Journal of Educational and Behavioral Statistics
September 2007, Vol. 32, No. 3, pp. 233–251
DOI: 10.3102/1076998607300422

An Importance Sampling EM Algorithm for Latent Regression Models

Matthias von Davier
Sandip Sinharay
Educational Testing Service

Reporting methods used in large-scale assessments such as the National Assessment of Educational Progress (NAEP) rely on latent regression models. To fit the latent regression model using the maximum likelihood estimation technique, multivariate integrals must be evaluated. In the computer program MGROUP used by the Educational Testing Service for fitting the latent regression model to data from NAEP and other assessments, the integral is computed either by numerical quadrature or approximated. CGROUP, the current operational version of MGROUP used in NAEP for problems with more than two dimensions, uses Laplace approximation that may not provide fully satisfactory results, especially if the number of items per scale is small. This article examines a stochastic expectation-maximization (EM) method that uses importance sampling to NAEP-like settings. A simulation study and a real data analysis show that the importance sampling EM method provides a viable alternative to CGROUP for fitting multivariate latent regression models.

Keywords: *latent regression; NAEP; stochastic integration*

Introduction

The National Assessment of Educational Progress (NAEP), the only regularly administered and congressionally mandated national assessment program (e.g., see Beaton & Zwick, 1992), is an ongoing survey of the academic achievement of school students in the United States in a number of subject areas such as reading, writing, and mathematics. For several reasons (e.g., Mislevy, Johnson, & Muraki, 1992; von Davier & Sinharay, 2004), NAEP reporting methods started using in 1984 a multilevel statistical model consisting of two components: (a) an item response theory (IRT) component at the first level and (b) a linear regression component at the second level (e.g., see Beaton, 1987; Mislevy et al., 1992).

The authors thank Andreas Oranje, John Mazzeo, Shelby Haberman, Alina von Davier, Neal Thomas, Ying Jin, Matthew Johnson, David Thissen, and the two anonymous reviewers for useful advice; Steve Isham for help with the data sets used in the analysis; and Kim Fryer for help with copyediting the article.

Other large-scale educational assessments such as the International Adult Literacy Study (Kirsch, 2001), Trends in Mathematics and Science Study (Martin & Kelly, 1996), and Progress in International Reading Literacy Study (Mullis, Martin, Gonzalez, & Kennedy, 2003) also adopted essentially the same model.

This model is often referred to as a *latent regression model.* An algorithm for estimating the parameters of this model is implemented in the MGROUP set of programs, which is an Educational Testing Service (ETS) product. MGROUP computes the maximum likelihood estimates of the parameters of the model using a version of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) suggested by Mislevy (1984, 1985). The algorithm requires the values of the posterior mean and the posterior standard deviation (*SD*) of the proficiency variable $\boldsymbol{\theta}$ for each examinee, computation of which involves integrating out the multivariate $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ has up to two dimensions, the integration is computed using numerical quadrature (e.g., Beaton, 1987) by the BGROUP version of the MGROUP program in operational settings in NAEP. When $\boldsymbol{\theta}$ has higher dimensions (which is true in a number of NAEP subject areas such as mathematics), no numerical integration routine is available operationally (although it may now be possible to perform numerical integration for higher dimensions; work on that is in progress), and an approximation of the integral is used. The CGROUP version of MGROUP, the current operational technique used in NAEP and other assessments since 1993, is based on the Laplace approximation (Kass & Steffey, 1989) that ignores the higher order derivatives of the examinee posterior distribution and may not provide accurate results, especially in higher dimensions. For example, a graphic plot for a data example in Thomas (1993) shows that CGROUP overestimates posterior variances for high-ability examinees. A number of extensions to the CGROUP version of MGROUP or proposals for alternative estimation methods have been suggested (e.g., Cohen & Jiang, 1999; Johnson & Jenkins, 2004; von Davier, 2003). None of these alternatives have been entirely satisfactory (e.g., see von Davier & Sinharay, 2004, p. 3). Therefore, there is opportunity for further research in this area.

Statisticians have often used stochastic EM methods (Broniatowski, Celeux, & Diebolt, 1983; Celeux & Diebolt, 1985) in algorithms where the expectation required in the E step is hard to compute analytically or numerically. A stochastic EM algorithm is an EM algorithm that computes the expectation in the E step using simulation. Depending on the nature of the simulation used in the E step, a stochastic EM method can be a Monte Carlo EM (MCEM; e.g., Wei & Tanner, 1990), Markov chain MCEM (MCMCEM; e.g., McCulloch, 1994), rejection sampling EM (e.g., Booth & Hobert, 1999), importance sampling EM (ISEM; e.g., Booth & Hobert, 1999), and so forth. Important applications of the ISEM algorithm in psychometrics include Clarkson and Gonzalez (2001); Fox (2003); Lee, Song, and Lee (2003); and Meng and Schilling (1996), each of which use the Markov chain Monte Carlo algorithm (e.g., Gilks, Richardson, & Spiegelhalter, 1996) to simulate in the E step. The ISEM (e.g., Booth & Hobert, 1999) is a

stochastic EM method where integrals in the E step are approximated using *importance sampling* (Marshall, 1956), which approximates an integral by an average of a random sample. To our knowledge, the ISEM algorithm has not been used in psychometrics yet.

This article uses the ISEM method to fit the latent regression model used in NAEP-like applications. We present a comparison of the results from the ISEM algorithm with those from the CGROUP version of MGROUP, which is the technique currently used mostly in NAEP. For a low-dimensional real data example, the results from the suggested method are also compared to the BGROUP version of MGROUP, which uses numerical integration directly and may be viewed as the gold standard method in such settings. A simulation study and a real data analysis show that the ISEM method provides a viable alternative to CGROUP for fitting latent regression models.

The following section gives some background and describes the current NAEP statistical model and estimation procedure. We then review the ISEM method and discuss how it is applied to the NAEP estimation procedure. Next, we provide a simulated data and a real data example, respectively. Finally, the conclusion offers a summary and suggestions for future work.

The NAEP Model, Estimation, and the Current MGROUP Method

NAEP's Latent Regression Model

NAEP and other educational large-scale survey assessments implement a latent regression model that can be seen as a multilevel IRT model. Assume that the unique p-dimensional latent proficiency variable for examinee i is $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots \theta_{ip})'$. In NAEP, p could be between 1 and 5.

