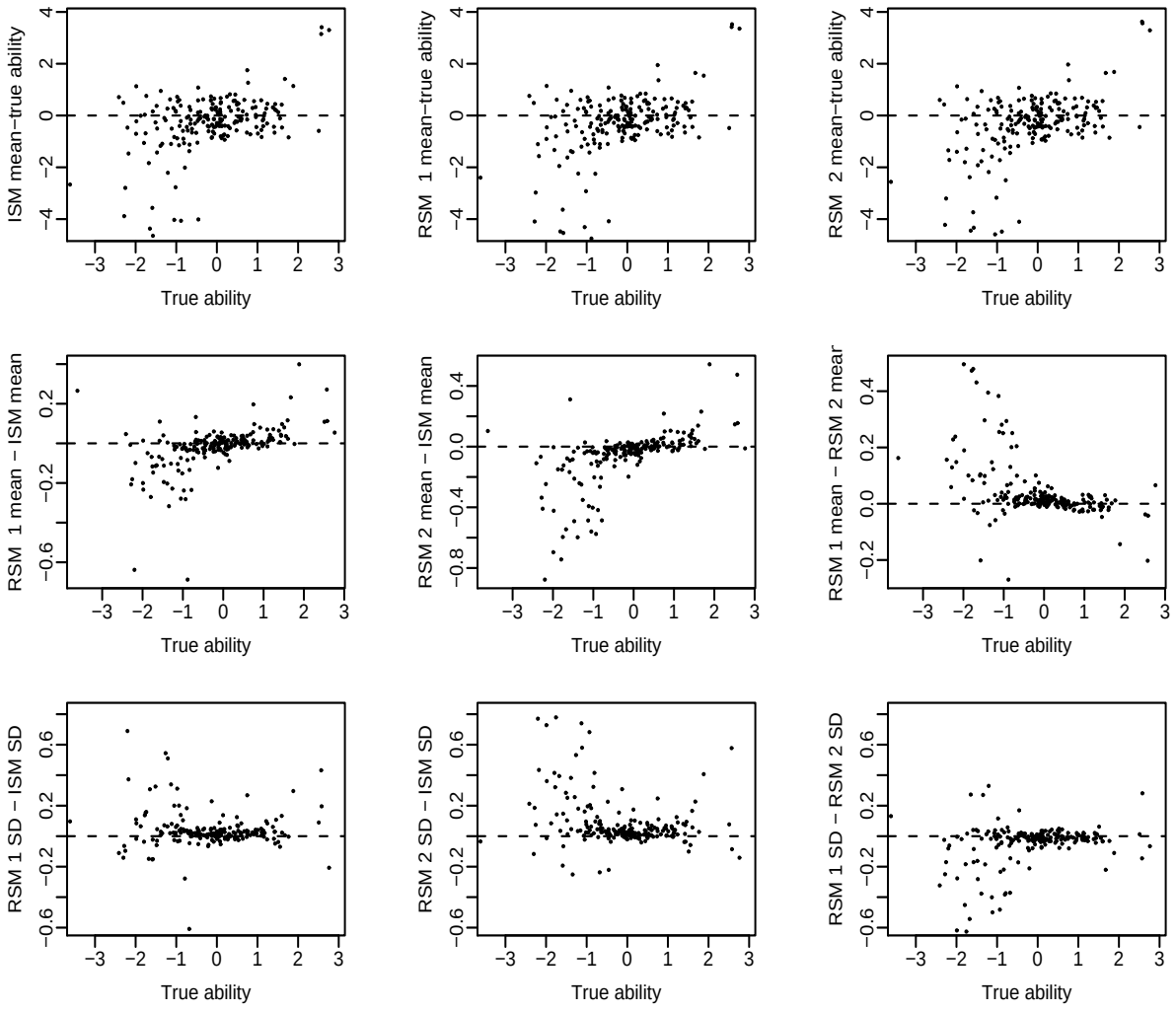

***Figure* 9. The comparison of posterior means and SDs of examinee abilities under the ISM and the RSM for 16-item test for large prior variance for $\mathcal{N}(0, 5^2)$ population distribution.**

6.3 Discussion

One more relevant factor for comparing ISM and RSM is computing time. The RSM takes about twice as much as the ISM for calibration (both were calibrated using Fortran 77 programs) and takes about 8 to 10 times as much as the ISM for scoring.

