Improving the Quality of Ability Estimates Through Multidimensional Scoring and Incorporation of Ancillary Variables

Jimmy de la Torre

Applied Psychological Measurement 2009 33: 465 originally published online 29 June 2009
DOI: 10.1177/0146621608329890

Applied Psychological Measurement
Volume 33 Number 6
September 2009 465-485
© 2009 SAGE Publications
10.1177/0146621608329890
http://apm.sagepub.com
hosted at
http://online.sagepub.com

Improving the Quality of Ability Estimates Through Multidimensional Scoring and Incorporation of Ancillary Variables

Jimmy de la Torre
Rutgers, The State University of New Jersey

For one reason or another, various sources of information, namely, ancillary variables and correlational structure of the latent abilities, which are usually available in most testing situations, are ignored in ability estimation. A general model that incorporates these sources of information is proposed in this article. The model has a general formulation that allows incorporation of either source or both sources of information in scoring the examinees using various item response models and subsumes the traditional method of expected a posteriori as a special case. Results show that using the different sources of information singly or simultaneously provides better ability estimates (i.e., higher correlation with the true abilities and smaller posterior variance and mean squared error). The optimal condition occurs when several short tests measuring highly correlated abilities that also correlate highly with the covariates are used. Markov chain Monte Carlo parameter estimation algorithms corresponding to the different model formulations are also developed. Simulated and actual data are analyzed to establish the usefulness and feasibility of the proposed models. Several practical considerations in using these models are also discussed.

Keywords: *item response theory; Markov chain Monte Carlo; ability estimation; Bayesian estimation; multidimensionality; ancillary variables; covariates*

Introduction

In most testing situations, several tests that measure different but highly correlated traits are commonly administered at the same time. For example, in the National Assessment of Educational Progress (NAEP) mathematics assessment, the correlations for the five traits are all greater than 0.70, and they average approximately 0.85 when the correlations are estimated conditional on the demographic subgroups. The marginal correlation between the traits are even higher when the conditioning variables are ignored (Johnson & Carlson, 1994). A more efficient method of ability estimation can result when the correlational structure of the latent traits is taken into account. The ability estimates obtained by de la

Author's Note: This article is based on the author's PhD dissertation at the University of Illinois at Urbana-Champaign, which was supported by a Spencer Foundation Dissertation Fellowship, an American Psychological Association Dissertation Research Award, and an Applied Psychological Measurement grant. The author is grateful to CTB/McGraw-Hill for making the data available for this study.

Torre and Patz (2005) and Wang, Chen, and Cheng (2004) by considering a multidimensional approach to ability estimation were more precise compared to ability estimates obtained using one test at a time. In addition, Li and Schafer (2005), Segall (1996), and Wang and Chen (2004) showed the advantages of taking a multidimensional perspective in item selection and scoring in computerized adaptive testing. Their results demonstrate that a multidimensional approach to adaptive testing produces substantial gains in efficiency compared to the unidimensional procedure. Augmenting a score from a particular test by information obtained from other tests is not limited to the item response theory (IRT) framework. Wainer et al. (2001) illustrated how the accuracy of conventional summed scores and scale scores can be improved through borrowing strength using test reliabilities and intertest correlations.

In addition to examinees' responses to the test questions, other ancillary information about the examinees is also available in many testing situations. Examples of this ancillary information include demographic variables, such as sex, age, and race, and educational variables, such as grade level, courses taken, and previous test scores. In computer-based testing, response time can also be a source of ancillary information (e.g., Fox, Klein Entink, & van der Linden, 2007; van der Linden, 2007). Depending on the design, incorporating ancillary variables that differentiate the examinee populations can lead to unbiased estimates of population parameters, more precise ability estimates, and consistent parameter estimates (Little & Rubin, 1983; Mislevy, 1984, 1987; Mislevy & Sheehan, 1989). Another example of the use of ancillary information is in computing plausible values in the IRT-based scales of NAEP. The method proceeds by sampling from the conditional distribution of examinee i, $p(\theta|\mathbf{x}_i,\mathbf{y}_i)$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the responses to the items and the background questions, respectively. Imputed values are obtained by using either univariate or multivariate regression of the ability on the ancillary variables (Mislevy, Johnson, & Muraki, 1992).

The examples above illustrate that although the sources of additional information (i.e., responses to other tests and ancillary variables) have been used to improve test scoring, they are typically used separately. Better ability estimates can be obtained by using models that incorporate the various sources of ancillary information simultaneously. One such model is the mixed-coefficient multinomial logit model (Adams & Wu, 2006), which can be used in conjunction with item response and latent regression models. However, in its current formulation, the item response model is limited to the Rasch family. Therefore, the primary goal of this article is to propose a general framework for scoring examinees that (a) allows for information from responses to other tests and ancillary variables to be simultaneously incorporated in the model and (b) can be used in conjunction with various item response models. It seeks to examine how the correlational structure of the latent traits and the auxiliary information affect the accuracy of ability estimation, separately and jointly. With all the additional sources of information incorporated, the resulting model is necessarily more complex compared to the conventional models (i.e., models that use responses to single tests only), and parameter estimation would involve maximization in high-dimensional space that may not be easily amenable to traditional methods of estimation. To reduce the complexity of the problem, the proposed models are formulated using a hierarchical Bayesian framework, and the ability and structural parameters (i.e., correlation matrix and regression coefficients) are estimated via Markov chain Monte Carlo (MCMC) simulation.

Higher-Order Models

IRT Models as Higher-Order Models

Many research settings involve measured observations that have hierarchical structures (Draper, 1995; Kreft & de Leeuw, 1998). These structures can be integrated in a single framework using modeling approaches that allow specification of different models at the different levels of the hierarchy. Examples of such an approach are IRT models. Essentially, IRT models integrate two models specified at two levels. At the first level is the item response function that relates the examinee's ability and the item characteristics to the probability of a particular response; at the second level is the distribution function that characterizes how the ability is distributed in the population. One can view the former as modeling the within-person variability and the latter as modeling the between-person variability (Adams, Wilson, & Wu, 1997).

In the usual formulation, the item response function is given by $p(\mathbf{x}|\theta, \boldsymbol{\beta})$, where $\mathbf{x}$ denotes the response vector, θ is the examinee's ability, and $\boldsymbol{\beta}$ is the vector of item parameters. The density function of ability is written as $p(\theta|\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ is the vector of distribution parameters.

For a randomly sampled θ, the conditional probability of response vector $\mathbf{x}$ can be obtained by integrating out the ability parameter. That is, the two levels of the models are combined as

$$p(\mathbf{x}|\boldsymbol{\beta}) = \int p(\mathbf{x}|\theta, \boldsymbol{\beta}) p(\theta|\boldsymbol{\omega}) d\theta. \tag{1}$$

Whereas the first-level model is obvious (i.e., $p(\mathbf{x}|\theta, \boldsymbol{\beta})$), the second-level model needs some explanation. For examinee i, ability can be expressed as

$$\theta_i = \mu + \varepsilon_i, \tag{2}$$

where μ is the expected value of ability and ε_i follows some distribution. In the usual case where $\mu = 0$ and ε_i *iid* $N(0, 1)$, $p(\theta|\boldsymbol{\omega}) = N(0, 1)$, and is the same for all i. By not regressing the ability on some explanatory variables (i.e., keeping the examinees undifferentiated), Equation (2) represents the simplest way of modeling θ.