Let us denote the response vector to the test items for examinee i as $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{ip})$, where $\boldsymbol{y}_{ik}$, a vector of responses, contributes information about θ_{ik}. The likelihood for an examinee is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}_i) = \prod_{q=1}^{p} f_1(\boldsymbol{y}_{iq}|\theta_{iq}) \equiv L(\boldsymbol{\theta}_i; \boldsymbol{y}_i) \cdot \tag{1}$$

The expressions $f_1(\boldsymbol{y}_{iq}|\theta_{iq})$ above consist of terms contributed by a univariate IRT model, usually the two- or three-parameter logistic model for dichotomous items, and the generalized partial-credit model for polytomous items.

Suppose $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots x_{im})$ are m fully measured demographic and educational characteristics for the examinee. Conditional on $\boldsymbol{x}_i$, the examinee proficiency vector $\boldsymbol{\theta}_i$ is assumed to follow a multivariate normal prior distribution; that is,

$$\boldsymbol{\theta}_i|\boldsymbol{x}_i \sim N(\boldsymbol{\Gamma}'\boldsymbol{x}_i, \boldsymbol{\Sigma}). \tag{2}$$

Together, Equations 1 and 2 form the NAEP latent regression model or *conditioning model.*

NAEP's Estimation Process and the MGROUP Program

NAEP uses a three-stage estimation process for fitting the aforementioned latent regression model and making inferences. The first stage, *scaling*, fits the model given by Equation 1 to the examinee response data and estimates the item parameters. The prior distribution used in this step is not the one given by Equation 2; that is, the subscales are assumed to be independent a priori. The second stage, *conditioning*, has two parts. The first part assumes that the item parameters are fixed at the estimates found in scaling and fits the model given by Equations 1 and 2 to the data, that is, estimates $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$. In the second part of the conditioning step, plausible values (which are imputed values of the proficiency variables) for all examinees are obtained using the parameter estimates acquired in scaling and the first part of conditioning—the plausible values are used to estimate examinee subgroup averages. The third stage of the NAEP estimation process, called *variance estimation*, estimates the variances corresponding to the examinee subgroup averages using a jackknife approach (e.g., see Johnson & Jenkins, 2004). Our research focuses on the conditioning step and assumes that the scaling has already been done (i.e., the item parameters are fixed); this is the reason we suppress the dependence of Equation 1 on the item parameters.

Because we are concerned with the conditioning step, the remaining part of this section provides a more detailed discussion of it. Mislevy (1984, 1985) shows that the maximum likelihood estimates of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ can be obtained using an EM algorithm (Dempster et al., 1977). The EM algorithm iterates through a number of expectation steps (E steps) and maximization steps (M steps). Suppose $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n)'$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_n)'$. The expression for $(\mathbf{\Gamma}_{t+1}, \mathbf{\Sigma}_{t+1})$, the updated value of the parameters in the $(t+1)^{th}$ M step, is obtained as

$$\mathbf{\Gamma}_{t+1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\widetilde{\boldsymbol{\theta}_{1t}}, \widetilde{\boldsymbol{\theta}_{2t}}, \ldots, \widetilde{\boldsymbol{\theta}_{nt}})', \tag{3}$$

$$\mathbf{\Sigma}_{t+1} = \frac{1}{n}\left[\sum_i \mathrm{Var}(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t) + \sum_i (\widetilde{\boldsymbol{\theta}_{it}} - \mathbf{\Gamma}'_{t+1}\boldsymbol{x}_i)(\widetilde{\boldsymbol{\theta}_{it}} - \mathbf{\Gamma}'_{t+1}\boldsymbol{x}_i)'\right], \tag{4}$$

where $\widetilde{\boldsymbol{\theta}_{it}} = E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ and $\mathrm{Var}(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ are the posterior mean and variance of the proficiency variable of examinee i given the preliminary parameter estimates of iteration t. Correspondingly, the $(t+1)^{th}$ E step computes $E(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ and $\mathrm{Var}(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \mathbf{\Gamma}_t, \mathbf{\Sigma}_t)$ for the examinees. The process is repeated until convergence of the estimates $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$. The MGROUP set of programs at ETS performs the EM algorithm mentioned above.

There are different versions of the MGROUP program depending on the method used in the E step: BGROUP using numerical quadrature, NGROUP (Mislevy, 1985) using Bayesian normal theory, CGROUP (Thomas, 1993) using

Laplace approximations, and YGROUP (von Davier, 2003) using seemingly unrelated regressions (SURs; Zellner, 1962).

The BGROUP version of MGROUP, applied when the dimension of $\boldsymbol{\theta}_i$ is less than or equal to 2, applies numerical quadrature to approximate the integral. Thus, this version is the gold standard in MGROUP. When the dimension of $\boldsymbol{\theta}_i$ is larger than 2, CGROUP is the most appropriate and is used operationally in several large-scale assessments including NAEP. This approach uses the Laplace approximation, which involves a Taylor-series expansion of an integrand and ignores higher order derivatives of examinee posterior distributions, of the posterior mean and variance. Details about the method can be found in Thomas (1993, pp. 316–317). The Laplace method does not provide an unbiased estimate of the quantity it is approximating and may provide inaccurate results if higher order derivatives of the examinee posterior distributions (that the Laplace method assumes to be equal to zero) are not negligible. The error of approximation for each component of the mean and covariance of $\boldsymbol{\theta}_i$ is of order $O(\frac{1}{k^2})$ (e.g., Kass & Steffey, 1989), where k is the number of items measuring skill corresponding to the component. Because the number of items given to each examinee in large-scale assessments such as NAEP is not too large (making k rather small), the error in the Laplace approximation may become nonnegligible, especially for high-dimensional $\boldsymbol{\theta}_i$s. Furthermore, if the posterior distribution of $\boldsymbol{\theta}_i$s is multimodal (which is not impossible, especially for a small number of items), the method can perform poorly. Therefore, the CGROUP version of MGROUP is not entirely satisfactory. Figure 1 in Thomas (1993), where the posterior variance estimates of 500 randomly selected examinees using BGROUP and CGROUP for two-dimensional $\boldsymbol{\theta}_i$ are plotted, shows that the CGROUP provides inflated variance estimates for examinees with large posterior variance (see also the section on an analysis of a real NAEP data set later in this report). The departure may be more severe for $\boldsymbol{\theta}_i$s in higher dimensions.

