Journal of Educational and Behavioral Statistics
Winter 1997, Vol. 22, No. 4, pp. 425–445

Generating Multiple Imputations for Matrix Sampling Data Analyzed With Item Response Models

Neal Thomas
University of North Carolina

Nianci Gan
North Carolina State University

Keywords: *EM algorithm, matrix sampling, measurement error, multiple imputation, National Assessment of Educational Progress (NAEP), SIR algorithm*

Sample survey designs in which each participant is administered a subset of the items contained in a complete survey instrument are becoming an increasingly popular method of reducing respondent burden (Mislevy, Beaton, Kaplan, & Sheehan, 1992; Raghunathan & Grizzle, 1995; Wacholder, Carroll, Pee, & Gail, 1994). Data from these survey designs can be analyzed using multiple imputation methodology that generates several imputed values for the missing data and thus yields several complete data sets. These data sets are then analyzed using complete data estimators and their standard errors (Rubin, 1987b). Generating the imputed data sets, however, can be very difficult. We describe improvements to the methods currently used to generate the imputed data sets for item response models summarizing educational data collected by the National Assessment of Educational Progress (NAEP), an ongoing collection of samples of 4th, 8th, and 12th grade students in the United States. The improved approximations produce small to moderate changes in commonly reported estimates, with the larger changes associated with an increasing amount of missing data. The improved approximations produce larger standard errors.

Matrix sampling designs involve administering several survey instruments, each of which contains only a subset of the items surveyed, thereby creating a potentially large amount of item nonresponse. This item nonresponse, in contrast with typical nonresponse, is known to be missing completely at random. By administering a subset of questions to each individual, the time and effort needed to complete the survey instrument can be reduced. Much of the reduced burden on the sampled individuals is replaced by increased effort on the part of the data analysts.

We thank Eddie Ip, Charles Lewis, Tom Leonard, Bob Mislevy, the associate editor, and two reviewers for many useful suggestions. This work was supported by NCES Grant R999B40014 and the Research Statistics Group at Educational Testing Service.

We describe and assess the missing data methods currently used to analyze data from matrix sampling designs implemented by the National Assessment of Educational Progress (NAEP) and develop several improved methods. The matrix sampling designs are also utilized by the Adult Literacy Survey and the International Mathematics and Science Studies. Testing data and background information (e.g., demographic and educational environment variables) are summarized using item response models and multivariate multiple regression. These models are evaluated using an EM algorithm to obtain maximum likelihood estimates followed by multiple imputation of "complete" data sets. The imputed data sets allow the use of standard estimators of summary statistics by NAEP staff and secondary users. Mislevy (1991) explains the decision to use multiple imputation to analyze NAEP data.

We describe the models and the methods currently used by NAEP to generate imputed data sets in the following section. The section after that contains a description of a method for approximating standard errors for the model parameters that is computationally feasible in high-dimensional problems. We then discuss improved methods for generating the imputed values, and in the final section we compare these methods with current methods using 1990 NAEP mathematics data.

NAEP Implementation of Multiple Imputation

An Item Response Model for Matrix Sampled Educational Data

NAEP collects probability samples of U.S. students and obtains information predictive of academic performance such as demographic and educational environment variables. We denote these variables for each student in a sample of size n by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$. Each $\mathbf{x}_i$ is composed of M scalar variables, $\mathbf{x}_i' = (x_{i1}, \ldots, x_{iM})$, and for modeling convenience, the first component of $\mathbf{x}_i$ is a constant intercept term. A short (typically 45 minutes) examination is also administered to each student, and the results of the examination for the ith student are denoted by $\mathbf{y}_i$.

A model representing the data is specified in two stages. First, a latent proficiency vector, $\boldsymbol{\theta}_i' = (\theta_{i1}, \ldots, \theta_{ip})$, is hypothesized for the ith student, and the $\boldsymbol{\theta}_i$ vectors are assumed to be normally distributed conditional on the variables $\mathbf{x}_i$. The mean of this conditional distribution is given by the multivariate multiple linear regression $\boldsymbol{\Gamma}'\mathbf{x}_i$, where $\boldsymbol{\Gamma} = [\gamma_1 | \ldots . | \gamma_p]$ are $M \cdot p$ unknown regression parameters, and a common (unknown) p-dimensional variance-covariance matrix $\boldsymbol{\Sigma}$ with elements Σ_{jk} is assumed yielding the normal distribution $\phi(\boldsymbol{\theta}_i; \boldsymbol{\Gamma}'\mathbf{x}_i, \boldsymbol{\Sigma})$ (Mislevy, Johnson, & Muraki, 1992). Examples of the intended interpretation of the proficiency variables from the NAEP mathematics assessments are the four proficiency variables representing (a) numbers and operations, (b) measurement, (c) geometry, and (d) algebra and functions.

The distribution of $\boldsymbol{\theta}_i$ can be viewed as a prior distribution conditional on knowing $\mathbf{x}_i$ and the parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ before observing the examination data $\mathbf{y}_i$. Note that the distribution of $\mathbf{x}$ is not specified here, because as is typical in

regression applications, any unknown parameters associated with the distribution of $\mathbf{x}$ are assumed independent of the $\boldsymbol{\theta}_i$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$, so that their estimation is not required for the estimation of the conditional distribution of $\boldsymbol{\theta}$. The proficiency vectors of different students are assumed to be independent despite the clustering in the sampling design. This simplifying assumption is motivated by noting that much of the effect of the clustering is eliminated by conditioning on the extensive background data $\mathbf{x}$.

The second part of the model is an item response model for the examination data $\mathbf{y}_i$ conditional on $\boldsymbol{\theta}_i$ and $\mathbf{x}_i$. The $\mathbf{y}_i$ data are partitioned by the test designers into p content areas, $\mathbf{y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ip})$, corresponding to the p latent proficiency variables. Each $\mathbf{y}_{ij}$ is composed of item scores, y_{ijk}, which are binary or ordinal with values coded as $0, 1, \ldots, m_{jk}$ for the kth item measuring the jth trait. Each student receives only a small subset of the items in the matrix sampling design, and a value of -1 is assigned to indicate that an item has not been administered. The model invokes several independence assumptions conditional on the latent proficiency variables: (a) the item responses are independent of $\mathbf{x}_i$; (b) the responses of a student to different items are independent; (c) responses from different students are independent; and (d) $\mathbf{y}_{ij}$ are independent of θ_{ik} conditional on θ_{ij}, $j \neq k$, that is, item scores depend only on the proficiency to which they are assigned. With these independence assumptions, the density of $\mathbf{y}_i$ conditional on $\boldsymbol{\theta}_i$ can be represented as

$$\prod_{j=1}^{p} f_j\left(\mathbf{y}_{ij} \mid \theta_{ij}\right), \tag{1}$$

and each $f_j(\mathbf{y}_{ij} \mid \theta_{ij})$ is in turn the product of the response probabilities for the examination items. The product in (1) can be viewed as the likelihood function for $\boldsymbol{\theta}_i$ based on the $\mathbf{y}_i$ data. Items not presented to a student do not contribute to the likelihood function because they are missing completely at random.

