PSYCHOMETRIKA—VOL. 67, NO. 1, 33–48
MARCH 2002

THE ROLE OF SECONDARY COVARIATES WHEN ESTIMATING LATENT TRAIT POPULATION DISTRIBUTIONS

NEAL THOMAS

DATAMETRICS, INC. AND BRISTOL-MYERS SQUIBB

The U.S. National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study (TIMSS), and the U.S. Adult Literacy Survey collect probability samples of students (or adults) who are administered brief examinations in subject areas such as mathematics and reading (cognitive variables), along with background demographic (primary) and educational environment (secondary) questions. The demographic questions are used in the primary reporting, while the numerous "explanatory" secondary variables, or "covariates", are only directly utilized in subsequent secondary analyses. The covariates are also used indirectly to create the plausible values (multiple imputations) that are an integral part of analyses because of the use of sparse matrix sampling of cognitive items. The improvement in the precision of the primary reporting due to the inclusion of the covariates is assessed here and contrasted with the precision of reporting using plausible values created using only the primary demographic variables.

The results demonstrate that the improvement in precision depends on the matrix sampling designs for the cognitive assessments. The improvements range from essentially none for the most common designs, to moderate for some less common designs. Consequently, two potential changes in the reporting procedures that could improve the statistical and operational efficiency of primary reporting are (a) eliminate or reduce the collection of covariates and increase the number of cognitive items, (b) to avoid delays, eliminate the covariates from the creation of plausible values used for the primary reports, but include them later when creating public-use files for secondary analyses. The potential improvements in statistical and operational efficiency must be weighed against the intrinsic interest in the covariates, and the potential for small discrepancies in the primary and secondary reporting.

Key words: national assessment (NAEP), item response theory, matrix sampling, multiple imputation, sample surveys, covariate design.

1. Introduction

1.1. Data Collection

The U.S. National Assessment of Educational Progress (NAEP) collects national and state probability samples of students who are administered brief examinations in cognitive subject areas such as mathematics and reading, along with background demographic and educational environment questions. Because of concerns about student motivation and the need to limit classroom disruption, examination and survey time is typically limited to one hour. Consequently, each sampled student can be administered only a few items, and sparse matrix samples are utilized to improve the generalizability of its results (Beaton & Zwick, 1992; Mislevy, Johnson, & Muraki, 1992). Because of the matrix sampling of cognitive test items, responses to most potential items are missing, but the responses to items not administered are known to be missing completely at random. Similar methods are also utilized by the Third International Mathematics and Science Study (TIMSS) and the U.S. Adult Literacy Survey.

Two common item sampling designs involve the "balanced" and "split" allocation of items (Zeger & Thomas, 1997). In a balanced design, approximately the same distribution of items are administered to each student to measure their proficiencies in different topics within a subject (e.g., 15 algebra and 5 geometry items to each student within mathematics). In a split design, the number of items measuring proficiency in each topic may differ substantially between students;

Thanks to Donald Rubin, Robert Mislevy, and John Barnard for their helpful comments and computing assistance. This work was supported by NCES Grant 84.902B980011.

Requests for reprints should be sent to Neal Thomas, 61 Dream Lake Drive, Madison CT 06443-1600. E-Mail: snthomas99@yahoo.com

0033-3123/2002-1/2000-0795-A $00.75/0

in fact, some students may not be administered items measuring some of the proficiencies (e.g., some students receive 20 algebra items, other students receive 20 geometry items). Zeger and Thomas showed that the balanced designs are the more efficient matrix sampling designs. Split designs are sometimes employed because the administration of an extended item type measuring one proficiency, such as a long reading passage, may preclude the administration of items measuring all other proficiencies.

In contrast to the matrix sampling of cognitive items, each student is administered the same extensive set of background questions. In addition, teachers and principals at the sampled schools also respond to numerous background questions. The background questions can be classified as "primary" or "secondary" variables. Primary variables involve classifications of students used in the initial reporting of results. The set of primary variables is not fixed or mandated, but for NAEP typically includes gender, race, region of the country, parent education, school type (public vs. private), and type of community (i.e., low metropolitan, high metropolitan, other). The secondary questions include the number of courses taken in a subject, home environment (e.g., "How much time do you spend watching TV?"), attitudes about the subject area (e.g., "Do you think mathematics is useful?"), school quality indicators such as the annual turn-over rate for teachers, et cetera. The number of secondary variables from all three sources (students, teachers, principals) typically exceeds 50. Reports on the secondary variables and their relationship to cognitive performance are produced following the initial publicized reporting of surveys by the primary contractor and other researchers from government, academics, and nonprofit research organizations.

Multidimensional latent trait item response (IRT) models are used to summarize and report results. To account for the incomplete information about the latent proficiencies due to the sparse matrix sampling, multiple imputation missing data methods are used (Mislevy, 1991; Rubin, 1987). These likelihood-based multivariate methods use the observed item responses and background variables to impute the missing latent proficiencies for each student. The focus here is on the role of the secondary variables (which will be called covariates) in reducing the missing information about the proficiencies to improve the precision of the primary reports.

1.2. Overview

The improvement in the precision of population estimates from the inclusion of the extensive covariates in the multivariate procedure that imputes values for the missing proficiencies is of considerable practical interest because the collection and inclusion of the extensive background variables results in:

- delays and additional expense in reporting the primary results due to the need to process the many additional variables and merge different data sources,
- more difficult numerical computation due to the high dimensionality and colinearities of the background variables,
- a reduction in the examination time available for cognitive items.