Application of the ISEM Method to NAEP

Importance Sampling

Suppose we want to compute the expected value of a function $g(\boldsymbol{\omega})$ where the expectation is taken with respect to a probability density $f(\boldsymbol{\omega})$. One can express the expectation of $g(\boldsymbol{\omega})$ as

$$E_f[g(\boldsymbol{\omega})] = \int_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) f(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_{\boldsymbol{\omega}} \frac{g(\boldsymbol{\omega}) f(\boldsymbol{\omega})}{h(\boldsymbol{\omega})} h(\boldsymbol{\omega}) d\boldsymbol{\omega} \tag{5}$$

for any probability density $h(\boldsymbol{\omega})$ defined on the same sample space that is zero only if $f(\boldsymbol{\omega})$ is zero. If it is possible to generate a random sample $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_n$ from the distribution with density $h(\boldsymbol{\omega})$, the importance sampling method (Geweke, 1989; Marshall, 1956) suggests approximating $E_f[g(\boldsymbol{\omega})]$, using Equation 5, as

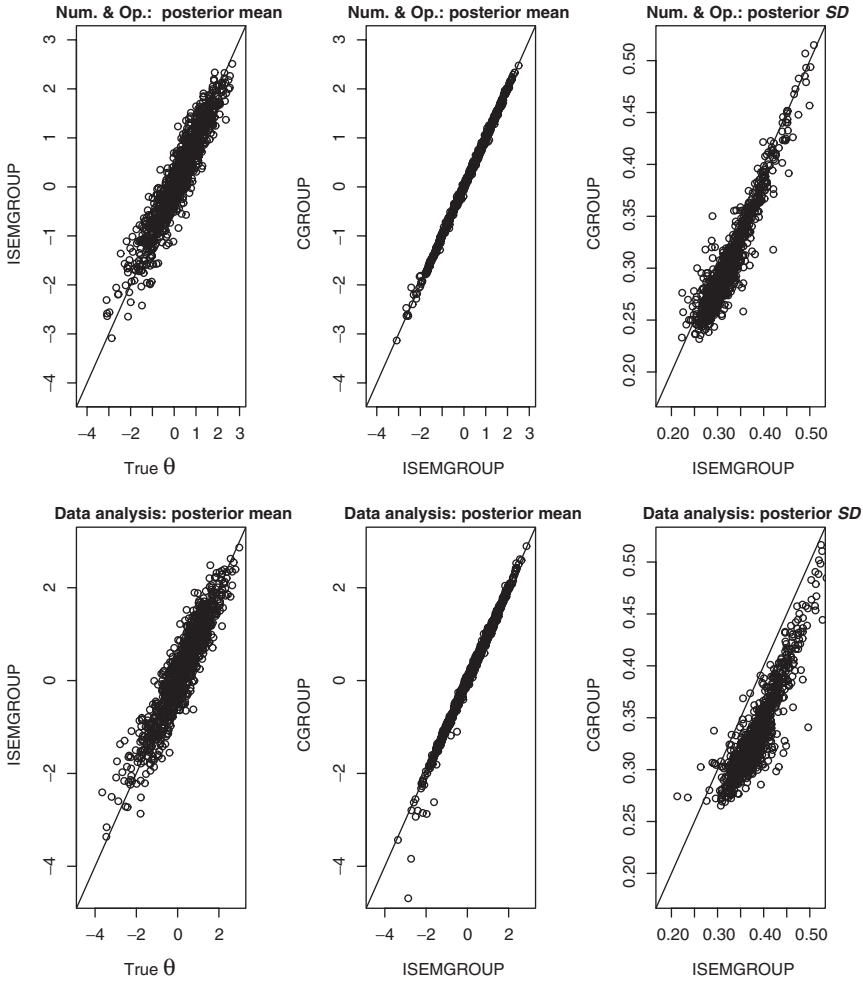

FIGURE 1. *Posterior moments from CGROUP and ISEMGROUP compared to the generating values and the posterior means and standard deviations (*SD*s) from CGROUP and ISEMGROUP compared to each other.*

$$E_f[g(\boldsymbol{\omega})] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{g(\boldsymbol{\omega}_i)f(\boldsymbol{\omega}_i)}{h(\boldsymbol{\omega}_i)}. \tag{6}$$

The standard error corresponding to the above estimate is obtained by dividing the *SD* of the importance ratios $\frac{g(\boldsymbol{\omega}_i)f(\boldsymbol{\omega}_i)}{h(\boldsymbol{\omega}_i)}$ by $\sqrt{n}$. The density $h(\boldsymbol{\omega})$ is called the *importance sampling density.*

If $h(\boldsymbol{\omega})$ is chosen so that the *importance ratio* $\frac{g(\boldsymbol{\omega})f(\boldsymbol{\omega})}{h(\boldsymbol{\omega})}$ is roughly constant across the range of possible values of $\boldsymbol{\omega}$ (which will make the *SD* of the importance ratios small), fairly accurate approximation of the integral may be obtained. In particular, $h(\boldsymbol{\omega})$ should have heavier tails than the product $g(\boldsymbol{\omega})f(\boldsymbol{\omega})$ because otherwise there may be a few $\boldsymbol{\omega}_i$s for which $h(\boldsymbol{\omega})$ will be much smaller than $g(\boldsymbol{\omega})f(\boldsymbol{\omega})$ and the estimate will be inaccurate.

The main advantages of importance sampling are that it provides an unbiased estimate of the integral and the accuracy can be monitored by examining the *SD* of the $\frac{g(\boldsymbol{\omega}_i)f(\boldsymbol{\omega}_i)}{h(\boldsymbol{\omega}_i)}$ values.