Binary scored items contributing to $f_j(\mathbf{y}_{ij} \mid \theta_{ij})$ in (1) are represented by three-parameter logistic item response models. For a binary item, y_{ijk}, the model is

$$P(y_{ijk} = 1 \mid \theta_{ij}) = c_{jk} + (1 - c_{jk}) / \{1 + \exp[a_{jk}(\theta_{ij} - b_{jk})]\}. \tag{2}$$

The response probabilities contributing to (1) for an ordinal item, y_{ijk}, are modeled by a partial credit model,

$$P(y_{ijk} = l \mid \theta_{ij}) = \frac{\exp\left\{\sum_{h=1}^{l} a_{jk}(\theta_{ij} - b_{jkh})\right\}}{\sum_{q=1}^{m_{jk}} \exp\left\{\sum_{h=1}^{q} a_{jk}(\theta_{ij} - b_{jkh})\right\}}, \tag{3}$$

$l = 1, \ldots, m_{jk}$. We represent all of the item-specific parameters (a_{jk}, b_{jk}, c_{jk}) associated with items measuring the jth trait by a parameter vector $\boldsymbol{\beta}_j$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$, and we include the $\boldsymbol{\beta}_j$ in the likelihood function for θ_{ij}, $f_j(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j)$, to explicitly denote the dependence of (1) on each $\boldsymbol{\beta}_j$. Mislevy and Bock (1982)

and Muraki (1992) give details of these models. The methods discussed in the following sections can be easily adapted for other item response models, provided they retain the same independence assumptions.

When the independence assumptions are applied, the regression model and the item response model fully specify the distribution of observed data. The likelihood function for the parameters based on the distribution of the data is

$$\text{lik}(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \prod_{i=1}^{N} \int \phi(\boldsymbol{\theta}_i; \boldsymbol{\Gamma}'\mathbf{x}_i, \boldsymbol{\Sigma}) \prod_{j=1}^{p} f_j\left(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j\right) d\boldsymbol{\theta}_i . \tag{4}$$

The integrand in (4) is proportional to the posterior distribution of $\boldsymbol{\theta}_i$ with $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ regarded as known, which we will denote by

$$f(\boldsymbol{\theta}_i \mid \mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto \phi(\boldsymbol{\theta}_i; \boldsymbol{\Gamma}'\mathbf{x}_i, \boldsymbol{\Sigma}) \prod_{j=1}^{p} f_j(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j) . \tag{5}$$

Because we observe $\mathbf{x}_i$ and $\mathbf{y}_i$ and not $\boldsymbol{\theta}_i$ for sampled students, $\boldsymbol{\theta}_i$ is regarded as missing data. The $\boldsymbol{\theta}_i$ are missing completely at random, because none are observed, and they are imputed for each sampled student as described in the next section. The missing item responses are not directly imputed.

Generation of the Imputations

Multiple imputation requires that the missing data, $\boldsymbol{\theta}_i$, be generated from their posterior distribution given the observed data. As is typical of multiple imputation, we form the posterior in two stages, first generating values for the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ from their posterior distribution and then generating $\boldsymbol{\theta}_i$ from their posterior distribution in (5), substituting the generated values for $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$.

NAEP currently employs several approximations in its implementation of this two-stage imputation procedure. We outline the procedures and approximations used by NAEP and then outline methods for assessing and improving the current NAEP methods.

(1) An EM algorithm is used to find maximum likelihood estimates (MLEs) for the item parameters, $\hat{\boldsymbol{\beta}}_j$, separately for each proficiency, $j = 1, \ldots, p$, based on the likelihood function for the subset of examination data associated with the jth proficiency, $\mathbf{y}_{ij}$. Muraki (1992) presents details of this algorithm. In addition, equating methods are often employed, and data from several different samples (typically from different years) are used to rescale and center the $\hat{\boldsymbol{\beta}}_j$.

(2) An EM algorithm is used to estimate $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ using the $\mathbf{x}_i$, $\mathbf{y}_i$ data with the estimated item parameters from Step 1 regarded as fixed and known. We will denote these estimates by $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$ and refer to them as MLEs, although they are not jointly maximized with $\hat{\boldsymbol{\beta}}$ using all of the data.

(3) The posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ is approximated by fixing them at $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Gamma}}$, and $\hat{\boldsymbol{\Sigma}}$; in effect, the parameters are not actually generated from their posterior distribution. Beginning with the 1992 assessments, the $\boldsymbol{\Gamma}$ have been generated from an undocumented multivariate normal approximation developed as a precursor of the newer methods described in this article. For comparisons between our new methods and the existing methods in the final section, we generate the imputations for the existing methods with $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ fixed at their MLEs. Note that no explicit prior distribution is specified for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$.

(4) For each $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ generated in Step 3, a $\boldsymbol{\theta}$ is generated for each student from a multivariate normal distribution that has the same mean and variance-covariance matrix as the actual posterior distribution in (5). The mean and variance-covariance matrix are accurately approximated using methods developed for the EM calculations (Thomas, 1993). The approximation is motivated by the asymptotic normality of the posterior distribution in (5) as the number of examination items is increased (Chang & Stout, 1993).

Steps 3 and 4 are repeated several (typically five) times to produce the data sets with the imputed $\boldsymbol{\theta}$ values that are subsequently analyzed to produce NAEP reports.

Overview

The failure to generate the parameters from their posterior distribution in the first stage of the imputation process produces imputation-based standard errors that are too small (Rubin, 1987b). Because the NAEP sample sizes range from 1,000 to 15,000, the standard asymptotic normal approximation for the posterior distribution of the parameters formed with a diffuse prior distribution should be good, especially when it is applied to an appropriate transformation of the $\boldsymbol{\Sigma}$ parameters. In the next section we compute the asymptotic standard errors directly from the second derivatives of the log likelihood function and show that these standard errors can be adequately approximated using a much simpler approach based on a normal measurement error model. We then show that the simplified standard error structure allows computationally simple generation of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$, even when there are many parameters, as is typical in real applications. We also introduce a transformation of $\boldsymbol{\Sigma}$ that ensures that the generated variance matrices are positive definite. The development of computationally feasible methods to account for the estimation of the item parameters and the consequences of various equating techniques remains unsolved. Mislevy, Sheehan, and Wingersky (in press), Tsutakawa and Johnson (1990), and Adams, Wilson, and Wu (1997) discuss closely related topics. In a later section we use an importance sampling algorithm to improve the normal approximation to the posterior distribution in (5). Finally we evaluate the changes in NAEP reporting statistics due to the improved approximations.

An alternative method for jointly generating $\mathbf{\Gamma}$, $\mathbf{\Sigma}$, and $\boldsymbol{\theta}$ for item response models adapting a fully Bayesian approach utilizes Gibbs sampling algorithms (Albert, 1992; Albert & Chib, 1993; Meng & Schilling, 1996). S. Zeger and Karim (1991) present another example of Gibbs sampling applied to dependent observations represented by generalized linear models. Research on closely related models for correlated binary and ordinal data is reviewed by Pendergast et al. (1996). Because of the large number of items (typically greater than 200) and the numerous parameters associated with the regressor variables, computationally and operationally feasible iterative simulation algorithms for large-scale matrix sampling surveys have not yet been developed. The computationally simpler methods we develop could be used to create good starting values for iterative simulation methods.

Computing Standard Errors for $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$

Computing the second derivatives of the log of the likelihood function in (4) involves the evaluation of integrals that are not analytically tractable. When $\boldsymbol{\theta}$ is one- or two-dimensional, these integrals can be computed by standard quadrature methods with operationally acceptable speed. When the dimension of $\boldsymbol{\theta}$ is three or greater (six is the current maximum), the number of integrals required is very large, and each integral has the same dimension as $\boldsymbol{\theta}$. We created a computer program to evaluate the standard errors of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ for one- and two-dimensional $\boldsymbol{\theta}$, primarily for the purpose of evaluating approximations to these standard errors. For operational implementation, computationally simpler alternatives are required. We develop an approximation by introducing a model that has measurement errors that are normally distributed but still has many properties similar to the more complex NAEP model, whose measurement errors are implied by the item response part of that model.