This model can be extended easily by proposing a linear regression model of $\boldsymbol{\theta}_i$ on $\mathbf{y}_i$,

$$\boldsymbol{\theta}_i = \mathbf{y}_i'\boldsymbol{\xi} + \varepsilon_i, \tag{3}$$

where the elements of $\mathbf{y}_i$ are known values and $\boldsymbol{\xi}$ is the vector of regression coefficients. With the same distributional assumption about ε_i as above, the more general expression in (3) leads to $p(\boldsymbol{\theta}_i|\boldsymbol{\omega}) = N(\mathbf{y}_i'\boldsymbol{\xi}, 1 - \psi^2)$, where ψ^2 is the proportion of variance in θ accounted for by the covariates. Subsequently, when dealing with a multidimensional space, the model can further be generalized to allow for the regression of multivariate latent variables on some known explanatory variables.

From a Bayesian perspective, $p(\theta|\boldsymbol{\omega})$ is simply the prior distribution of the latent ability. Estimating the parameters of the prior distribution as a function of some explanatory variables is a straightforward extension of the model. An early example of hierarchical IRT modeling from a Bayesian framework that uses covariates for the latent variables can be found in the work of Mislevy (1987). These covariates are viewed as auxiliary information that can be exploited to differentiate the prior distributions of the various groups of examinees.

Several studies have explicitly employed a hierarchical IRT paradigm. To assess both school and students simultaneously, Mislevy and Bock (1989) used a hierarchical approach to combine school-level and student-level effects in an IRT model. Adams, Wilson, and Wu (1997) used a class of models belonging to the Rasch family as the structural model in regressing the latent ability variables on several demographic variables, including sex, grade, and socioeconomic status. A more general approach was presented by Fox and Glas (2001). In their multilevel models, aside from predictor variables of the latent variable at the student level, predictor variables were used at the group level as well. Aside from covariates for the examinees, Patz and Junker (1999) described a generic hierarchical IRT model that incorporates a set of covariates as predictors of the item parameters. These covariates include the condition under which the tests were administered, differences in the wording of similar items, and characteristics of the raters. Bradlow, Wainer, and Wang (1999) and Du, Wainer, Bradlow, and Rogers (2001) used a hierarchical item response model to include a random effect for items nested on the same testlet to address the problem of dependencies among items from a common stimuli.

One advantage of IRT is its flexibility in incorporating into the model various information, both about the examinees and about the items. However, as more information is integrated, the model becomes too complex and unwieldy for traditional estimation methods that require maximization. For instance, with increasing model complexity, obtaining the derivatives required for maximization may become exceedingly prohibitive. Nonetheless, marginal maximum likelihood estimation is feasible when simpler IRT models (e.g., Rasch) are involved (Adams, Wilson, & Wang, 1997).

A method that can be used in conjunction with high-dimensional complex models is MCMC simulation. Although the problem to be solved need not be expressed in a Bayesian context, the results obtained from MCMC from a Bayesian formulation will allow any inference about the entire posterior distribution, not just about one of its characteristics (e.g., the mode). The dimensionality of the model can be greatly reduced if, in addition to item parameters, the latent regression parameters and the covariance structure of the latent traits, are known or if some approximations are employed to estimate these parameters (e.g., replacing the likelihood by a normal approximation to obtain a posterior distribution in closed form; Mislevy et al., 1992). Under these conditions, the use of MCMC and a hierarchical Bayesian formulation may not be necessary.

Parametrization and Estimation

Objectives

This study seeks to provide a general framework for ability estimation where ancillary information found in the covariates and correlational structure of the abilities can be incorporated in the estimation process using an integrated framework. The primary objective of this study was to investigate the magnitude of improvements in the ability estimates when responses to different tests and the ancillary information about the examinees are considered singly and jointly. In addition, the study also examined how the number of tests, length of test, the correlation between the different abilities, and the correlation between the ancillary variables and the latent trait affect the quality of the estimates. A full Bayesian hierarchical formulation was used to accommodate all the factors in the modeling process. The parameters were estimated using MCMC. The procedure that uses simultaneous estimation and ancillary information was compared to procedures that estimate abilities one at a time or ignores the ancillary variables. It should be noted that this approach is applicable not only when multiple tests are considered but also when scores across different subsections based on skills or objectives measured by a single test are of interest. In addition, although this article focuses on the three-parameter logistic (3PL) model, the framework was formulated such that other item response models can be used in its place.

Model and Parameterization

The extension of the 3PL model to the multidimensional context (Reckase, 1997) is given by

$$P(X_{ij}=1|\boldsymbol{\theta}_i)=\gamma_j+(1-\gamma_j)\frac{\exp(\alpha_j'\boldsymbol{\theta}_i+\beta_j)}{1+\exp(\alpha_j'\boldsymbol{\theta}_i+\beta_j)}. \tag{4}$$

For this article, the independent-cluster assumption is made (i.e., each item measures one ability, and thus, α_j contains only one non-zero element). Under this assumption, the tests taken together can be viewed as a multi-unidimensional test.

The model above, (4), can be re-expressed as follows:

$$P(X_{ij(d)}=1|\theta_{i(d)})=\gamma_{j(d)}+(1-\gamma_{j(d)})\frac{\exp(\alpha_{j(d)}\theta_{i(d)}+\beta_{j(d)})}{1+\exp(\alpha_{j(d)}\theta_{i(d)}+\beta_{j(d)})}, \tag{5}$$

where

$X_{ij}(d)$ is the response of examinee i to the jth item of dimension d;
$\alpha_j(d)$, $\beta_j(d)$, and $\gamma_j(d)$ are the discrimination, difficulty-related, and guessing parameters;
$\theta_i(d)$ is the dth component of the vector $\boldsymbol{\theta}_i$ (i.e., $\boldsymbol{\theta}_i=\{\theta_{i(d)}\}$);
$d=1,\ldots,D$ (the number of dimensions or tests);
$j(d)=1,\ldots,J(d)$; and
$\sum_{d=1}^{D}J(d)=J$.