In some situations, the researcher knows a distribution only up to a constant but still needs to compute the moments of the distribution. For example, in a Bayesian analysis, the researcher often knows only a multiple of the posterior (by multiplying the likelihood and the prior), but not the posterior itself. Importance sampling can be used in this situation as well. Suppose, in the above setup, one does not know $f(\boldsymbol{\omega})$ but does know $q(\boldsymbol{\omega}) \equiv c \cdot f(\boldsymbol{\omega})$. The expectation of a function $g(\boldsymbol{\omega})$ is given by $E_f[g(\boldsymbol{\omega})] = \frac{\int_{\boldsymbol{\omega}} g(\boldsymbol{\omega})q(\boldsymbol{\omega})d\boldsymbol{\omega}}{\int_{\boldsymbol{\omega}} q(\boldsymbol{\omega})d\boldsymbol{\omega}}$. The numerator in the right-hand side the equation above is approximated by

$$\int_{\boldsymbol{\omega}} g(\boldsymbol{\omega})q(\boldsymbol{\omega})d\boldsymbol{\omega} = \int_{\boldsymbol{\omega}} \frac{g(\boldsymbol{\omega})q(\boldsymbol{\omega})}{h(\boldsymbol{\omega})} h(\boldsymbol{\omega})d\boldsymbol{\omega} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{g(\boldsymbol{\omega}_i)q(\boldsymbol{\omega}_i)}{h(\boldsymbol{\omega}_i)} = T_1,$$

whereas the denominator is approximated by $T_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{q(\boldsymbol{\omega}_i)}{h(\boldsymbol{\omega}_i)}$. Thus, $\frac{T_1}{T_2}$ provides an estimate of $E_f[g(\boldsymbol{\omega})]$. Geweke (1989) shows that $\frac{T_1}{T_2}$ (and, hence, the estimate in Equation 6, if applicable) converges almost surely to $E_f[g(\boldsymbol{\omega})]$ under weak assumptions and a central limit theorem, establishing that $n^{1/2}(\frac{T_1}{T_2} - E_f[g(\boldsymbol{\omega})]) \rightarrow \mathcal{N}(0, \tau^2)$, where τ^2 can be estimated consistently, applies under stronger assumptions.

Importance Sampling in the E Step of the MGROUP Program

The primary goal of this report is to approximate the posterior expectation and variance of the examinee proficiencies $\boldsymbol{\theta}_i$ in Equations 3 and 4 using the importance sampling method. The posterior distribution of $p(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}_t)$, is given by

$$p(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}_t) \propto f(\boldsymbol{y}_{i1}|\theta_{i1}) \ldots f(\boldsymbol{y}_{ip}|\theta_{ip})\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}_t'\boldsymbol{x}_i, \boldsymbol{\Sigma}_t), \tag{7}$$

using Equations 1 and 2. The proportionality constant in Equation 7 is a function of $\boldsymbol{y}_i$, $\boldsymbol{\Gamma}_t$, and $\boldsymbol{\Sigma}_t$. Let us denote

$$q(\boldsymbol{\theta}_i|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Gamma}_t, \boldsymbol{\Sigma}_t) \equiv f(\boldsymbol{y}_{i1}|\theta_{i1}) \ldots f(\boldsymbol{y}_{ip}|\theta_{ip})\phi(\boldsymbol{\theta}|\boldsymbol{\Gamma}_t'\boldsymbol{x}_i, \boldsymbol{\Sigma}_t) \tag{8}$$

We drop the subscript i for convenience for the rest of this section and let $\boldsymbol{\theta}$ denote the proficiency of an examinee.

We have to compute the mean and variance of $p(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)$, that is, the quantities

$$\widetilde{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t) \equiv \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t) d\boldsymbol{\theta}, \tag{9}$$

and

$$\text{Var}(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t) \equiv \int (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})' p(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t) d\boldsymbol{\theta} \cdot \tag{10}$$

Using ideas from the above description of the importance sampling method, if we can generate a random sample $\boldsymbol{\theta}^1,\ \boldsymbol{\theta}^2, \ldots,\ \boldsymbol{\theta}^n$ from a distribution $h(\boldsymbol{\theta})$ approximating $p(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)$ reasonably, we can approximate Equation 9 by the ratio of

$$U_1 = \frac{1}{n} \sum_{j=1}^{n} \frac{\boldsymbol{\theta}^j q(\boldsymbol{\theta}^j|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)}{h(\boldsymbol{\theta}^j)}$$

and

$$U_2 = \frac{1}{n} \sum_{j=1}^{n} \frac{q(\boldsymbol{\theta}^j|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)}{h(\boldsymbol{\theta}^j)}.$$

Similarly, we can approximate Equation 10 as the ratio of

$$U_3 = \frac{1}{n} \sum_{j=1}^{n} \frac{(\boldsymbol{\theta}^j - E(\boldsymbol{\theta}^j|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t))(\boldsymbol{\theta}^j - E(\boldsymbol{\theta}^j|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t))' q(\boldsymbol{\theta}^j|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)}{h(\boldsymbol{\theta}^j)}$$

and U_2.

Following Booth and Hobert (1999), who apply the ISEM method for fitting generalized linear mixed models, we use a multivariate t importance sampling density with small degrees of freedom; that is, $df = 4$. Booth and Hobert (1999) argue that in general, a multivariate t density is a better choice as an importance sampling density than a normal density. One reason is that the former is heavy tailed whereas the latter is not, and a good importance sampling density should be heavy tailed when the density it approximates may be so; also, the t density is easy to sample from. The degrees of freedom of the t density are taken to be 4 because low degrees of freedom of a t density ensure that it has a heavy tail.

(The choice of optimal degrees of freedom is an issue requiring further investigation.)

It is also possible to estimate variances of the parameter estimates obtained from the EM algorithm as suggested by, for example, Booth and Hobert (1999).

Note that a number of studies have successfully applied other forms of stochastic EM methods, such as rejection sampling EM (Booth & Hobert, 1999) and MCMCEM or MCMC-EM (e.g., McCulloch, 1994), but we do not explore those approaches, mainly because of the difficulties in their application to our problem.

Implementing ISEMGROUP

The ISEM algorithm was integrated into the YGROUP (von Davier, 2003) version of MGROUP, which performs SURs (Zellner, 1962) in the E step of the EM algorithm; this implementation will be referred to as the ISEMGROUP version of MGROUP henceforth. The posterior means and the posterior residual (co-)variance matrix obtained using SUR, computation of which is fast and efficient, are used as the mean and variance of the multivariate t_4 importance sampling density.

As in all versions of MGROUP, optional starting values for the latent regression $(\mathbf{\Gamma}, \mathbf{\Sigma})$ can be provided but are not essential. As an alternative, initial iterations with a nonstochastic E step of one of the other versions can be carried out to generate starting values. The initial sample size of the importance sample can be chosen to be much smaller than the final sample size, in order to speed up initial iterations where individual accuracy is less important than approaching a good starting value for the aggregate parameters. The examples below use 500 as the initial importance sample size and 6,000 as the final importance sample size.