A Normal Measurement Error Model

The NAEP imputation methods described in the previous section use the asymptotic normality of (5) based on an increasing number of items to approximate analytically intractable densities. Additional simplification can be achieved by using normal approximations to the likelihood function in (1), motivated by similar asymptotic considerations. If we approximate $f_j(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j)$ by $\phi(\mu_{ij}, \tau_{ij})$ for appropriately chosen mean (μ_{ij}) and variance (τ_{ij}), which depend on the $\mathbf{y}_i$ and the (supposed) known item parameters, $\boldsymbol{\beta}_j$, then the integrals in (4) have simple closed forms that yield much simpler likelihood-based estimation and standard errors for $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ (Mislevy, 1990).

Further simplification can be achieved by setting each τ_{ij} equal to τ_j, for some choice of τ_j. We will demonstrate that using a model with a constant measurement error to compute standard errors produces adequate approximations for the NAEP mathematics data, despite the fact that the τ_{ij} vary considerably. We can express the formal similarity of the likelihood function in (4) with that of the likelihood function arising in a standard multiple regression setting in more

familiar terms by assigning $y^*_{ij} = \mu_{ij}$, $j = 1, \ldots, p$, and $\mathbf{y}^*_i = (y^*_{i1}, \ldots, y^*_{ip})'$, $i = 1, \ldots, N$, and then regarding the $\mathbf{y}^*_i$ as our dependent variable in the regression model

$$y^*_i = \Gamma' \mathbf{x}_i + e^*_i, \qquad e^*_i \sim \phi \{\mathbf{0}, \Sigma + \text{diag}(\tau_1, \ldots, \tau_p)\}, \tag{6}$$

where the τ are known. Substituting the likelihood function for the model in (6) for the likelihood function in (4) and evaluating it at the MLE based on (4) yields well-known standard error formulas. The variance-covariance matrix for γ_j is $(\mathbf{X}'\mathbf{X})^{-1}(\hat{\Sigma}_{jj} + \tau_j)$, and the covariance matrix for γ_j and γ_k is $(\mathbf{X}'\mathbf{X})^{-1}\hat{\Sigma}_{jk}$. Define $\boldsymbol{\sigma} = \text{vech}(\Sigma)$, where vech is the vector of unique elements of Σ ordered by columns, and denote the variance-covariance matrix of $\boldsymbol{\sigma}$ by $\mathbf{V}_{\boldsymbol{\sigma}}$; the covariance of Σ_{jk} and Σ_{lm} is $(1/N)\,(\hat{\Sigma}^*_{jl}\hat{\Sigma}^*_{km} + \hat{\Sigma}^*_{jm}\hat{\Sigma}^*_{kl})$, where $\hat{\Sigma}^*_{jj} = \hat{\Sigma}_{jj} + \tau_j$, $j = 1, \ldots, p$, $\hat{\Sigma}^*_{jk} = \hat{\Sigma}_{jk}$, $j \neq k$ (Press, 1972). Note that we express the asymptotic standard errors from the Bayesian perspective in which the parameters are random variables and not the estimators, because this is the way the distributions are used in the imputation procedures.

In the following section we propose a method for selecting τ_j and compare the standard errors based on the approximating model in (6) with results obtained by direct evaluation of the second derivatives of the likelihood function in (4).

Approximating Standard Errors

The method we propose for selecting τ_j is motivated by the fact that the MLE and their standard errors obtained from univariate analyses applied separately to each proficiency variable and its associated examination data are very similar to the MLE and standard errors produced by the multivariate procedure. The multivariate procedure also estimates the correlations between the proficiency variables that are not directly available from the univariate analyses. When the measurement errors are normally distributed and $\tau_{ij} = \tau_j$, the univariate and multivariate MLEs of the parameters coincide. For some matrix sampling designs, such as the 1992 NAEP reading assessments, the univariate and multivariate MLEs differ moderately, and the univariate procedure tends to produce larger standard errors. L. Zeger and Thomas (1997) discuss the close relationship between the univariate and multivariate results.

We select a common measurement error variance for each proficiency variable by computing the standard error for the grand mean of each proficiency variable directly from the exact univariate likelihood function for the variable, which is not a computational burden, and then solve a simple equation for the τ_j to make the standard error based on the approximating constant normal measurement error model agree with the univariate standard error for the grand mean. More specifically, after the multivariate EM algorithm for Γ and Σ is completed, we compute the variance-covariance matrix for the regression coefficients based on the univariate log likelihood function evaluated at each multi-

variate MLE, $\hat{\gamma}_j$. We denote these variance matrices by $\mathbf{V}_j$. Letting $\bar{\mathbf{x}}$ denote the mean of the $\mathbf{x}_i$, we find τ_j so that

$$\bar{\mathbf{x}}' (\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}} (\hat{\Sigma}_{jj} + \tau_j) = \bar{\mathbf{x}}'\mathbf{V}_j\bar{\mathbf{x}}. \tag{7}$$

For examination designs in which the univariate standard errors are larger than the multivariate standard errors, the proposed method of selecting the τ_j will tend to produce conservative approximate standard errors.

We demonstrate the standard error approximation using the 1990 NAEP 4th grade mathematics data. To reduce the number of comparisons, we present results for a model that includes only a small subset of the NAEP background $\mathbf{x}$ variables, which are described in Table 1. Two math proficiency distributions are estimated, Numbers and Operations (θ_1) and Measurement (θ_2). The second column in the upper portion of Table 2 contains the estimates of γ_1 for the Numbers and Operations proficiency obtained from the bivariate EM algorithm, the middle portion of the table contains the corresponding estimates of γ_2 for the Measurement proficiency, and the lower portion contains the estimates of Σ. The standard errors computed using numerical quadrature to compute the derivatives of the bivariate log likelihood function are displayed in the third column, and the approximate standard errors from (6) are displayed in the fourth column. The quality of the approximations is typical of that observed in several large examples. The approximate standard errors tend to be too large for variables indicating higher-achieving populations and too small for lower-achieving populations, because the measurement errors implied by the item response model decrease with increasing proficiency. The larger measurement errors are a consequence of the fact that the NAEP examinations, which contain some multiple-choice items with potential guessing, are too difficult for the lower-achieving students. The standard errors that would result if the estimates were based on "complete" data—for example, $(\mathbf{X}'\mathbf{X})^{-1}\Sigma_{jj}$—are displayed in the fifth column of Table 2.

TABLE 1
Description of regressor variables

Parameter	Description
Female	1 if examinee is female
Afric	1 if examinee is African American
Hisp	1 if examinee is Hispanic
Asian	1 if examinee is Asian
EMALMG	Equal to modal age, less than modal grade
EMAGMG	Equal to modal age, greater than modal grade
GMAEMG	Greater than modal age, equal to modal grade
LMAEMG	Less than modal age, equal to modal grade

Note. Modal age/grade is 9 years/4th grade.