Let $\mathbf{X}$ and $\boldsymbol{\Theta}$ be the matrices of item responses and ability parameters, respectively; then the likelihood of the data is given by

$$p(\mathbf{X}|\boldsymbol{\Theta})=\prod_{i=1}^{I}\prod_{d=1}^{D}\prod_{j(d)=1}^{J(d)}p(X_{ij(d)}|\theta_{i(D)}), \tag{6}$$

where

$$p(X_{ij(d)}|\theta_{i(d)}) = (P(X_{ij(d)} = 1|\theta_{i(d)})^{X_{ij}(d)}(1 - P(X_{ij(d)} = 1|\theta_{i(d)}))^{1-X_{ij(d)}}. \tag{7}$$

Prior Distributions

The prior distributions of the ability parameters are given below. For examinee i with ability $\boldsymbol{\theta}_i$,

$$\boldsymbol{\theta}_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma}, \mathbf{y}_i \sim MVN_D(\boldsymbol{\Xi}'\mathbf{y}_i, \boldsymbol{\Sigma}), \tag{8}$$

$$\boldsymbol{\Sigma} \sim Inv - Wishart_{\nu_0}(\boldsymbol{\Lambda}_0^{-1}), \tag{9}$$

$$\boldsymbol{\Xi} \sim (-\infty, +\infty), \tag{10}$$

where $\mathbf{y}_i = \{Y_{i1}, Y_{i2}, \ldots, Y_{ip}\}$ is the vector the p observable covariates of examinee i; $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_d, \ldots, \boldsymbol{\xi}_D\}$ is the $p \times D$ matrix of regression parameters; $\mu_i = \boldsymbol{\Xi}'\mathbf{y}_i$ and $\boldsymbol{\Sigma}$ are the mean vector and the common (i.e., undifferentiated by examinee) covariance matrix of the multivariate normal distribution, respectively; ν_0 are the degrees of freedom, and $\boldsymbol{\Lambda}_0$ is the $D \times D$ symmetric positive-definite scale matrix of the inverse-Wishart distribution. A uniform, but improper, prior was used for $\boldsymbol{\Xi}$ to facilitate sampling from the posterior distribution.

The functional forms of the prior distributions are chosen out of convenience, and the associated hyperparameters are selected to be reasonably vague within the range of realistic item parameters. Figure 1 gives the directed acyclic diagram of the response of examinee i on item j of the dth test. The variables inside the boxes are known.

Joint and Conditional Posterior Distributions

In addition to $\mathbf{X}$ and $\boldsymbol{\Theta}$, let $\mathbf{Y}$ be the matrix of covariates (i.e., design matrix). The joint distribution of $\boldsymbol{\Theta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Xi}$ given $\mathbf{X}$ and $\mathbf{Y}$ is

$$p(\boldsymbol{\Theta}, \boldsymbol{\Xi}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{X}|\boldsymbol{\Theta})\ p(\boldsymbol{\Theta}|\boldsymbol{\Xi}, \boldsymbol{\Sigma}, \mathbf{Y})p(\boldsymbol{\Sigma})p(\boldsymbol{\Xi}). \tag{11}$$

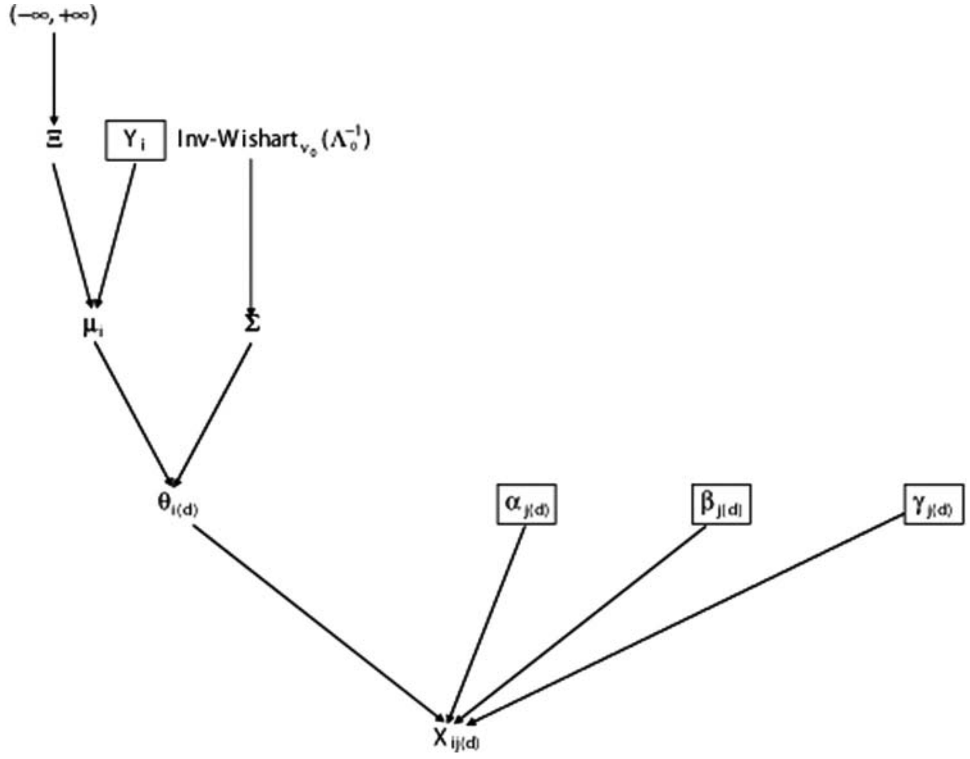
Below are the full conditional distributions of $\boldsymbol{\Theta}$, $\boldsymbol{\Xi}$, and $\boldsymbol{\Sigma}$.
For the ability parameters, $\boldsymbol{\Theta}|\boldsymbol{\Xi}, \boldsymbol{\Sigma}, \mathbf{X}, \mathbf{Y}$,

$$p(\boldsymbol{\Theta}|\boldsymbol{\Xi}, \boldsymbol{\Sigma}, \mathbf{X}, \mathbf{Y}) \propto\ p(\mathbf{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{\Xi}, \boldsymbol{\Sigma}, \mathbf{Y}). \tag{12}$$

But because the examinees are considered random, the full conditional distribution of each examinee can be evaluated as

$$p(\boldsymbol{\theta}_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma}\ \mathbf{X}, \mathbf{Y}) \propto p(\mathbf{x}_i|\boldsymbol{\theta}_i)\ p(\boldsymbol{\theta}_i|\boldsymbol{\Xi}, \boldsymbol{\Sigma}, \mathbf{y}_i) \tag{13}$$

Figure 1
A Directed Acyclic Graph of the Model for Examinee *i* and Item *j* in Dimension *d*



Hence, (13) can be written as

$$p(\mathbf{\Theta}|\Xi, \mathbf{\Sigma}, \mathbf{X}, \mathbf{Y}) \propto \prod_{i=1}^{I} p(\mathbf{x}_i|\boldsymbol{\theta}_i)\; p(\boldsymbol{\theta}_i|\Xi, \mathbf{\Sigma}, \mathbf{y}_i). \tag{14}$$

For the covariance matrix, $\mathbf{\Sigma}|\mathbf{\Theta}, \Xi, \mathbf{X}, \mathbf{Y}$,

$$p(\mathbf{\Sigma}|\mathbf{\Theta}, \Xi, \mathbf{X}, \mathbf{Y}) \propto p(\mathbf{\Theta}|\Xi, \mathbf{\Sigma}, \mathbf{Y})\; p(\mathbf{\Sigma}). \tag{15}$$

For the regression parameters, $\Xi|\mathbf{\Theta}, \mathbf{\Sigma}, \mathbf{X}, \mathbf{Y}$,

$$p(\Xi|\mathbf{\Theta}, \mathbf{\Sigma}, \mathbf{X}, \mathbf{Y}) \propto \; p(\mathbf{\Theta}|\Xi, \mathbf{\Sigma}, \mathbf{Y})\; p(\Xi). \tag{16}$$

The above model specification simultaneously accounts for responses to other tests and ancillary variables in scoring examinees' responses. In the traditional method of scoring, where tests are scored one at a time and no covariates are used, $\Xi = 0$ and $\mathbf{\Sigma} = \mathrm{I}$. When tests are scored simultaneously without the benefit of ancillary variables, only $\mathbf{\Sigma}$ is estimated whereas $\mathbf{\Xi} = 0$. Finally, when ancillary variables are used to score one test at a time, only $\mathbf{\Xi}$ is estimated whereas $\mathbf{\Sigma}$ is treated as a diagonal matrix with the dth diagonal element equal to $1 - \psi_d^2$, the proportion of reduction in the variance of ability d due to the covariates. The framework presented in this article assumes that the item parameters have been previously calibrated and are treated as known. Hence, the metric of the ability estimates was determined by the scale of item parameters, which has a mean of 0 and standard deviation of 1.