Determining Convergence of the ISEM Algorithm

When applying an ISEM algorithm, it is possible that the EM algorithm has not converged but appears to have done so after an EM step (i.e., updated parameters and likelihoods do not change much from the previous step) because of an unlucky random sample. To ensure the convergence of the algorithm, one should therefore use a stricter convergence criterion or a larger variety of criteria that all have to be met simultaneously. We monitor the likelihood function and the parameter vector for convergence and conclude that convergence has occurred only when the relative change in both these quantities is less than ϵ in five successive iterations.

The likelihood increases for the iterations in a nonstochastic EM algorithm (e.g., see Dempster et al., 1977) whereas it may not be so for an ISEM algorithm because of Monte Carlo error. To assess convergence in stochastic optimization algorithms, the inevitable nonmonotonicity of these algorithms has to be taken

into account. Each E step in our implementation uses a different set of points (the importance sample) in the multivariate quadrature grid in order to perform the integration. This means that the likelihood of each observed response vector may vary; hence, the likelihood of the whole sample may vary even if all other parameters are unchanged from one importance sample to the next.

Therefore, in our implementation, we monitor absolute changes in the log likelihood, the regression parameters $\mathbf{\Gamma}$, and the residual (co-)variance matrix $\mathbf{\Sigma}$ simultaneously. The maximum absolute changes are checked against user-defined stopping criteria, which are used to decide whether a significant change has occurred that demands further iterations toward the optimum.

The algorithm stops if, and only if, all three absolute changes are smaller than user-defined numbers. In addition to that, the absolute changes of previous iterations are integrated with the current change by using the following approach: For cycle t, define the average absolute maximum change (AAMC) as $\overline{AAMC_{\boldsymbol{x},\,t}} = p \times AMC_{\boldsymbol{x},\,t} + (1-p) \times \overline{AAMC_{\boldsymbol{x},\,t-1}}$, which averages the current absolute maximum change (AMC) and the previous AAMC. The $\boldsymbol{x}$ denotes a parameter (vector) or a function of parameters (e.g., the maximum change in regression parameters in our case); t denotes the current iteration (cycle) of the algorithm; and $0 \leq (1-p) \leq 1.0$, the weight of the averaged previous cycles' criterion.

This criterion ensures that if $p < 1$, more than the current change (which might be small due to the stochastic nature of the algorithm) is taken into account when deciding whether to stop iterations. If $p = 1$, we have a regular (no memory) stopping criterion, which stops whenever the current iteration alone meets the stopping rule.

In the current implementation, the stopping rule fades out the past absolute changes rather slowly. We found that $p = .4$ and the AAMC bound of .045 for relative change in likelihood, regression weights, and variances work reasonably well. This choice balances infinite iterations against premature termination of the algorithm in the examples presented here. The rule for stopping the algorithm is that $\overline{AAMC_{\boldsymbol{x},\,t}} < 0.045$ has to be satisfied for termination, a maximum change bound similar to what Booth and Hobert (1999) report.

Analysis of a Simulated Data Set

This section presents the first proof-of-concept example. The simulated data set used for this example matches the structure and size of the 2000 NAEP mathematics assessment for Grade 4. The assessment has five scales: (a) Number and Operations, (b) Measurements, (c) Geometry, (d) Data Analysis, and (e) Algebra. This section applies the ISEM method to find the parameter estimates for the simulated data set, which involves responses of 13,511 students. Each student responds to 1 of 26 sets of items (booklets). Each booklet contains three blocks of 12–15 items (both multiple choice and constructed response) out of a total of 145

TABLE 1
Residual Variance–Covariance Estimates for the Simulated Data

	NumOp	Measurement	Geometry	DA	Algebra
CGROUP estimates					
NumOp	.326	.305	.329	.299	.324
Measurement	.938	.325	.319	.286	.312
Geometry	.946	.918	.371	.302	.336
DA	.925	.886	.875	.320	.306
Algebra	.948	.915	.922	.902	.359
ISEMGROUP estimates					
NumOp	.339	.298	.317	.291	.316
Measurement	.863	.351	.304	.277	.300
Geometry	.876	.824	.388	.290	.319
DA	.841	.786	.782	.355	.296
Algebra	.879	.820	.831	.804	.381

Note: NumOp = Number and Operations; DA = Data Analysis; Residual variances = main diagonals; covariances = upper off-diagonals; correlations = lower off-diagonals.

items. Each examinee has information on 381 predictors that are used in the latent regression model. The latent regression model is fitted to the data using the CGROUP and ISEMGROUP versions of the MGROUP separately.

Estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ for the Simulated Data

The correlation coefficients between the CGROUP and ISEMGROUP regression parameters estimates (i.e., estimates of the individual parameters in $\boldsymbol{\Gamma}$) for the five subscales are all above .99, which indicates the extreme closeness of the estimates from the two approaches.

Table 1 shows the residual variance–covariance estimates (i.e., estimates of the individual parameters in $\boldsymbol{\Sigma}$) from CGROUP and ISEMGROUP for the simulated data set. The table shows slightly larger estimated residual variances for ISEMGROUP as compared to CGROUP. The pattern is the opposite for the estimated residual covariances. As an outcome, the estimated correlations are comparatively lower for ISEMGROUP. The differences between the covariance estimates of CGROUP and ISEMGROUP seem smaller than the differences between the variance estimates of the two approaches.

Subgroup Estimates for the Simulated Data

Figure 1 presents recovery plots for 1,000 randomly selected examinees for ISEMGROUP for two subscales: (a) Number and Operations, which is the subscale covered by the greatest number of items, and (b) Data Analysis, which is the subscale covered by the least number of items.

In the recovery plots, the generating values of the proficiency variables are the reference, and the conditional posterior moments (given the responses, the background information, and the maximum likelihood estimates of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$, the moments are the outputs of the EM algorithm required to fit the latent regression model and do not use the plausible values) produced by ISEMGROUP are plotted against these reference values. The figure also shows, for the same examinees and the same two subscales, subplots comparing the examinee posterior means and *SD*s from CGROUP and ISEMGROUP. Each plot shows a diagonal line for convenience. The ISEMGROUP posterior means are highly correlated (correlation coefficient being .93 and .94 for the two subscales shown) with the generating values; the plots of CGROUP posterior means versus the generating values, not shown here, look very similar. The posterior means generated by CGROUP and ISEMGROUP fall nicely around the diagonal lines, and the plots are much narrower than the "truth" recovery plots. This indicates that CGROUP and ISEMGROUP agree quite well with regard to posterior means.

The situation is different for the posterior *SD*s. The posterior *SD*s given by ISEMGROUP are considerably larger than those for CGROUP for all subscales. It is to be determined whether this is a systematic difference, due to Monte Carlo inflation of the variance, or whether this depends on some other structural variable such as the correlation between scales.