TABLE 2
Standard error comparisons

Regressor	Estimate	Quadrature *SE*	Normal error *SE*	"Complete" data *SE*
		Numbers and Operations		
Intercept	−0.538	0.015	0.015	0.013
Female	−0.024	0.016	0.016	0.013
Afric	−0.536	0.022	0.022	0.018
Hisp	−0.428	0.021	0.020	0.017
Asian	0.086	0.044	0.046	0.038
EMALMG	−0.607	0.020	0.019	0.016
EMAGMG	0.546	0.149	0.157	0.131
GMAEMG	−0.203	0.019	0.019	0.016
LMAEMG	0.116	0.144	0.151	0.126
		Measurement		
Intercept	−0.330	0.017	0.018	0.013
Female	−0.112	0.018	0.018	0.013
Afric	−0.667	0.027	0.025	0.018
Hisp	−0.463	0.024	0.024	0.017
Asian	0.051	0.049	0.053	0.038
EMALMG	−0.572	0.023	0.023	0.016
EMAGMG	0.461	0.172	0.183	0.132
GMAEMG	−0.165	0.021	0.022	0.016
LMAEMG	−0.377	0.181	0.175	0.126
		Variance-covariance matrix		
Σ_{11}	0.378	0.0087	0.0082	0.0057
Σ_{22}	0.379	0.0083	0.0077	0.0056
Σ_{12}	0.358	0.0124	0.0110	0.0057

The numerous covariance terms between the different estimates, which are not displayed, were also adequately approximated. In particular, the covariances between the estimates of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ based on quadrature are nearly zero, in good agreement with the zero covariance predicted by the approximation in (6).

Generating the Model Parameters

We generate $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ from an asymptotic normal approximation with mean at the MLE, $(\hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\Sigma}})$, and an approximate variance-covariance matrix obtained from the simplified regression model in (6). There are two potential difficulties in the generation of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$. First, the proficiency variables are often highly correlated, so that the $\boldsymbol{\Sigma}$ generated from a normal approximation may not be positive definite. Second, the dimension of $\boldsymbol{\Gamma}$ can be excessively high, for example,

$200 \cdot 6 = 1{,}200$ in some of the NAEP mathematics models. These problems are addressed by first transforming the parameters, then generating the transformed parameters from their more computationally tractable distributions, and then back-transforming the generated parameters.

We use the matrix log to transform $\boldsymbol{\Sigma}$, as proposed by Leonard and Hsu (1992) and Chiu, Leonard, and Tsui (1996). The matrix log of $\hat{\boldsymbol{\Sigma}}$ is computed by forming the spectral decomposition $\hat{\boldsymbol{\Sigma}} = \mathbf{EDE}'$, where $\mathbf{E}$ is the matrix of normalized eigenvectors of $\hat{\boldsymbol{\Sigma}}$, and $\mathbf{D}$ is the diagonal matrix of corresponding eigenvalues of $\hat{\boldsymbol{\Sigma}}$. The matrix log is defined as

$$\hat{\mathbf{A}} = \log(\hat{\boldsymbol{\Sigma}}) = \mathbf{E} \log(\mathbf{D})\, \mathbf{E}',$$

and the matrix exponential is computed by exponentiating the eigenvalues. Set $\hat{\boldsymbol{\alpha}} = \text{vech}(\hat{\mathbf{A}})$, and denote the corresponding values of $\log(\boldsymbol{\Sigma})$ by $\mathbf{A}$ and $\boldsymbol{\alpha}$; the resulting $\boldsymbol{\alpha}$ vector has elements that are unconstrained in the set of real numbers, and their distribution should be better approximated by a multivariate normal distribution than vech $(\boldsymbol{\Sigma})$, based on the experience of Leonard and Hsu. Note that when $\boldsymbol{\Sigma}$ is diagonal, the matrix log reduces to the usual normalizing log transformation of a variance. We use the matrix log transformation instead of the simpler univariate log transformation of the variances and the Fisher Z transformation of the correlations, because the latter transformations are not assured of producing positive definite matrices when the dimension is greater than two and the correlations are close to one, as is the case with many NAEP applications.

The mean of the approximate posterior distribution for $\boldsymbol{\alpha}$ is the transformed MLE, $\hat{\boldsymbol{\alpha}}$, and the asymptotic variance of $\boldsymbol{\alpha}$ is approximated by $\mathbf{JV}_{\sigma}\mathbf{J}'$, where $\mathbf{J}$ is the Jacobian of the vech operator applied to the matrix log transform evaluated at $\hat{\boldsymbol{\Sigma}}$. The inverse of $\mathbf{J}$—that is, the Jacobian of the vech operator applied to the matrix exponential—is derived in the Appendix and given in Equation A1. After each $\boldsymbol{\alpha}$ is drawn from its approximate distribution, it is expanded to form a symmetric matrix and exponentiated to produce a $\boldsymbol{\Sigma}$ drawn from its approximate distribution.

The changes in the $\boldsymbol{\Sigma}$ generated using the transformation are typically modest, although using the transformation of $\boldsymbol{\Sigma}$ avoids the substantial operational problems that arise when direct generation of $\boldsymbol{\Sigma}$ produces a matrix that is not positive definite, which can happen when there are several proficiency variables. The plot in Figure 1 displays the effect of the transformation of $\boldsymbol{\Sigma}$ on the correlation coefficients. The plot displays the densities, smoothed by the Splus function DENSITY, based on 2,000 draws of $\boldsymbol{\Sigma}$ and 2,000 back-transformed values of $\log(\boldsymbol{\Sigma})$. The matrix log transformation produces a slight skewing away from the boundary, which is similar to the effect produced by a normal approximation to the Fisher Z transformation (which is not displayed). Similar agreement was found for the diagonal elements of $\boldsymbol{\Sigma}$. The close agreement is a source of confirmation of the adequacy of the normal approximation resulting from the large NAEP sample sizes and the assumption of a common residual variance-covariance matrix.

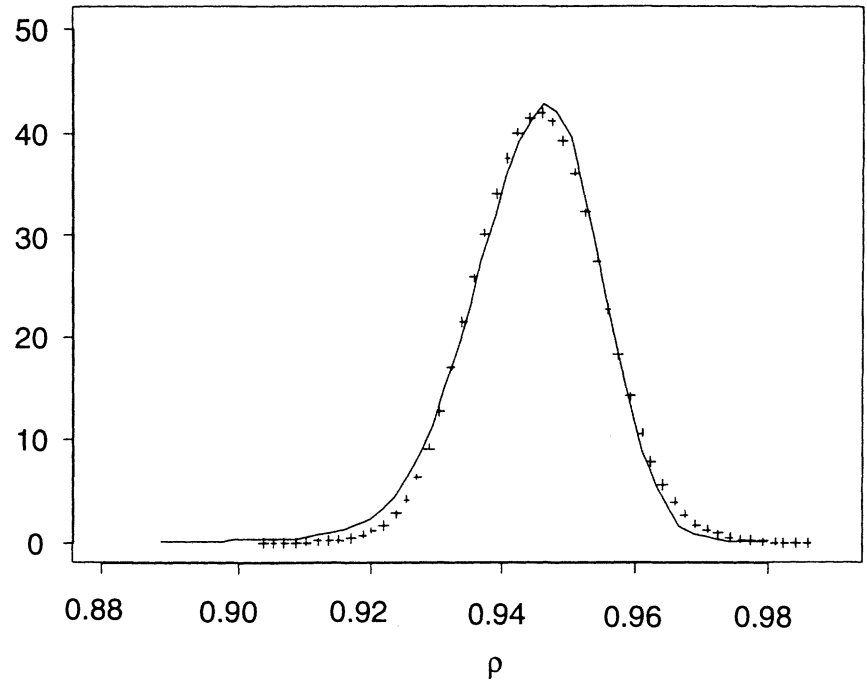

FIGURE 1. *Densities of the correlation between the Numbers and Operations proficiency and the Measurement proficiency*

Note. The density obtained from the 2,000 back-transformed values of $\log(\boldsymbol{\Sigma})$ is represented by the smooth curve, and the crosses represent the density obtained from direct generation of the $\boldsymbol{\Sigma}$.