Simulation Study

Design and Analysis

For the simulation study, the multi-unidimensional tests were composed of either two or five tests. Each test was made up of 10 or 20 items. For the purposes of this article, the former represented a short test, whereas the latter, a medium-length test. The relationship between the covariates and the underlying abilities was given by the squared multiple correlation. This is equivalent to the proportion of variance accounted for (PVAF), denoted by ψ, because the mean abilities as linear functions of covariates were considered. The levels of PVAF were 0.25 and 0.50. For all conditions, two independent covariates from a standard normal distribution were used. Finally, two different levels of correlation between the abilities were studied: 0.50 and 0.90.

The different levels of the factors were crossed completely to yield 16 conditions. Of the 1 billion 10-item tests randomly constructed from a pool of 550 nationally standardized mathematics items, a 10-item test was selected that provided a mean information function that was closest to the mean information function of the entire pool in the following sense:

$$\int [\bar{I}_t(\theta) - \bar{I}_T(\theta)]^2 g(\theta) d\theta, \tag{17}$$

where

$\bar{I}_t(\theta)$ is the mean information of a 10-item test at a particular θ,
$\bar{I}_T(\theta)$ is the mean information of the entire pool at the same θ, and
$g(\theta)$ is the standard normal density.

These items were used for the 10-item condition. To ensure that the two test lengths had the same average quality (i.e., the mean item information functions were identical), the 10 items were replicated to create the 20-item test.

To fix the metric of $\boldsymbol{\theta}$ and to facilitate the comparisons between the different methods of estimation, the examinees were drawn from a multivariate normal distribution with the following two constraints: The marginal distributions of $\boldsymbol{\theta}$ are standard normal distributions, and the conditional correlation of $\boldsymbol{\theta}$ given $\mathbf{y}$ is ρ.

The mean vector of each examinee was a linear combination of the covariates where the weights were based on the levels of the PVAF. The simulation was designed so that the PVAF was the same for all the dimensions and was equal to the level of PVAF. Furthermore, because PVAF is simply the scaled sum of the squared regression coefficients, no unique set of the regression parameters exists except when PVAF is zero. Hence, for simplicity, the same weights were used for all the covariates. A common conditional covariance matrix for all the examinees was used. In addition, without loss of generality, the conditional correlations between all pairs of abilities were set to be the same and equal to the level of the correlation. Given $\boldsymbol{\theta}$ and the item parameters, responses to the tests were simulated. The sample size for this study was fixed at $I = 2{,}000$ for each condition.

In each condition, estimates were obtained using four different estimation methods representing instances of the general model: A, B, C, and D. Using Method A, covariation between the different abilities and the ancillary information were ignored in the estimation process. Method B estimated the latent traits using the correlational structure of abilities only, whereas Method C used the ancillary information only. Finally, Method D was a method of estimation that simultaneously used additional information that can be found in the correlational structure of the abilities and the ancillary variables. Method A represents the more extensively used method of estimation at present (i.e., abilities are estimated one test at a time and no covariates are used) and served as the baseline by which other methods were compared.

Parameter Estimation

Parameters for all the conditions were estimated using MCMC. Following is an outline of the MCMC algorithm for Method D.

Iteration 0:

1. Assign the following initial values to the parameters: $\boldsymbol{\Xi} = \mathbf{0}, \boldsymbol{\Sigma} = I$, and $\boldsymbol{\Theta}$, random draws from MVN ($\mathbf{0}$, I).

Iteration t:

2. For the regression parameters, the full conditional distribution of $\boldsymbol{\Xi}$, $p(\boldsymbol{\Xi}|\boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{Y})$, given an improper prior, is the matrix-normal MVN $((\mathbf{Y}'\ \mathbf{Y})^{-1}\ \mathbf{Y}'\boldsymbol{\Theta}, \boldsymbol{\Sigma} \otimes (\mathbf{Y}'\mathbf{Y})^{-1})$. This allows $\boldsymbol{\Xi}^{(t)}$ to be sampled directly from $p(\boldsymbol{\Xi}|\boldsymbol{\Theta}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}, \mathbf{Y})$.
3. The full conditional distribution of $\boldsymbol{\Sigma}, p(\boldsymbol{\Sigma}|\boldsymbol{\Xi}, \boldsymbol{\Theta}, \mathbf{Y})$, is an $Inv-Wishart_{\nu_I}(\boldsymbol{\Lambda}_I^{-1})$, where $v_I = v_0 + I$, and $\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_0 + \Sigma(\boldsymbol{\theta}_i - \boldsymbol{\Xi}^{'}\mathbf{Y})\ (\boldsymbol{\theta}_i - \boldsymbol{\Xi}'\mathbf{Y})'$ (Gelman, Carlin, Stern, & Rubin, 2003). Therefore, $\boldsymbol{\Sigma}^{(t)}$ can be sampled directly from $p(\boldsymbol{\Sigma}|\boldsymbol{\Xi}^{(t)}, \boldsymbol{\Theta}^{(t-1)}, \mathbf{Y})$.
4. Finally, since $\boldsymbol{\theta}$ has independent components, the sampling can be done one examinee at a time. For examinee i, sample $\boldsymbol{\theta}_i$* from MVN ($\boldsymbol{\theta}_i^{(t-1)}$, $\boldsymbol{\Sigma}_c$), where $\boldsymbol{\Sigma}_c$ is the fixed scale of the candidate-generating distribution. Accept the draw with probability

$$\alpha(\boldsymbol{\theta}_i^{(t-1)}, \boldsymbol{\theta}_i^{*}) = \min\left\{\frac{p(\mathbf{x}_i|\boldsymbol{\theta}_i^{*})\ p(\boldsymbol{\theta}_i^{*}|\boldsymbol{\Xi}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \mathbf{y}_i)}{p(\mathbf{x}_i|\boldsymbol{\theta}_i^{(t)})\ p(\boldsymbol{\theta}_i^{(t)}|\boldsymbol{\Xi}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \mathbf{y}_i)}, 1\right\}. \qquad (18)$$

To sample from $p(\boldsymbol{\Xi}\,|\,\boldsymbol{\Theta}, \boldsymbol{\Sigma}, \mathbf{Y})$ in vector format, one can use the following algorithm:

$$\mathrm{Vec}(\boldsymbol{\Xi}) = [\boldsymbol{\Sigma} \otimes (\mathbf{Y}'\mathbf{Y})^{-1}]^{1/2}\mathbf{Z} + \mathrm{Vec}((\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\boldsymbol{\Theta}), \tag{19}$$

where $\mathbf{Z}_{(D \times P) \times 1} \sim MVN(\mathbf{0}, \mathrm{I})$, and $\mathrm{Vec}(\cdot)$ stacks the vectors of the argument.