It is possible to perform a more rigorous analysis here by simulating a full CAT using a whole-item pool and examining the difference caused by the use of the ISM in the θ-scale or in the scaled score, but this paper does not delve into that.

The above results suggest that if the prior distribution of the ability distribution is $\mathcal{N}(0,1)$, then there is little difference between the ISM and RSM—therefore, there is probably no need to use the more complicated RSM. Glas and van der Linden (2003), using the same $\mathcal{N}(0,1)$ prior distribution, obtained very similar results; in 48 cases (the cases determined by the number of items, ratio of within-family and between-family variances, etc.) of scoring that they consider, the mean absolute error (MAE) in ability estimation for ISM is less than or equal to that of RSM in 19 cases and the MAE for ISM is never much higher than that of RSM. Therefore, the evidence of misfit of the ISM in Section 5. is not of much practical consequence. However, the above results also show that if one uses a more diffuse prior distribution, like the $\mathcal{N}(0,5^2)$ distribution (which leads to the posterior mean being very close to the MLE), then the ISM may result in underestimation of variability in ability estimation. Scoring was also performed by generating data from families with family variances that are larger than the values obtained for the GRE data; in that case, the difference in the posterior SDs of ISM and RSM is more prominent; also, the patterns of difference between the two methods are quite similar to those observed by Lewis (2001), such as the posterior SD for RSM is almost always larger than that of ISM and posterior mean for ISM is smaller for that for RSM for large θs.

There is little difference between RSM Scheme 1 and RSM Scheme 2, that is, there is no penalty for ignoring the variation of $\widehat{\boldsymbol{\mu}}_{\mathcal{I}(j)}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{I}(j)}$.

The issue of scoring with model-based items needs further investigation. It is important to figure out when ISM is preferable over RSM and vice-versa.

7. Conclusions

Due to flexible administration times, large-scale assessments such as the GRE require large pools of items from which to draw a given test-taker's examination, otherwise the security of the items may be compromised. Item modeling initiatives develop numerous items from a common item model as one way to populate these vast item pools. The two main advantages to using these item models are: (a) they reduce the burden on test developers, because the developers only need to create the item stem rather than all the individual items; and (b) the family structure implied by the item generation models can

be utilized in calibration; only the families need to be calibrated, not each individual item. This paper focuses on pretest data from the GRE that involved item models and attempts to answer the research questions that are involved with these unique data.

There are three calibration techniques that can be utilized for item models: the unrelated sibling model (USM), which ignores similarities within families; the identical sibling model (ISM), which ignores variability within families; and the related siblings model (RSM), which incorporates a hierarchical structure to relate items within a family. Although the USM is the ideal calibration model, it requires that all items within a family be calibrated separately, precluding its use in most practical situations. The ISM and RSM have the advantage that the *family* can be calibrated once and the future items from these families need not be calibrated separately.

This paper demonstrates by using simple descriptive statistics such as the variation of proportion corrects within item families, and with more formal measures like posterior predictive checks that the pretest items that were administered on the GRE do not behave isomorphically, that is, there is some within-family variation in the behavior of the items. In fact, there were some families where items behaved quite differently from their siblings (closer inspection revealed intuitive reasons for the extra variation). A model such as the ISM, which ignores within-family variation, will not be able to handle such unexpected variation in the items within a family. As a contrast, the RSM handles such extra variation quite nicely by simply returning a larger estimate for the within-family variation.