Table 2 compares the subgroup means (obtained using the posterior means computed above, not using plausible values) from CGROUP and ISEMGROUP for relevant subgroups. There seems to be little difference between the two methods from this aspect as well.

It may be concluded that ISEMGROUP does not perform much differently from CGROUP when reproducing group-level statistics.

Analysis of a Real NAEP Data Set: 2003 Reading Assessment at Grade 4

The 2003 NAEP reading assessment at Grade 4 has data from 187,581 examinees in the fourth grade or 9 years of age. There are literary blocks (to assess the ability to read for literary experience) and informative blocks (to assess the ability to read for information). Each block consists of 10 to 12 items, with a mixture of multiple-choice and constructed-response items. Each student answers (a) a literary block and an informative block, (b) two literary blocks, or (c) two informative blocks. We will refer to the two skills (subscales) measured by the assessment as *literary* and *information.* The combined item sample for these two scales has 111 items in total; 688 background variables are used in the latent regression model. The latent regression model is fitted to the data using the BGROUP, the version utilizing numerical integration that serves as the gold standard here (and can be used here because the $\boldsymbol{\theta}_i$s are two-dimensional), the CGROUP, and the ISEMGROUP versions of MGROUP separately.

TABLE 2
Comparison of Subgroup Estimates From CGROUP and ISEMGROUP

	CGROUP					ISEMGROUP				
Subgroup	NumOp	Meas	Geom	DA	Alg	NumOp	Meas	Geom	DA	Alg
Overall	.033	.006	−.023	.047	.037	.031	.009	−.013	.060	.039
Male	.069	.083	−.043	.061	.089	.066	.086	−.030	.075	.092
Female	−.004	−.072	−.003	.032	−.016	−.006	−.071	.004	.045	−.016
White	.272	.310	.248	.361	.270	.270	.311	.254	.369	.269
Black	−.618	−.838	−.748	−.794	−.624	−.617	−.833	−.738	−.777	−.623
Hispanic	−.454	−.566	−.550	−.525	−.420	−.459	−.560	−.527	−.493	−.411
Asian	.605	.388	.532	.369	.680	.605	.392	.538	.382	.688
AmInd	−.407	−.287	−.693	−.515	−.429	−.412	−.271	−.630	−.154	.014

Note: NumOp = Number and Operations; Meas = Measurements; Geom = Geometry; DA = Data Analysis; Alg = Algebra; AmInd = American Indian.

Estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ for the Real Data

The correlation coefficients between the BGROUP and CGROUP regression parameters estimates for the two subscales are above .99985, and the same is true for the correlation coefficient between the BGROUP and ISEMGROUP regression parameters estimates. These numbers indicate that the $\hat{\boldsymbol{\Gamma}}$ estimates produced by BGROUP, CGROUP, and ISEMGROUP are virtually identical.

Table 3 shows the residual variance estimates $\hat{\boldsymbol{\Sigma}}$ as generated by BGROUP, ISEMGROUP, and CGROUP for the NAEP data. The three sets of estimates for the residual correlations and the residual variances are close. Interestingly, the estimates generated by ISEMGROUP are slightly smaller than CGROUP estimates for the reading data, whereas the residual variances estimates for the simulated math were larger for ISEMGROUP than for CGROUP. As an outcome, the correlation is higher for ISEMGROUP than for CGROUP (which is the opposite of what was found in the analysis of the simulated data set). Whether this is an effect of the different sample sizes or the number of scales needs further investigation.

Subgroup Estimates for the Real Data

Figure 2 presents plots of the posterior means produced by ISEMGROUP compared to the posterior means generated by BGROUP for 1,000 randomly selected examinees. Figure 3 compares the corresponding posterior *SD*s. The posterior moments were computed in the same manner as described in Analysis of the Simulated Data Set. There is a near-perfect agreement between the posterior means from BGROUP and CGROUP (correlations for the two subscales are .9996 and .9995; plots are not shown). Except for a few outliers, the posterior

TABLE 3
Residual Variances, Covariances, and Correlations for the 2003 NAEP Reading Assessment Grade 4 Data

	BGROUP		ISEMGROUP		CGROUP	
	Literary	Information	Literary	Information	Literary	Information
Literary	.506	.418	.485	.413	.515	.418
Information	.821	.512	.844	.494	.802	.526

Note: Residual variances = main diagonals; covariances = upper off-diagonals; correlations = lower off-diagonals.

means produced by ISEMGROUP agree closely with the values produced by BGROUP; correlations for the two subscales are .975 and .976. The outliers observed in these plots are subject to ongoing research aimed at optimizing and stabilizing the current implementation of ISEMGROUP. It will be investigated whether the stochastic integration failed to produce meaningful results due to poor coverage of the true posterior by the importance sampling distribution or whether these values are caused by other less obvious reasons.

In contrast to the first example (that involving simulated data), there is no obvious large difference between the estimates of the posterior means or *SD*s generated by CGROUP and ISEMGROUP. Both approaches differ slightly from the estimates generated by BGROUP. CGROUP overestimates the posterior *SD* for larger values when compared to BGROUP. This phenomenon was also observed in Figure 1 of Thomas (1993).

ISEMGROUP, on the other hand, stays closer to BGROUP for large posterior *SD*s than CGROUP does. For small *SD*s, ISEMGROUP seems to produce slightly larger values than BGROUP, whereas the estimates generated by CGROUP and BGROUP agree better with smaller *SD*s.

Table 4 shows the subgroup estimates and the corresponding *SD*s (in parentheses) provided by the three methods. There are hardly any differences between CGROUP and ISEMGROUP so far as subgroup estimates are concerned. However, when compared to BGROUP results, both CGROUP and ISEMGROUP slightly underestimate the subgroup means.