Gelman, Carlin, Stern, and Rubin (1995, p. 409) provide an alternative method of sampling from this full conditional distribution that avoids the use of the Kronecker product. By recasting the matrices as follows,

$$\begin{aligned}
\boldsymbol{\Xi}^*_{(D \times P) \times 1} &= \mathrm{Vec}(\boldsymbol{\Xi}),\\
\boldsymbol{\Theta}^*_{(I \times D) \times 1} &= \mathrm{Vec}(\boldsymbol{\Theta}),\\
\boldsymbol{\Sigma}^*_{(I \times D) \times (I \times D)} &= \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{pmatrix},\\
\mathbf{Y}^*_{(I \times D) \times (D \times P)} &= \begin{pmatrix} \mathbf{y}_1' & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{y}_1' & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{y}_1' \\ \mathbf{y}_2' & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{y}_2' & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{y}_2' \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{y}_I' & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{y}_I' & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{y}_I' \end{pmatrix},
\end{aligned}$$

where $\mathbf{y}_i$ is the covariate vector of examinee i (i.e., the transpose of the ith row of the design matrix $\mathbf{Y}$), one can sample from

$$\boldsymbol{\Xi}^* \sim MVN((\mathbf{Y}^{*\prime}\boldsymbol{\Sigma}^{*-1}\mathbf{Y}^*)^{-1}\mathbf{Y}^{*\prime}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\Theta}^*, (\mathbf{Y}^{*\prime}\boldsymbol{\Sigma}^{*-1}\mathbf{Y}^*)^{-1}). \tag{20}$$

They also suggested the use of matrix factorization to avoid inversion of large matrices in this algorithm.

The MCMC algorithms for Methods A, B, and C can be obtained from the algorithm above by constraining $\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ as described in the Design and Analysis section. The following hyperparameters were used for the covariance matrix: $v_0 = D + 2$, and $\mathbf{\Lambda}_0 = 0.5I_{D \times D} + 0.5$. For better numerical stability, the acceptance probabilities were computed in the log scale (Carlin & Louis, 2000). For each condition, 10,000 iterations were run with the first 2,000 iterates serving as the burn-in. The estimates of the parameters were based on the mean of the remaining observations. For example, $\boldsymbol{\theta}_i$ was estimated as

$$\hat{\boldsymbol{\theta}}_i = \sum_{t=2001}^{10000} \boldsymbol{\theta}_i^{(t)}. \tag{21}$$

$\boldsymbol{\Xi}$ and $\boldsymbol{\Sigma}$ and were estimated in the same manner. To allow direct sampling from the posterior distribution using an Inverse-Wishart distribution, the MCMC draws of $\boldsymbol{\Sigma}$ were based on the covariance matrix. However, after obtaining the sample mean of the draws, the covariance matrix was standardized to obtain an estimate of the correlation matrix. In addition to graphical methods, it was verified using the criterion suggested by Raftery and Lewis (1992), as implemented in the software CODA (Best, Cowles, & Vines, 1995), that the chain length used in conjunction with the algorithm for estimating the parameters of the most complex model, Method D, was sufficiently long.

Results

Estimation of Regression Parameters

Estimates of the regression coefficient, ξ, correlation between abilities, ρ, and ability, $\boldsymbol{\theta}$, were compared with the generating parameters to determine the accuracy of the estimation methods in recovering the parameters. Estimation of the coefficients in regressing the ability on the covariates involved comparison of Methods C and D, the two estimation methods that incorporated the auxiliary information about the examinees found in the covariates. The design of the study resulted in a matrix of regression parameters that have the same coefficients. Hence, comparison based on a single entry in the matrix of regression coefficients was sufficient. For the different levels of ψ, the regression parameter ξ can be computed as $\xi = \sqrt{\psi/p}$, where p is the number of covariates. Hence, for $\psi = 0.25$, the regression coefficient to be estimated is $\xi = 0.35$, whereas the regression coefficient for $\psi = 0.50$ is $\xi = 0.50$. The mean biases and absolute deviations from these ξ values are the same for Methods C and D: 0.00 and 0.02 when $\psi = 0.25$ and 0.01 and 0.00 when $\psi = 0.50$. These results indicate that the latent regression parameters can be accurately estimated by either Method C or D, particularly when the covariates can account for a large proportion of the variability in the abilities being estimated.

Estimation of Correlation Between Abilities

Although both Methods B and D used the information contained in the correlational structure of the abilities in the estimation process, the two methods estimate different correlation coefficients. Method D estimates the partial or conditional correlation of the abilities given the covariates, whereas Method B estimates the marginal correlation of the

Table 1
Correlation Between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$

		$\psi = 0.25$		$\psi = 0.50$	
	Method	$\rho = 0.50$	$\rho = 0.90$	$\rho = 0.50$	$\rho = 0.90$
$D = 2, J = 10$	A	0.84	0.83	0.85	0.84
	B	0.86	0.89	0.87	0.90
	C	0.86	0.85	0.88	0.88
	D	0.87	0.90	0.89	0.91
$D = 2, J = 20$	A	0.92	0.90	0.91	0.91
	B	0.92	0.93	0.92	0.94
	C	0.92	0.91	0.92	0.93
	D	0.93	0.93	0.93	0.95
$D = 5, J = 10$	A	0.85	0.85	0.85	0.85
	B	0.88	0.93	0.90	0.94
	C	0.86	0.86	0.89	0.88
	D	0.88	0.93	0.91	0.94
$D = 5, J = 20$	A	0.92	0.91	0.91	0.91
	B	0.93	0.96	0.93	0.96
	C	0.92	0.92	0.92	0.93
	D	0.93	0.96	0.93	0.97

abilities. The conditional correlation of $\boldsymbol{\theta}$ given $\mathbf{Y}$ is ρ regardless of the value of ψ; the marginal correlation of $\boldsymbol{\theta}$ for specific values of ψ and ρ is given by $\rho_{\theta} = \rho + \psi - \rho \times \psi$. That is, the marginal correlations are 0.63 and 0.73 when ψ is 0.25 and 0.50, respectively, for $\rho = 0.50$. These values are 0.93 and 0.95 for $\rho = 0.90$.

By design, both the marginal and conditional correlation matrices of $\boldsymbol{\theta}$ have compound symmetry structures; hence, investigation of a single coefficient sufficed. Results show that both the marginal and the conditional correlations can be estimated accurately by Methods B and D, respectively. When $\rho = 0.50$, the mean biases and absolute deviations of the two methods are not greater than 0.01 and 0.02, respectively. The same results were obtained when $\rho = 0.90$.