As discussed in Section 4.3, Graf et al. (2004) note that in order to make multiple-choice items as isomorphic as possible, extreme care must be taken when selecting the distractors. The RSM can be extended to include such collateral information, which would try to further explain the within-family variation. For example, the model in (2) could be extended to incorporate item-level predictors $\boldsymbol{x}$ (e.g., distractor information) with a regression model of the form $\boldsymbol{\eta}_j \sim \mathcal{N}_3(\boldsymbol{\lambda}_{\mathcal{I}(j)} + \boldsymbol{\zeta}'\boldsymbol{x}, \boldsymbol{T}_{\mathcal{I}(j)})$. This is a possible area for future research.

One surprising result found in this paper is that the more intuitive RSM provides little gain over the rather restrictive ISM in scoring. Our original conjecture was that the RSM correctly accounts for the variation within the item families (which is supported by the statistical test in Section 5. that finds the ISM to have an inadequate fit to the data and

finds the RSM to have a good fit), and hence scoring under the RSM would produce more accurate scores, that is, the RMSE would be smaller (compared to the ISM) when the RSM is utilized for scoring individuals. However, that conjecture is not supported by our results. RSM does not perform much better than ISM regarding ability estimation in the simulation results of Glas and van der Linden (2003), either. These findings may indicate that the ISM is adequate (and that there is no need of the complicated RSM) for ability estimation for the level of within-family variation present in GRE data, that is, the test developers were able to control the within-family variation within acceptable limits so that the assumption of interchangeable siblings is nearly satisfied so far as ability estimation is concerned. The added value of the RSM then is to provide a more faithful accounting of the data (as clear from the results of the posterior predictive checks) and to find item families that are far from isomorphic so that item writers can use that knowledge to find factors causing lack of isomorphicity and to remove them as far as possible. The RSM also should provide an easy way of including collateral information. However, this is an area that requires more research. On a related note, it will be useful to find a measure of the level of lack of isomorphicity that can be tolerated for a particular problem—such a measure will benefit the test administrators.

References

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS RR-96-13). Princeton, NJ: ETS.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning and Assessment, 2*(3).

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

Deane, P., & Sheehan, K. (2003). *Automatic item generation via frame semantics: Natural language generation of math word problems.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York: Chapman & Hall.

Glas, C. A. W., & van der Linden, W. J. (2001). *Modeling variability in item parameters in CAT.* Paper presented at the North American Psychometric Society meeting, King of Prussia, PA.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*(4), 247-261.

Graf, E. A. (2003). *A cognitive analysis of item model instances developed for the quantitative GRE: A collaborative effort.* Paper presented at the ETS Item Model Workshop, Princeton, NJ.

Graf, E. A., Steffen, M. & Lawless, R. (2004). *Statistical and cognitive analysis of quantitative item models.* Paper presented at the 30th annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, PA.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B, 29*, 83-100.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement, 5*(4), 275-290.

Hombo, C., & Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Hornke, L. F. (2002). Item-generation models for higher order cognitive functions. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Irvine, S. H., & Kyllonen, P. C. (Eds.) (2002). *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Johnson, M. S., & Sinharay, S. (in press). *Calibration of polytomous item families using Bayesian hierarchical modeling. Applied Psychological Measurement.*

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modeling procedure for constructing content-equivalent multiple-choice questions. *Medical education, 20*(1), 53-56.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known item response functions.* Paper presented at the annual meeting of the Psychometric Society, Nashville, TN.

Lewis, C. (2001). Expected response functions. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory.* New York: Springer.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55-78.

Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (ETS RR-94-28-0NR). Princeton, NJ: ETS.

Rohatgi, V. K. (1976). *An introduction to probability theory and mathematical statistics.* New York: Wiley.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12*, 1151-1172.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of

schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*(4), 295–313.

Steffen, M., Graf, E. A., & Levin, J. (2004). *An investigation of the psychometric equivalence of quantitative isomorphs: Phase 1.* Manuscript in preparation.

Thissen-Roe, A., & Hunt, E. (2004). *A scaling technique for analyzing data from formative testing.* Paper presented at the 30th annual conference of the International Association for Educational Assessment (IAEA), Philadelphia, PA.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55,* 371-390.

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