For each $\boldsymbol{\Sigma}$ that is generated, a corresponding value of $\boldsymbol{\Gamma}$ is drawn from its approximate distribution implied by (6) conditional on the generated value of $\boldsymbol{\Sigma}$. We could use the single value $\hat{\boldsymbol{\Sigma}}$ when generating $\boldsymbol{\Gamma}$, but doing so fails to create the finite-sample t distribution for $\boldsymbol{\Gamma}$. Because we have numerous degrees of freedom, we expect the t correction to be very small, but the computational effort required to use the generated $\boldsymbol{\Sigma}$ is also small. We form the upper triangular square root for each generated variance-covariance matrix, $(\boldsymbol{\Sigma}^{1/2})'\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$, and work with linearly transformed $\boldsymbol{\Gamma}$ parameters, $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_p] = \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1/2}$ and $\hat{\boldsymbol{\delta}} = [\hat{\delta}_1, \ldots, \hat{\delta}_p] = \hat{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}^{-1/2}$. Using standard multivariate calculations for the regression model in (6), $\delta_j \sim \phi(\hat{\delta}_j, (\mathbf{X}'\mathbf{X})^{-1})$, and δ_j is independent of δ_k when $j \neq k$ (Rao, 1965). As a consequence, the $M \cdot p$ parameters can be generated from independent normal random variables using only the inversion of a single $M \cdot M$ matrix. As before, the $\boldsymbol{\Gamma}$ are then computed by back-transforming the generated $\boldsymbol{\delta}$.

Generating $\boldsymbol{\theta}_i$ From Their Conditional Posterior Distributions

The final step in creating the imputations is the generation of the $\boldsymbol{\theta}_i$ from their posterior distribution in (5). The normal approximation to (5) is a useful initial

distribution for developing more accurate methods. The typical deviations of (5) from normality are left skewness for low-scoring students and right skewness for high-scoring students. The density for the Measurement proficiency of a low-scoring student is displayed in Figure 2 along with its approximating normal density. The marginal density was obtained by numerically integrating (5) with respect to the Numbers and Operations proficiency variable. The densities are normed so that they add to one on the grid of points indicated by the density lines representing the normal approximation. The normal approximation is usually better than in Figure 2 for students with less extreme performance.

We use sampling importance resampling (SIR) to improve the normal approximation (Gelfand & Smith, 1990; Rubin 1987a, 1991; Tanner, 1993). This method is appealing because it requires very little programming or computing once the existing EM calculations are performed, and it is noniterative, so that the amount of computation required can be accurately determined in advance and set within feasible limits. The SIR method consists of the following steps.

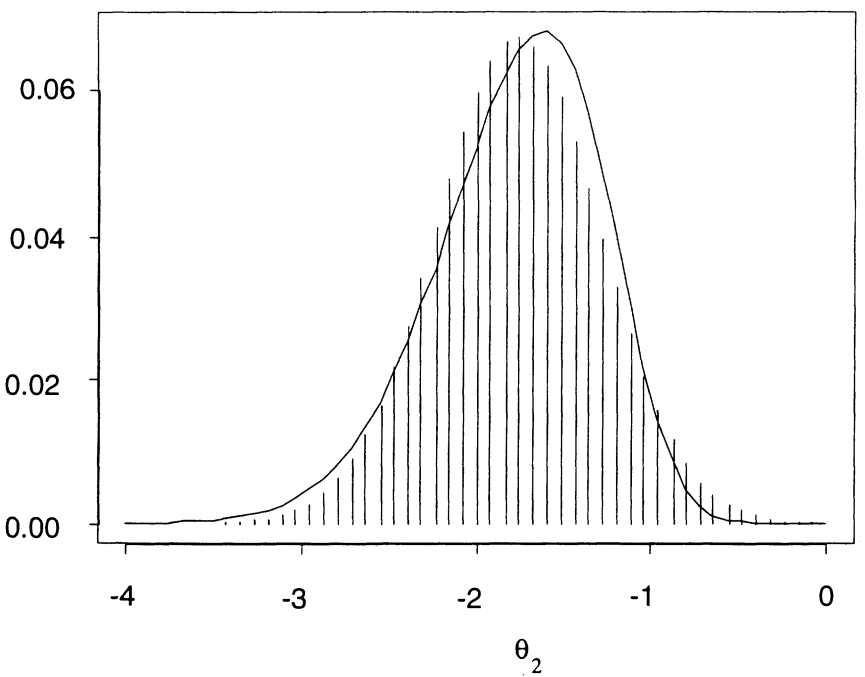

FIGURE 2. *Posterior density (curve) of the measurement proficiency of a low-scoring student, computed by numerical integration, shown with its normal approximation (discretized histogram)*

Note. The discretized histogram is computed from the normal approximation on a grid of 51 points.

(1) Select an importance density, g, that is similar to f in (5), and draw $\boldsymbol{\theta}_*^1, \ldots, \boldsymbol{\theta}_*^s$ random values from g.

(2) Compute the ratio of f and g evaluated at $\boldsymbol{\theta}_*^1, \ldots, \boldsymbol{\theta}_*^s$, and norm the s ratios to add to one: $r_1, \ldots, r_s \propto f(\boldsymbol{\theta}_*^1) / g(\boldsymbol{\theta}_*^1), \ldots, f(\boldsymbol{\theta}_*^s) / g(\boldsymbol{\theta}_*^s)$, $\Sigma_{k=1}^s r_k = 1$.

(3) Select one of the $\boldsymbol{\theta}_*$ by drawing an index from $(1, \ldots, s)$ with probabilities $r_1, \ldots, r_s$.

The SIR improvement to the normal approximation in Figure 2 is displayed in Figure 3. To obtain a very accurate estimate of the SIR generated density, 75,000 independent draws were selected from the SIR algorithm, with $s = 10$. We used a multivariate t_{20} importance density with the same mean as in the normal approximation and with the same variance-covariance matrix inflated by $1.5 \cdot (20/18) = 1.667$. The use of a more dispersed distribution for the importance function is widely recommended in the importance sampling literature. We obtained a good approximation with a small value of s because our importance function is a good approximation to the actual posterior distribution. The importance weights for the example in Figure 2, as well as for all of the students in the

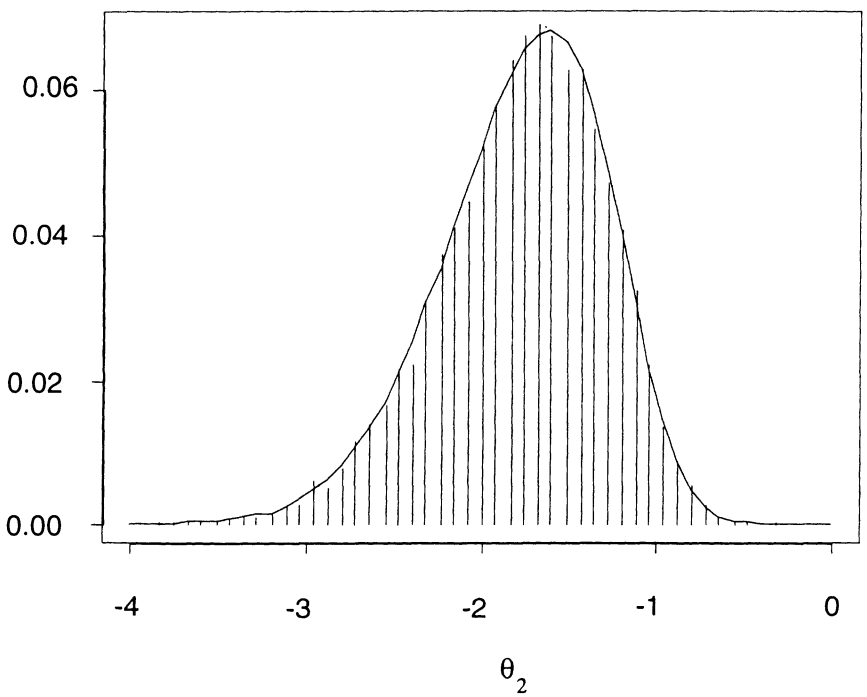

FIGURE 3. *SIR improvement to the normal approximation shown in Figure 2*

Note. The curve is the same posterior density displayed in Figure 2, but the discretized histogram is estimated from 70,000 independent draws from the SIR method.

mathematics sample, are not highly variable. The importance weights should be examined as part of the diagnostic output of the estimation procedure to ensure that a small number of weights do not dominate the importance sampling, which could happen if the initial normal approximations are poor.