Ability Estimation

All the four methods were used to estimate the abilities of the examinees. The symmetry of the design allows for examination of the abilities on a single dimension. Various measures were computed to compare the quality of the ability estimates obtained using the different methods. These measures are correlation between θ and $\hat{\theta}$, root mean squared error of $\hat{\theta}$, posterior standard deviation of the estimates, and efficiency of the methods in estimating θ.

The correlations between the true and estimated abilities are given in Table 1. The results show that the use of either the correlational structure of the abilities or the covariates provided better ability estimates compared to the method that ignored this information. In addition, additional improvement can be achieved by simultaneously using these different sources of information.

Using Method A, the baseline correlations between true and estimated abilities were 0.85 and 0.91, when a 10-item test and a 20-item test are used, respectively. The quality of estimates was not expected to change by varying the number of tests, degree of correlation between abilities, or proportion of variance accounted for by the covariates using Method A; the fluctuations in the estimated correlations between true and estimated abilities can be attributed to sampling variations. And as expected, better ability estimates were obtained when conditions involved longer tests. This was true not only for Method A but also for the other methods.

As anticipated, increasing the number of tests, which resulted in a larger correlation matrix, benefited both Methods B and D because the two methods use the correlational structure in estimating θ. The same impacts on Methods B and D can be seen in increasing ρ, the conditional correlation between the abilities. Another anticipated result was that larger ψ allowed for better ability estimates using Method C. Similarly, because Method D uses the covariates in addition to the correlational structure, additional improvements were obtained with larger ψ. The improvement provided by simultaneously using both sources of information was greater than the improvement provided by using only a single source of information. Because the effect of ψ appeared in the marginal correlation of θ, increasing ψ also results in higher marginal correlations and, hence, better estimates for Method B, albeit to a lesser extent compared to Method D. The degree of improvement provided by Methods B and D can be ordered and summarized according to the degree of marginal correlation between the abilities. Finally, Method B outperformed Method C in all conditions, except when the conditions involved two tests that were moderately correlated and the proportion of variance accounted for by the covariates was high—the information provided by the covariates far outweighed the meager information found in the correlational structure of the abilities.

Estimates of another measure of the quality of ability estimates, the posterior standard deviation, which can be interpreted as the standard error of the estimate, are given in Table 2. The mean posterior standard deviation for each method was also computed as $\sqrt{\sum_{i=0}^{I} V(\theta_{i(d)}|\mathbf{x}_i)/I}$. For a 10-item test, the baseline posterior standard deviation of the ability estimate was about 0.53, whereas the baseline posterior standard deviation of the ability estimate was about 0.41 for a 20-item test. The impacts of the different factors on the posterior standard deviation were largely similar to the impacts on correlation between true and estimated abilities.

To keep track of the bias and variability of the estimates, the mean squared error (MSE) of each method across the different conditions was computed. For a certain method, MSE is computed as $\sum_{i=0}^{I}(\theta_{i(d)} - \hat{\theta}_{i(d)})^2/I$. In addition, the discrepancy between the MSE of the estimates was quantified by computing the relative efficiency of the different methods of estimation. Relative efficiency was defined as the ratio between the baseline MSE (i.e., Method A) and the MSE of the different methods (i.e., Methods B, C, and D). The ratio indicates the relative test length needed using the baseline method to obtain the same precision of an alternative method. The MSE and the corresponding efficiencies are listed in Table 3.

In general, the different factors affected the MSE of the different methods in the same manner they affected the correlation between true and estimated abilities and the posterior standard

Table 2
Mean Posterior Standard Deviation of $\hat{\theta}$ Across the Examinees

		$\psi = 0.25$		$\psi = 0.50$	
	Method	$\rho = 0.50$	$\rho = 0.90$	Method	$\rho = 0.50$
$D = 2, J = 10$	A	0.53	0.53	0.53	0.53
	B	0.49	0.46	0.49	0.44
	C	0.51	0.52	0.47	0.46
	D	0.48	0.45	0.45	0.41
$D = 2, J = 20$	A	0.41	0.41	0.41	0.41
	B	0.39	0.35	0.39	0.38
	C	0.40	0.40	0.38	0.38
	D	0.39	0.35	0.37	0.33
$D = 5, J = 10$	A	0.53	0.53	0.53	0.53
	B	0.47	0.34	0.43	0.34
	C	0.50	0.51	0.46	0.48
	D	0.47	0.33	0.42	0.33
$D = 5, J = 20$	A	0.41	0.41	0.41	0.41
	B	0.38	0.30	0.36	0.26
	C	0.40	0.40	0.38	0.37
	D	0.38	0.29	0.35	0.26

deviation. It should also be noted that for the ability estimates, the posterior standard deviation, which can be computed for real data, and the root MSE, which can be computed only when the generating parameters are available, were very similar. This congruence can have practical implications when relative efficiency needs to be computed for real data.

Relative efficiency allows for the comparison of the four methods across the different conditions. The best result was obtained when multiple ($D = 5$) tests that measure highly correlated abilities ($\rho = 0.90$) and covariates that highly correlate with the abilities ($\psi = 0.50$) were available. The highest efficiency, in the vicinity of 2.60, a 160% improvement, was obtained using Method D under this condition. This means that the precision of ability estimates from a 10-item test using Method D is equivalent to the precision of ability estimates from a 26-item test using Method A. For multiple highly correlated tests (i.e., $D = 5$ and $\rho = 0.90$), the relative efficiency of Method D was never lower than 2.00. This is equivalent to adding 12 to 16 items to a short (10-item) test or 21 to 30 items to a medium-length (20-item) test. Finally, for all conditions, Method B or C provided a lower bound for the efficiency of Method D relative to Method A.

Method B was almost as efficient as Method D when multiple tests measuring highly correlated abilities were involved and provided efficiencies that were greater than or equal to 2.00. These efficiencies are equivalent to adding 11 to 15 items to a short test and 20 to 28 items to a medium-length test. Under the ideal condition in this study, the highest efficiency using Method B is 2.48.

Although Method C did not afford the same magnitude of efficiency as Methods B and D, improvement was achieved by using Method C, particularly when the ability and covariates had moderately high correlation, (i.e., $\psi = 0.50$ for this study) and the tests were short. Under these conditions, the efficiency of Method C was at least 125%, which is

Table 3
Mean Squared Error of $\hat{\theta}$ Across the Examinees (Relative Efficiency in Parentheses)