Conclusion and Future Work

CGROUP is the current operational method used in large-scale assessments such as NAEP. Even though CGROUP provides more accurate results than its predecessor (NGROUP), there is room for improvement, as demonstrated by Thomas (1993). In particular, CGROUP is found to inflate variance estimates for examinees with large posterior variances. As of now, there is no entirely satisfactory alternative to CGROUP. As this work shows, an ISEM method using importance

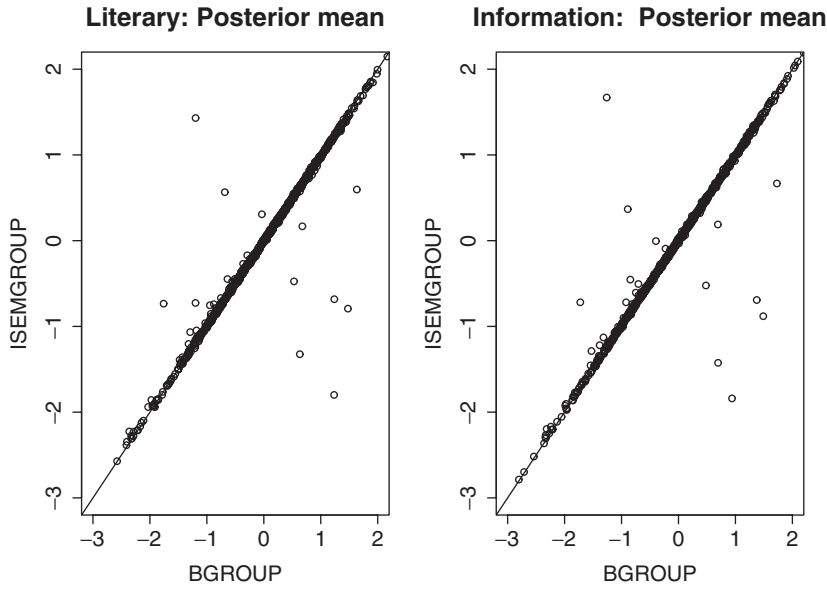

FIGURE 2. *Posterior means produced by ISEMGROUP and BGROUP for the NAEP 2003 Grade 4 reading data.*

sampling provides a viable alternative to CGROUP. Application of the suggested method to a simulated data set and a real data set reveals interesting facts.

The regression parameter estimates are extremely close for BGROUP (the gold standard), CGROUP (the current operational version of MGROUP used in NAEP and other similar assessments), and the newly proposed ISEMGROUP. The estimates of conditional posterior means seem to be quite close for CGROUP and ISEMGROUP. In the real data example, both of these sets of estimates are very close to the BGROUP estimates. The ISEM algorithm produced a few outlying estimated conditional posterior means, especially in the real data example, which may be due to Monte Carlo error; this issue needs further investigation. Statements about agreement between posterior *SD*s cannot be clearly made as of now. For the real data example, which has a large sample size, CGROUP provides inflated estimates of posterior *SD*s for examinees with large *SD*s, whereas ISEMGROUP does not; however, ISEMGROUP results in a few outliers in the middle of the range of the *SD*s. Interestingly, the pattern is quite different for the simulated data, where the ISEMGROUP estimates of posterior *SD*s are bigger than those from CGROUP. Overall, the ISEM method performs slightly better than the CGROUP version of MGROUP.

One problem with the ISEM method is that it is time consuming. For our real data example (2003 reading data), while BGROUP takes 8 hr and CGROUP takes 4 hr on a Pentium IV computer with 2.2 GHz, ISEMGROUP takes

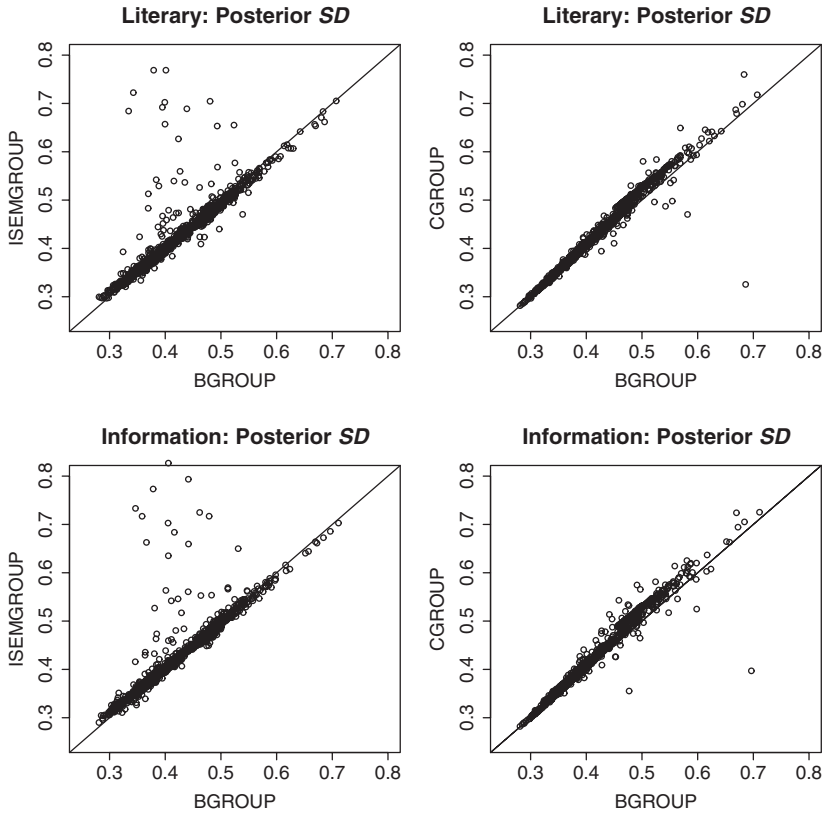

FIGURE 3. *Posterior* SD*s produced by CGROUP, ISEMGROUP, and BGROUP for the NAEP 2003 Grade 4 reading data.*

about 48 hr. However, with the advent of increasingly fast computers, application of ISEMGROUP will become more feasible. At the very least, the proposed method can be used to generate a second set of estimates based on a different approach to a complex estimation problem.