Although the SIR method produces noticeably better approximations to the distribution of each $\boldsymbol{\theta}_i$, the effect of the improved approximations on the estimates of aggregate summary statistics is unclear due to the varying quality of the normal approximations and the potential for cancellation of errors. We evaluate the improved approximations using NAEP mathematics data in the following section.

An Example of Estimation From Imputed Data Sets

We compare the NAEP methods described earlier with the improved approximations using data from the 8,790 students in the 1990 4th grade national mathematics sample. Four proficiency variables were measured: (a) Numbers and Operations, (b) Measurement, (c) Geometry, and (d) Algebra and Functions. These variables are listed in decreasing order of the number of items obtained for each, and therefore in order of decreasing precision of measurement. In addition, the mathematics reporting includes a composite variable that is a linear combination of the proficiency variables, with the best measured proficiencies more heavily weighted. Unlike in our earlier example, we include all of the background variables in the analyses to make them comparable to real NAEP analyses.

We describe results for estimates of the mean, variance, and 5th, 25th, 75th, and 95th percentiles of the proficiency distributions, which are a subset of the commonly reported summary statistics. We denote the estimator of any one of the summary statistics by the generic symbol $\hat{e}$. It is computed by averaging the corresponding complete data summary statistics, $\hat{e}_i$, from each of the $i = 1, \ldots, n_{\text{imp}}$ imputed data sets,

$$\hat{e} = \frac{1}{n_{\text{imp}}} \sum_{i=1}^{n_{\text{imp}}} \hat{e}_i \, .$$

The standard error of $\hat{e}$ is computed by the usual multiple imputation formula (Rubin, 1987b):

$$V_{\text{t}}(\hat{e}) = \bar{V}_{\text{c}}(\hat{e}) + (1 + n_{\text{imp}}^{-1}) V_{\text{b}}(\hat{e}) \, ,$$

where $V_{\text{b}}(\hat{e})$ is an estimate of the between-imputation variance of the $\hat{e}_i$,

$$V_{\text{b}}(\hat{e}) = \frac{1}{n_{\text{imp}} - 1} \sum_{i=1}^{n_{\text{imp}}} (\hat{e}_i - \hat{e})^2 \, .$$

The other component of the standard error, $\bar{V}_{\text{c}}(\hat{e})$, is the average of the complete-data squared standard errors. The complete-data standard errors for the complex probability samples collected by NAEP are computed using a jackknife proce-

dure, or, alternatively, by the use of design effects obtained from the study of numerous jackknife standard errors (Johnson & Rust, 1992). The design effects vary considerably for estimates from different subpopulations and vary somewhat between different estimands. We use a small and a large design effect to compute $\bar{V}_c(\hat{e})$ in our study to indicate the magnitude of likely changes (as a function of standard errors) for different subpopulations.

To improve the accuracy of our comparisons, we generated 250 imputed data sets instead of the 5 typically used in practice. The large number of imputations reduces the simulation variability in $\hat{e}$ and also provides a more stable estimate of $V_b(\hat{e})$. When we compute $V_t(\hat{e})$, however, we use $n_{imp} = 5$ to retain comparability with typical NAEP analyses.

The differences between the estimates based on the full sample of 8,790 students computed by the older methods and the improved methods are displayed in Table 3. The differences are divided by their standard errors, $V_t^{1/2}(\hat{e})$, which are computed using the old methods with a design effect of three. The changes range from small, for the composite variable which has the smallest measurement error, to moderate, for the Algebra proficiency, which has the largest measurement errors. The improved approximations produce longer-tailed distributions for the proficiency variables, as seen in the increased variance estimates, as well as in the lower 5th percentile and higher 95th percentile estimates. This behavior is consistent with the skewness toward more extreme positive (negative) values in the posterior distributions of high- (low-) scoring students, which is more accurately approximated with the improved methods, as displayed in Figure 3. Subpopulations with higher proportions of exceptional (high or low) students tend to exhibit bigger differences, but they also have larger standard errors due to the smaller sample sizes for subpopulation estimates. Most of the difference in the estimators is due to the SIR method for generating the $\boldsymbol{\theta}$. Generating $\boldsymbol{\Sigma}$ produces a small portion of the difference, and generation of $\boldsymbol{\Gamma}$ does not produce noticeable changes.

Table 4 contains the ratio of the between-imputation variance, $V_b(\hat{e})$, computed from the improved approximation to the between-imputation variance

TABLE 3
Difference in estimators

Estimand	Composite	Num and Op	Measurement	Geometry	Algebra
Mean	0.10	0.11	0.11	0.03	0.10
Variance	−0.08	−0.13	−0.37	−0.50	−0.80
5th percentile	0.06	0.14	0.28	0.30	0.52
25th percentile	0.00	0.01	0.11	0.09	0.30
75th percentile	0.14	0.14	0.02	−0.08	−0.20
95th percentile	0.00	−0.02	−0.21	−0.40	−0.47

Note. Each entry is the difference (old – new) between the two approximating methods divided by the standard error based on the older approximation.
Num and Op = Numbers and Operations.

TABLE 4
Ratios of error variances

Estimand	Between imputation	Total ($DE = 2$)	Total ($DE = 5$)
		Composite	
Mean	1.12	1.01	1.01
Variance	1.68	1.11	1.05
5th percentile	1.20	1.09	1.05
25th percentile	0.97	0.99	1.00
75th percentile	1.10	1.02	1.01
95th percentile	0.96	0.99	1.00
		Numbers and Operations	
Mean	2.01	1.08	1.04
Variance	3.35	1.40	1.18
5th percentile	2.09	1.42	1.22
25th percentile	1.47	1.13	1.07
75th percentile	1.59	1.11	1.05
95th percentile	1.26	1.08	1.04
		Measurement	
Mean	5.23	1.35	1.15
Variance	6.99	2.10	1.50
5th percentile	2.59	1.63	1.34
25th percentile	2.44	1.43	1.22
75th percentile	2.80	1.37	1.18
95th percentile	2.92	1.56	1.27
		Geometry	
Mean	6.12	1.52	1.23
Variance	9.80	2.77	1.82
5th percentile	3.32	1.99	1.54
25th percentile	3.13	1.65	1.32
75th percentile	3.83	1.71	1.34
95th percentile	4.41	2.06	1.53
		Algebra and Functions	
Mean	6.66	1.53	1.24
Variance	11.33	2.81	1.83
5th percentile	2.92	1.84	1.46
25th percentile	2.49	1.48	1.25
75th percentile	4.60	1.76	1.36
95th percentile	4.42	2.07	1.54