		$\psi = 0.25$		$\psi = 0.50$	
	Method	$\rho = 0.50$	$\rho = 0.90$	Method	$\rho = 0.50$
$D = 2, J = 10$	A	0.29 (—)	0.30 (—)	0.27 (—)	0.29 (—)
	B	0.26 (1.12)	0.21 (1.48)	0.24 (1.16)	0.19 (1.50)
	C	0.26 (1.11)	0.27 (1.11)	0.22 (1.26)	0.22 (1.30)
	D	0.25 (1.18)	0.19 (1.57)	0.21 (1.33)	0.16 (1.76)
$D = 2, J = 20$	A	0.16 (—)	0.18 (—)	0.18 (—)	0.17 (—)
	B	0.15 (1.06)	0.13 (1.42)	0.16 (1.12)	0.11 (1.48)
	C	0.15 (1.04)	0.17 (1.06)	0.15 (1.17)	0.14 (1.18)
	D	0.15 (1.09)	0.13 (1.45)	0.14 (1.23)	0.10 (1.62)
$D = 5, J = 10$	A	0.28 (—)	0.28 (—)	0.28 (—)	0.28 (—)
	B	0.22 (1.28)	0.13 (2.12)	0.19 (1.43)	0.11 (2.48)
	C	0.26 (1.11)	0.26 (1.09)	0.21 (1.28)	0.22 (1.27)
	D	0.22 (1.30)	0.13 (2.15)	0.18 (1.51)	0.11 (2.59)
$D = 5, J = 20$	A	0.16 (—)	0.17 (—)	0.18 (—)	0.17 (—)
	B	0.14 (1.19)	0.09 (2.01)	0.13 (1.31)	0.07 (2.40)
	C	0.15 (1.07)	0.16 (1.08)	0.15 (1.20)	0.14 (1.20)
	D	0.13 (1.20)	0.08 (2.06)	0.13 (1.37)	0.07 (2.51)

equivalent to adding about three items to the tests. Furthermore, when $\rho = 0.50$ under the same conditions, using the covariates provided greater improvements than using the tests simultaneously in scoring the examinees (i.e., Method B).

Example

Data Description and Analysis

The data analyzed in this study were obtained from CTB/McGraw-Hill and consisted of 41 math and 36 science items taken by 1,500 Grade 8 students. For illustration purposes, two 10-item math and two 10-item science tests were constructed on the basis of the particular objectives of the items, whereas the remaining items were used to estimate two separate abilities, math (θ_{M0}) and science (θ_{S0}), which were used as covariates. In addition, a transformed variable indicating the percentage of adult residents in the school zip code with a bachelor's degree or higher, %BD, was used as the third covariate. The three covariates were standardized. It should be noted that this example is contrived and was used primarily to illustrate the proposed method.

For this example, the same formulation of MCMC as in the previous section was employed to analyze the data using the four methods, except the chains were run for 25,000 iterations, and inferences were based on the last 20,000 draws. Diagnostics performed on the chains indicated that although the chains were not as efficient as those in the simulation study, the burn-in and chain length were sufficiently long for the chains to have reached stationarity and for the Monte Carlo error to be negligible.

Table 4
Estimates of the Regression Coefficients

		Ability			
Method	Covariate	θ_{M1}	θ_{M2}	θ_{S1}	θ_{S2}
C	θ_{M0}	0.073	0.085	0.058	0.061
	θ_{S0}	0.21	0.21	0.47	0.55
	%BD	0.05	–0.03	–0.01	–0.02
D	θ_{M0}	0.75	0.86	0.59	0.63
	θ_{S0}	0.23	0.22	0.47	0.58
	%BD	0.05	–0.03	–0.01	–0.02

Note: %BD = percentage of adult residents in the school zip code with a bachelor's degree or higher.

Table 5
Estimates of the Correlation Between Abilities

	Ability Pair					
Correlation (Method)	$\{\theta_{M1}, \theta_{M2}\}$	$\{\theta_{M1}, \theta_{S1}\}$	$\{\theta_{M1}, \theta_{S2}\}$	$\{\theta_{M2}, \theta_{S1}\}$	$\{\theta_{M2}, \theta_{S2}\}$	$\{\theta_{S1}, \theta_{S2}\}$
Marginal (B)	0.95	0.84	0.85	0.76	0.77	0.94
Conditional (D)	0.78	0.49	0.55	0.31	0.35	0.81

Results

The following subsections give the results obtained using the different estimation methods. Measures used in the simulation study that are applicable to real data were computed.

Estimation Regression Parameters

Table 4 gives the latent regression parameters obtained using Methods C and D for this example. The results show that the estimates obtained using the two methods were very similar to each other. Unlike the simulation study, the covariates were not designed to be orthogonal. Consequently, the regression parameters do not have straightforward interpretations. However, these estimates still indicate that compared to the other two covariates, %BD had the smaller partial correlations with the abilities being estimated. In contrast, θ_{M0} had the highest partial correlations of the three covariates. Based on these estimates, the covariates as a whole account for approximately 91%, 88%, 77%, and 77% of the variabilities in θ_{M1}, θ_{Ms}, θ_{S1}, and θ_{S2}, respectively.

Estimation of Correlation Between Abilities

Estimates of the correlations between the four abilities obtained using Methods B and D are given in Table 5. Again, Method B estimates the marginal correlation between the abilities, whereas Method D estimates the partial correlation between the abilities given the covariates. Because of the sizeable impact of the covariates, the marginal correlations

Table 6
Mean Posterior Standard Deviation of $\hat{\theta}$ (Approximate Relative Efficiency in Parentheses)

	Content Area			
Method	Math 1	Math 2	Science 1	Science 2
A	0.52 (—)	0.60 (—)	0.65 (—)	0.60 (—)
B	0.41 (1.67)	0.50 (1.49)	0.52 (1.56)	0.56 (1.17)
C	0.33 (2.50)	0.44 (1.92)	0.55 (1.35)	0.54 (1.23)
D	0.31 (2.80)	0.43 (1.99)	0.49 (1.71)	0.53 (1.31)

are much higher than the conditional correlations. The marginal correlation between similar abilities is about 0.94, whereas the conditional correlation is 0.80. Finally, the marginal and conditional correlations of differing abilities are approximately 0.81 and 0.42, respectively. It can be noted that of the four abilities, θ_{M1} has the highest marginal correlations with other abilities.

Ability Estimation

The four methods were used in estimating the different abilities. The simulation study indicated that for ability estimates, the root MSE and the posterior standard deviation are similar. Hence, in the absence of the MSE, the ratio of the posterior variances was used as an approximate measure of the relative efficiencies of the alternative methods compared to Method A. The mean posterior deviations across the examinees and the approximate relative efficiencies are given in Table 6.

Some results of this example are similar to the results obtained in the simulation study. In particular, the mean posterior standard deviation of the estimates using Method A was always larger than those of the alternative methods. In addition, the largest improvement in ability estimation is provided by Method D across the four content areas.

The math abilities benefited more from the use of the correlational structure and the ancillary information compared to the science abilities. These improvements are equivalent to 11 to 15 additional items. Finally, because the covariates accounted for a large proportion of the variability in the math abilities, the efficiencies of Method C in these areas were higher than those of Method B.

Summary and Discussion

In most testing situations, in addition to responses to a specific test that measures an examinee's particular ability, other sources of information about the examinee's ability are commonly available. However, for various reasons, which include computational complexities and validity issues, these ancillary sources of information are not usually used for ability estimation purposes. The general formulation of the model proposed in this study incorporates the information obtained from demographic and educational variables and responses to other tests into a cohesive model for ability estimation. The approach proposed in the article is a general framework that encompasses the traditional approach of estimating abilities one at a time solely from the appropriate test responses. The simulation section of this study has shown that although the model is more complex in terms of its formulation and estimation, worthwhile gains can be obtained from considering such an approach.