It may be possible to improve the efficiency of the ISEM method even further. Let us consider estimation of $E(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)$, which is given in Equation 9. Let us denote $\boldsymbol{\theta}_0$ to be the mode of $p(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)$. It can be shown that

$$E(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t) = e^{u(\boldsymbol{\theta}_0)} \int \boldsymbol{\theta} \exp\left\{\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)' u''(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\} \exp(\Delta(\boldsymbol{\theta}))d\boldsymbol{\theta},$$

where $u(\boldsymbol{\theta}) = \log[p(\boldsymbol{\theta}|\boldsymbol{X},\ \boldsymbol{Y},\ \boldsymbol{\Gamma}_t,\ \boldsymbol{\Sigma}_t)]$, and $\Delta(\boldsymbol{\theta}) = u(\boldsymbol{\theta}) - u(\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)' u''(\boldsymbol{\theta}_0)$ $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$. The quantity $\Delta(\boldsymbol{\theta})$ is much less variable (and close to zero) across the

TABLE 4
Comparison of Subgroup Estimates From BGROUP, CGROUP, and ISEMGROUP

	BGROUP		CGROUP		ISEMGROUP	
Subgroup	Literary	Information	Literary	Information	Literary	Information
Overall	.025 (.95)	−.003 (.98)	.020 (.96)	−.008 (.99)	.020 (.94)	−.003 (.98)
Male	−.093 (.96)	−.065 (1.00)	−.100 (.96)	−.072 (1.01)	−.099 (.94)	−.066 (.99)
Female	.147 (.94)	.060 (.97)	.143 (.94)	.057 (.98)	.142 (.93)	.061 (.96)
White	.277 (.87)	.278 (.89)	.273 (.87)	.277 (.90)	.272 (.86)	.279 (.89)
Black	−.478 (.92)	−.549 (.92)	−.487 (.92)	−.560 (.93)	−.482 (.91)	−.547 (.91)
Hispanic	−.402 (.94)	−.493 (.94)	−.409 (.94)	−.503 (.95)	−.407 (.93)	−.492 (.93)
Asian	.213 (.93)	.187 (.97)	.210 (.94)	.186 (.98)	.208 (.92)	.188 (.96)
AmInd	−.391 (.94)	−.400 (.94)	−.401 (.95)	−.408 (.95)	−.397 (.92)	−.400 (.93)

Note: Asian = Asian Pacific; AmInd = American Indian.

range of $\boldsymbol{\theta}$ than is $u(\boldsymbol{\theta})$. The first term under exponentiation in the last step above forms a normal distribution. Hence, use of this normal distribution as the importance sampling density should lead to more stable results, and the required importance sample size to reach the same level of accuracy might be greatly reduced.

Another possible area of future research is to combine the suggested method with the implementation of the sampling importance resampling method suggested by Thomas and Gan (1997) to generate the plausible values in the current NAEP estimation method. The method consists of the following steps: (a) Draw a random sample $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_n$ from an appropriately chosen importance sampling density; (b) calculate the importance ratios $r_1, r_2, \ldots r_n$ as described in Application of the Importance Sampling EM Method to NAEP above; and (c) select one among $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_n$ by drawing an index from $1, 2, \ldots n$ with probabilities $r_1, r_2, \ldots r_n$; this draw will be an improved plausible value. Thomas and Gan use a multivariate t_{20} importance density with the same mean as the examinee posterior mean and with a variance equal to the examinee posterior variance inflated by a factor of 1.667; their method produces up to moderate changes in commonly reported estimates in NAEP.

There are a number of issues that need to be answered before the method is implemented in practice. First, a study of the Monte Carlo error in estimating the conditional posterior means and variances is required. Second, we would like to study the effect of different importance sample sizes on the parameter estimates obtained, mainly to find out what is the optimum sample size required to obtain reasonable accuracy in the estimation process. Third, the issue of convergence of the EM algorithm used in ISEMGROUP needs to be explored further. Fourth, the optimal choice of the degrees of freedom of the t importance sampling density is an open question and needs further investigation.

References

Beaton, A. (1987). *The NAEP 1983–84 technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, *17*, 95-109.

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, *61*, 265-285.

Broniatowski, M., Celeux, G., & Diebolt, J. (1983). Reconnaissance de melanges de densites par un algorithme d'apprentissage probabiliste [Identification of mixture distributions by a probabilistic adaptive learning algorithm]. *Data Analysis and Informatics*, *3*, 359-373.

Celeux, G., & Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics*, *2*, 73-82.

Clarkson, D. B., & Gonzalez, R. (2001). Random effects diagonal metric multidimensional scaling models. *Psychometrika*, *66*, 25-43.

Cohen, J. D., & Jiang, T. (1999). Comparison of partially measured latent traits across normal populations. *Journal of the American Statistical Association*, *94*, 1035-1044.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1-38.

Fox, J. P. (2003). Stochastic EM for estimating the parameters of multilevel IRT model. *British Journal of Mathematical and Statistical Psychology*, *56*, 65-81.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317-1339.

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

Johnson, M. S., & Jenkins. F. (2004). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress* (ETS RR-04-38). Princeton, NJ: ETS.

Kass, R., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association*, *84*, 717-726.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (ETS RR-01-25). Princeton, NJ: ETS.

Lee, S. Y., Song, X. Y., & Lee, J. C. K. (2003). Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data. *Journal of Educational and Behavioral Statistics*, *28*, 111-124.

Marshall, A. W. (1956). The use of multi-stage sampling schemes in Monte Carlo computations. In M. A. Meyer (Ed.), *Symposium on Monte Carlo methods* (pp. 123-140). New York: John Wiley.

Martin, M. O., & Kelly, D. L. (1996). *TIMSS technical report: Vol. I. Design and development.* Chestnut Hill, MA: Boston College.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, *89*, 330-335.

Meng, X. L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*, 1254-1267.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika*, *44*, 358-381.

Mislevy, R. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993-997.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, *17*, 131-154.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools.* Chestnut Hill, MA: Boston College.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*, 309-322.

Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, *22*, 425-446.

von Davier, M. (2003). *Comparing conditional and marginal direct estimation of subgroup distributions* (ETS RR-03-02). Princeton, NJ: Educational Testing Service.

von Davier, M., & Sinharay, S. (2004). *Application of the stochastic EM method to latent regression models* (ETS RR-04-34). Princeton, NJ: Educational Testing Service.

Wei, G. C. G., & Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*, 699-704.

Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregate bias. *Journal of the American Statistical Association*, *57*, 348-368.

Authors

MATTHIAS VON DAVIER is Principal Research Scientist, Statistical and Psychometric Theory and Practice, Educational Testing Service, Rosedale Road, Princeton, NJ 08541; mvondavier@ets.org. His principal areas of interest are item response theory, latent class models, mixture distribution models, and applications of these methods to large-scale educational assessment data.

SANDIP SINHARAY is Senior Research Scientist, Statistical and Psychometric Theory and Practice, Educational Testing Service, Rosedale Road, Princeton, NJ 08541; ssinharay@ets.org. His principal areas of interest are Bayesian statistics, model checking and model selection, statistical computing, equating, differential item functioning, and item response theory.

Manuscript received June 29, 2004
Accepted June 16, 2005