computed by the older method. It also contains the corresponding ratios for the total estimation variance, $V_t(\hat{e})$, using a small design effect of two and a large design effect of five, which reduces the relative contribution of the between-imputation variance to the total variance. The increase in the between-imputation variance depends on the accuracy with which the proficiency variables are measured. The increases are relatively small for the composite variable, which is determined by nearly 1 hour of testing data, and substantial for the individual proficiency variables based on only a few minutes of direct testing. The ratios for the composite variable that are slightly less than one are the result of the small simulation variability that remains even when 250 imputed data sets are formed. Even with a large design effect, the increases in the total variance estimates resulting from the increased between-imputation variances produce increases large enough to be important for all but the most accurately measured NAEP proficiency variables. Most of the increase in the variance of the statistics reported in Table 4 is due to the generation of the $\boldsymbol{\Gamma}$. Generation of $\boldsymbol{\Sigma}$, however, produces large increases in the variances of estimates of correlations between the $\boldsymbol{\theta}$, which are not displayed in Table 4. The generation of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ from their approximate distributions should be included in future NAEP analyses to produce more realistic assessments of uncertainty.

APPENDIX

The Jacobian of the matrix exponential

Using Fuller's (1987, pp. 382–388) notation for representing a symmetric matrix $\mathbf{S}$ with elements s_{ij} in vector form, vec($\mathbf{S}$) is the vector of length p^2 obtained by concatenating the columns of $\mathbf{S}$ one beneath the other beginning with the leftmost column, and vech($\mathbf{S}$) is the vector of the $n_p = (1/2)p(p+1)$ sequentially selected distinct elements of vec($\mathbf{S}$). There exists a $p^2 \times n_p$ matrix $\boldsymbol{\Phi}_p$, such that

$$\text{vec}(\mathbf{S}) = \boldsymbol{\Phi}_p \, \text{vech}(\mathbf{S}), \tag{A1}$$

that can be constructed by noting that the matrix element s_{ij} should be the $[(j-1)p + i]$th element in vec($\mathbf{S}$) and that when $i \geq j$ the position of s_{ij} in vech($\mathbf{S}$) is

$$\begin{aligned} &\underbrace{p + (p-1) + \cdots}_{j-1} + i - j + 1 \\ &= [p + (p-1) + \cdots + (p-(j-1)+1)] + i - j + 1 \\ &= (j-1)(2p-j)/2 + i. \end{aligned}$$

Thus, when $i \geq j$,

$$\boldsymbol{\Phi}_p((j-1)p + i, (j-1)(2p-j)/2 + i) = 1.$$

Similarly, if $i < j$,

$$\boldsymbol{\Phi}_p((j-1)p + i, (i-1)(2p-i)/2 + j) = 1,$$

and all the other elements of $\boldsymbol{\Phi}_p$ are zero. By a similar construction, there exists an $n_p \times p^2$ matrix $\boldsymbol{\Psi}_p$ such that

$$\text{vech}(\mathbf{S}) = \mathbf{\Psi}_p \text{ vec}(\mathbf{S}), \tag{A2}$$

with

$$\mathbf{\Psi}_p((j-1)(2p-j)/2+i, (j-1)p+i) = \begin{cases} 1/2 & \text{if } i > j \\ 1 & \text{if } i = j \end{cases}$$

when $i \geq j$ and

$$\mathbf{\Psi}_p((i-1)(2p-i)/2+j, (j-1)p+i) = 1/2$$

when $i < j$, and all the other elements of $\mathbf{\Psi}_p$ are zero.

Consider the matrix logarithm transformation (Chiu et al., 1996; Leonard & Hsu, 1992) of the variance matrix $\mathbf{\Sigma}$, $\mathbf{A} = \log(\mathbf{\Sigma})$, and its inverse transformation, $\mathbf{\Sigma} = \exp(\mathbf{A})$. We will compute the Jacobian of the inverse transformation evaluated at the MLE, $\hat{\mathbf{\Sigma}}$, for use in computing standard errors for the MLE of the transformed parameters, $\hat{\mathbf{A}} = \log(\hat{\mathbf{\Sigma}})$. The spectral decomposition of $\hat{\mathbf{\Sigma}}$ is

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{E}}\hat{\mathbf{D}}\hat{\mathbf{E}}', \tag{A3}$$

where $\hat{\mathbf{D}}$ is the diagonal matrix of the eigenvalues of $\hat{\mathbf{\Sigma}}$, and the columns of the orthonormal matrix $\hat{\mathbf{E}}$ are the corresponding normalized eigenvectors. Define a function f as

$$f(x, y) = \begin{cases} x & \text{if } x = y, \\ (x-y)/(\log x - \log y) & \text{otherwise,} \end{cases} \tag{A4}$$

and let

$$\mathbf{F}_0 = (f(\hat{d}_{ii}, \hat{d}_{jj}))_{p\times p}, \tag{A5}$$

where $\hat{d}_{ii}$ is the ith element on the diagonal of $\hat{\mathbf{D}}$, and set

$$\hat{\mathbf{F}} = \text{diag}((\text{vech}(\mathbf{F}_0))'), \tag{A6}$$

the $n_p \times n_p$ diagonal matrix with elements of $(\text{vech}(\mathbf{F}_0))'$.

Theorem A.1. The Jacobian of the transformation vech $(\mathbf{\Sigma})$ = vech(exp($\mathbf{A}$)) evaluated at vech $(\hat{\mathbf{A}})$ is

$$\mathbf{\Psi}_p(\hat{\mathbf{E}} \otimes \hat{\mathbf{E}})\mathbf{\Phi}_p\hat{\mathbf{F}}\mathbf{\Psi}_p(\hat{\mathbf{E}}' \otimes \hat{\mathbf{E}}')\mathbf{\Phi}_p, \tag{A7}$$

where $\otimes$ represents the Kronecker product.

Proof of Theorem A.1. The proof utilizes methods suggested by Leonard and Hsu (1992) and Chiu et al. (1996). Applying the Volterra integral equation of Bellman (1960, pp. 170–171) to the matrices $\mathbf{A}$ and $(\mathbf{A} - \hat{\mathbf{A}})$, we have $X(t) = \exp(\mathbf{A}t)$ as the solution to the equation

$$X(t) = \exp(\hat{\mathbf{A}}t) + \int_0^t \exp(\hat{\mathbf{A}}(t-s))(\mathbf{A} - \hat{\mathbf{A}})X(s)\,ds, X(0) = \mathbf{I},$$

which becomes

$$X(t) = \exp(\hat{\mathbf{A}}t) + \int_0^t (\hat{\mathbf{\Sigma}})^{t-s}(\mathbf{A} - \hat{\mathbf{A}})X(s)\,ds, X(0) = \mathbf{I}, \tag{A8}$$

after substituting $\hat{\mathbf{\Sigma}} = \exp(\hat{\mathbf{A}})$. When $t = 1$, (A8) specializes to

$$\exp(\mathbf{A}) = \exp(\hat{\mathbf{A}}) + \int_0^1 (\hat{\boldsymbol{\Sigma}})^{1-s}(\mathbf{A} - \hat{\mathbf{A}})\, X(s)\, ds. \tag{A9}$$