Varying the formulations of the proposed general model allows for the two sources of ancillary information to be incorporated in the ability estimation process. When the impact of covariates that relate to the abilities needs to be estimated, two methods, C and D, can be used. When the different correlations between the abilities are to be estimated, Method B can be used to estimate the marginal correlation between the abilities, whereas Method D can be used to estimate the conditional correlation of the abilities given the covariates. Results of the simulation study indicate that by using the formulations and algorithms developed in this article, the parameters pertaining to the regression parameters and the correlational structures can be accurately estimated.

As expected, incorporating the ancillary variables and correlational structure in the estimation process provided better ability estimates. Different measures, correlation of true and estimated abilities, posterior standard deviation, MSE, and relative efficiency consistently indicated that incorporating either one or both sources of information can provide better results. De la Torre and Patz (2005) have shown that employing multidimensional scoring can further reduce the bias and standard error of the estimates of traditional unidimensional expected a posteriori, which already have smaller bias and standard error compared to other methods of estimation (Kim & Nicewander, 1993; Thissen & Orlando, 2001). This study has shown that by using Method D, which incorporates ancillary variables on top of the correlational structure, additional improvements in ability estimates can be obtained, particularly when large proportions of variances can be accounted for by the covariates. Based on the simulation results, the optimal condition occurs when several short tests that measure highly correlated abilities also correlate highly with the covariates. Under this situation, Method D can have an efficiency of 2.59 relative to the baseline method. The simulation study and real data analyses indicate that the optimal improvement in ability estimates can be obtained when all sources of ancillary information are accounted for. Moreover, depending on the conditions involved, the use of the covariates is more efficient than the use of the correlational structure, and vice versa.

In addition to improvements to ability estimates, the hierarchical framework presented in this article allows for the direct estimation of the correlations between the abilities and the contribution of the ancillary variables. Aside from accounting for the separate impacts of different sources of ancillary information, this framework avoids a two-step approach to correlation estimation, which results in biased estimates (Little & Rubin, 1983; Mislevy, 1984; Segall, 1996).

The example showed that to the extent that it can be verified, the analysis of actual responses of examinees provided results very similar to those obtained in the simulation study. Consequently, in real testing situations, better ability estimates can be expected when various sources of ancillary information are incorporated in the estimation process, and this can have practical implications. Without incurring additional expense, and by simply using information already available (i.e., scoring the same data differently), tests can be shortened without loss of reliability.

Although abilities estimated from ancillary information have better statistical properties, they also have more complex interpretations. Therefore, the valid use of estimates obtained using this approach requires careful consideration. In particular, as other authors have suggested (Mislevy, 1987; Wainer et al., 2001), the intended use of the test needs to be taken into account. When examinees are not to be compared with other examinees (i.e., comparison is within examinees), the estimates obtained from the proposed approach may be helpful. For example, when profile scores are needed to evaluate examinees' relative strengths and weaknesses in the different areas of a domain, such estimates may be appropriate. The specificity of the profile scores provide diagnostic information that can be used for directing instruction and learning.

However, when test scores are to be used in a contest on a specific domain (i.e., comparison is between examinees), estimates that are characterized by both performance on the domain and other ancillary information may not be appropriate. In these situations, a sufficiently long test that would warrant students to be ranked reliably on the domain of interest without reliance on other information is necessary.

As pointed out by de la Torre and Patz (2005), the estimates from the proposed approach need not replace the traditional unidimensional ability estimates but rather complement it. Traditional estimates may be reported at the domain level, whereas the enhanced estimates may be reported at the skills or the objective level. Currently, although finer-grain scores provide useful information, the traditional method of estimation does not warrant reporting such scores because of their insufficient reliability stemming from the small numbers of items directly measuring the specific skills or objectives. As shown by this study, one can capitalize on the availability of ancillary information to produce more reliable scores (i.e., lower standard errors) using the same set of items.

The conditions in this study involved ancillary variables that are measured without error. However, in real data analysis, this may not always be true. Consequently, the measurement error in the covariates may result in regression parameter estimates that are not consistent (Neter, Kutner, Nachtsheim, & Wasserman, 1996; Weisberg, 1985). Although the actual data presented in this article produced reasonable estimates even when covariates were measured with error, the exact extent to which measurement error in the covariates affects the ability estimations needs to be investigated further.

The method proposed in this study focuses on one particular IRT model and covariates that relate only to examinees. However, the generality of the method allows the 3PL model to be replaced by other models, such as the graded response model (Samejima, 1969), the generalized partial credit model (Muraki, 1992), or the generalized graded unfolding model (Roberts, Donoghue, & Laughlin, 2000). Moreover, if necessary, a combination of models, say, 3PL and generalized partial credit, can be used in the same analysis. In addition, the approach can be easily extended to include item parameter estimation, particularly when the calibration sample is small. Moreover, covariates that pertain to the item can also be incorporated. Finally, instead of assuming independent clusters, the method can be further generalized by using multidimensional IRT models that allow tests to measure more than one ability.

The code corresponding to the algorithm developed in this article was written in Ox (Doornik, 2003), a matrix programming language. The console version of the program can be used without fee for academic research purposes. The code used in this study can be obtained from the author by request.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

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

Adams, R. J., & Wu, M. L. (2006). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57-76). New York: Springer-Verlag.

Best, N. G., Cowles, M. K., & Vines, S. K. (1995). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output (Version 0.30) [Computer software]. Cambridge, MA: MRC Biostatistics Unit.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis.* London: Chapman and Hall.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of MCMC in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295-311.

Doornik, J. A. (2003). Object-oriented matrix programming using Ox (Version 3.1) [Computer software]. London: Timberlake Consultants.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, *20*, 115-147.

Du, Z., Wainer, H., Bradlow, E. T., & Rogers, H. J. (2001, April). *Modeling conditional item dependencies with a three-parameter logistic testlet model.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 221-288.

Fox, J. P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package CIRT. *Journal of Statistical Software*, *20*, 1-14.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, R. B. (1995). *Bayesian data analysis.* London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, R. B. (2003). *Bayesian data analysis* (2nd ed.). London: Chapman and Hall.

Johnson, E. G., & Carlson, J. (1994). *The NAEP 1992 technical report* (Report No. 23-TR-20). Washington, DC: National Center for Education Statistics.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587-599.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*, *29*, 3-25.

Little, R. J. A., & Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician*, *37*, 218-220.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359-381.

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

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Eds.), *Multilevel analysis of education data* (pp. 57-74). San Diego, CA: Academic Press.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedure in NAEP. *Journal of Educational Statistics*, *17*, 131-154.

Mislevy, R. J., & Sheehan, K. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661-679.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models.* Chicago: Erwin.

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

Raftery, A. E., & Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, *7*, 493-497.

Reckase, M. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement*, *24*, 3-32.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *17*.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331-354.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., et al. (2001). Augmented scores: ‘‘Borrowing strength’’ to compute score based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-388). Mahwah, NJ: Lawrence Erlbaum.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*, 295-316. Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*, 116-136.

Weisberg, S. (1985) *Applied linear regression.* New York: Wiley.