Substituting the representation of $X(s)$ in (A8) into (A9) and simplifying, we have

$$\exp(\mathbf{A}) = \exp(\hat{\mathbf{A}}) + \int_0^1 (\hat{\boldsymbol{\Sigma}})^{1-s}(\mathbf{A} - \hat{\mathbf{A}})(\hat{\boldsymbol{\Sigma}})^{s} ds + \mathbf{P}(\mathbf{A} - \hat{\mathbf{A}}), \tag{A10}$$

where each element of $\mathbf{P}$ is a polynomial of quadratic and higher terms in $\mathbf{A} - \hat{\mathbf{A}}$. Let

$$\mathbf{B} = [b_{ij}]_{p\times p} = \hat{\mathbf{E}}'(\mathbf{A} - \hat{\mathbf{A}})\hat{\mathbf{E}} \tag{A11}$$

and substitute (A3) into (A10) to obtain

$$\exp(\mathbf{A}) - \exp(\hat{\mathbf{A}}) = \hat{\mathbf{E}}\left[\int_0^1 (\hat{\mathbf{D}})^{1-s}\mathbf{B}(\hat{\mathbf{D}})^{s} ds\right]\hat{\mathbf{E}}' + \mathbf{P}(\mathbf{A} - \hat{\mathbf{A}}).$$

Integrating with respect to s yields

$$\exp(\mathbf{A}) - \exp(\hat{\mathbf{A}}) = \hat{\mathbf{E}}[b_{ij} f(\hat{d}_{ii}, \hat{d}_{jj})]_{p\times p}\hat{\mathbf{E}}' + \mathbf{P}(\mathbf{A} - \hat{\mathbf{A}}). \tag{A12}$$

Applying the result involving the Kronecker product and vec operations in Fuller (1987, p. 387) gives

$$\operatorname{vec}(\mathbf{B}) = (\hat{\mathbf{E}}' \otimes \hat{\mathbf{E}}')\operatorname{vec}(\mathbf{A} - \hat{\mathbf{A}}). \tag{A13}$$

Using the transformations between the vec and vech representations in (A1) and (A2), (A13) becomes

$$\operatorname{vech}(\mathbf{B}) = \boldsymbol{\Psi}_p(\hat{\mathbf{E}}' \otimes \hat{\mathbf{E}}')\boldsymbol{\Phi}_p\operatorname{vech}(\mathbf{A} - \hat{\mathbf{A}}). \tag{A14}$$

Similarly,

$$\begin{aligned}
&\operatorname{vech}(\hat{\mathbf{E}}[b_{ij} f(\hat{d}_{ii}, \hat{d}_{jj})]_{p\times p}\hat{\mathbf{E}}') \\
&\quad = \boldsymbol{\Psi}_p(\hat{\mathbf{E}} \otimes \hat{\mathbf{E}})\boldsymbol{\Phi}_p\operatorname{vech}([b_{ij} f(\hat{d}_{ii}, \hat{d}_{jj})]_{p\times p}) \\
&\quad = \boldsymbol{\Psi}_p(\hat{\mathbf{E}} \otimes \hat{\mathbf{E}})\boldsymbol{\Phi}_p\hat{\mathbf{F}}\operatorname{vech}(\mathbf{B}).
\end{aligned} \tag{A15}$$

Now, taking the vech operation on both sides of (A12) and applying (A14) and (A15),

$$\begin{aligned}
&\operatorname{vech}(\exp(\mathbf{A})) - \operatorname{vech}(\exp(\hat{\mathbf{A}})) \\
&\quad = \operatorname{vech}(\hat{\mathbf{E}}[b_{ij} f(\hat{d}_{ii}, \hat{d}_{jj})]_{p\times p}\hat{\mathbf{E}}') + \operatorname{vech}(\mathbf{P}(\mathbf{A} - \hat{\mathbf{A}})) \\
&\quad = \boldsymbol{\Psi}_p(\hat{\mathbf{E}} \otimes \hat{\mathbf{E}})\boldsymbol{\Phi}_p\hat{\mathbf{F}}\operatorname{vech}(\mathbf{B}) + \operatorname{vech}(\mathbf{P}(\mathbf{A} - \hat{\mathbf{A}})) \\
&\quad = \boldsymbol{\Psi}_p(\hat{\mathbf{E}} \otimes \hat{\mathbf{E}})\boldsymbol{\Phi}_p\hat{\mathbf{F}}\boldsymbol{\Psi}_p(\hat{\mathbf{E}}' \otimes \hat{\mathbf{E}}')\boldsymbol{\Phi}_p\operatorname{vech}(\mathbf{A} - \hat{\mathbf{A}}) + \operatorname{vech}(\mathbf{P}(\mathbf{A} - \hat{\mathbf{A}})),
\end{aligned}$$

yielding the first term (Jacobian) in the Taylor expansion of $\exp(\mathbf{A})$ evaluated at $\hat{\mathbf{A}}$.

References

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

Albert, J. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17*, 251–269.

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association, 88*, 669–679.

Bellman, R. (1960). *Introduction to matrix analysis*. New York: McGraw-Hill.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58*, 37–52.

Chiu, Y., Leonard, T., & Tsui, K. (1996). The matrix logarithm covariance model. *Journal of the American Statistical Association, 91*, 198–210.

Fuller, W. (1987). *Measurement error models.* New York: Wiley.

Gelfand, A., & Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

Johnson, E., & Rust, K. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*, 175–190.

Leonard, T., & Hsu, J. (1992). Bayesian inference for a covariance matrix. *The Annals of Statistics, 20*, 1669–1696.

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. (1990). Scaling procedures. In E. Johnson & R. Zwick (Eds.), *Focusing the new design: The NAEP 1988 technical report* (Tech. Rep. No. 19-TR-20). Princeton, NJ: Educational Testing Service.

Mislevy, R. (1991). Randomization-based inferences about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

Mislevy, R., & Bock, D. (1982). *BILOG: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

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

Mislevy R., Sheehan, K., & Wingersky, M. (in press). Dealing with uncertainty about item parameters: Expected response functions. *Journal of Educational and Behavioral Statistics.*

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

Pendergast, J., Gange, S., Newton, M., Lindstrom, M., Palta, M., & Fisher, M. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review, 64*, 89–118.

Press, J. (1972). *Applied multivariate analysis.* New York: Holt, Rinehart, and Winston.

Raghunathan, T., & Grizzle, J. (1995). A split questionnaire survey design. *Journal of the American Statistical Association, 90*, 54–63.

Rao, C. R. (1965). *Linear statistical inference and its applications.* New York: Wiley.

Rubin, D. (1987a). [Comment on "The calculation of posterior distributions by data augmentation," by M. Tanner & W. Wang]. *Journal of the American Statistical Association, 82*, 543–546.

Rubin D. (1987b). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. (1991). EM and beyond. *Psychometrika, 56*, 241–254.

Tanner, M. (1993). *Tools for statistical inference* (2nd ed.). New York: Springer-Verlag.

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

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

Wacholder, S., Carroll, R., Pee, D., & Gail, M. (1994). The partial questionaire design for case-control studies. *Statistics in Medicine, 13*, 623–634.

Zeger, L., & Thomas, N. (1997). Efficient matrix sampling instruments for correlated latent traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association, 92*, 416–425.

Zeger, S., & Karim, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79–86.

Authors

NEAL THOMAS is Senior Research Biostatistician, Bristol Myers Squibb Company, and can be reached at 61 Dream Lake Drive, Madison, CT 06443; neal_thomas@ibm.net. He specializes in missing data, observational studies, and clinical trials.

NIANCI GAN is a graduate student, Department of Statistics, North Carolina State University; ngan@eos.ncsu.edu. He specializes in spatial statistics.

Received December 13, 1995
Revision received July 17, 1995
Accepted August 12, 1996