Theoretical calculations and examples from NAEP show that the use of the covariates in the imputation methods produces very little improvement in primary reporting of means and proportions exceeding cutpoints when balanced matrix sampling designs are utilized. For the purpose of primary reporting, the covariates could be eliminated when balanced designs are utilized, yielding reduced costs and potentially improved precision due to the possibility of including additional cognitive items. When the less efficient split matrix sampling designs are utilized, the addition of the covariates to the imputation models produces small to moderate improvement in the precision of the primary reporting. Thus, when split designs are used, the benefits of excluding covariates from primary reporting must be weighed more carefully. Also, the covariates and their relationship to cognitive performance are of interest.

Johnson, Mislevy, and Thomas (1994) have shown that all covariates must be included in the imputation model if analyses involving those covariates are planned. A common NAEP practice is to use principal components to represent almost all of the variability in the covariates by orthogonal predictors to improve numerical stability. This is consistent with the advice in Rubin (1987). Section 2 contains a review of the statistical methods currently used, including formulas and notation referenced in the remaining sections. Section 3 begins with a review of previous results on the use of auxiliary data to reduce missing information on the latent proficiencies. Simplifying normal approximations to the IRT models are introduced that result in a simple analytic representation for the contribution of the covariates. The accuracy of the analytic predictions in section 3 are evaluated with two examples from NAEP in Section 4. Section 5 concludes with a summary of the results and their implications.

2. Models and Estimation Methods

A set of q background predictor variables for the i-th student in a sample of size n are denoted by $\mathbf{x}_i' = (x_{i1}, \ldots, x_{iq})$, and for modelling convenience, the first component of $\mathbf{x}_i$ is a constant intercept term, and the primary variables precede the covariates.

The variables representing the cognitive items for the i-th student are denoted by $\mathbf{y}_i$, and are partitioned by the test designers into p content areas, $\mathbf{y}_i' = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ip})$, corresponding to p hypothesized latent proficiencies in a multidimensional IRT model. Each $\mathbf{y}_{ij}$ is composed of item scores, $y_{ijk}, k = 1, \ldots, s_j$, that are binary or ordinal with values coded as $0, 1, \ldots, m_{jk}$, for the k-th item measuring the j-th proficiency.

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 i-th student, which determines the distribution of the item scores through logistic IRT models, where conditional on the latent proficiencies, all item responses are assumed independent of each other and the $\mathbf{x}_i$. Second, the $\boldsymbol{\theta}_i$ are assumed to follow a multivariate normal distribution conditional on the background variables.

2.1. Logistic IRT Model for Binary and Ordinal Cognitive Responses

A logistic item response model is used for the cognitive data $\mathbf{y}_i$ conditional on $\mathbf{x}_i$, $\boldsymbol{\theta}_i$, and the item parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$, $\boldsymbol{\beta}_j = \{\boldsymbol{\beta}_{jk}, k = 1, \ldots, s_j\}$. The probabilities of binary scored items are modeled by a three parameter logistic IRT model,

$$\Pr\left(y_{ijk} = 1 \mid \boldsymbol{\theta}_i, \mathbf{x}_i, \boldsymbol{\beta}\right) = c_{jk} + \left(1 - c_{jk}\right) / \left[1 + \exp\left\{a_{jk}\left(\theta_{ij} - b_{jk}\right)\right\}\right], \tag{1}$$

where $\boldsymbol{\beta}_{jk} = (a_{jk}, b_{jk}, c_{jk})$. The response probabilities of an ordinal item are modeled by a partial credit model,

$$\Pr\left(y_{ijk} = l \mid \boldsymbol{\theta}_i, \mathbf{x}_i, \boldsymbol{\beta}\right) = \frac{\exp\left\{\sum_{h=0}^{l} a_{jk}\left(\theta_{ij} - b_{jkh}\right)\right\}}{\sum_{q=0}^{m_{jk}} \exp\left\{\sum_{h=0}^{q} a_{jk}\left(\theta_{ij} - b_{jkh}\right)\right\}}, \qquad l = 0, \ldots, m_{jk}, \tag{2}$$

where $\boldsymbol{\beta}_{jk} = \{a_{jk}, b_{jkh}, h = 1, \ldots, m_{jk}\}$.

The model invokes several independence assumptions conditional on the $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}$: (a) the $\mathbf{y}_i$ are independent of the $\mathbf{x}_i$; (b) the responses of a student to different items are independent (i.e., the distribution of the $\mathbf{y}_i$ is the product of the probabilities in (1) and (2)); (c) responses from different students are independent; and (d) $\mathbf{y}_{ij}$ are independent of $\theta_{ij'}$ conditional on θ_{ij}, $j \neq j'$, that is, item responses depend only on the proficiency to which they are assigned. With these independence assumptions, the density of $\mathbf{y}_i, i = 1, \ldots, n$ conditional on $\boldsymbol{\theta}_i$ can be represented as

$$\prod_{i=1}^{n} \left\{ \prod_{j=1}^{p} f_j\left(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j\right) \right\}, \tag{3}$$

and each $f_j(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j)$ is in turn the product of the response probabilities for the cognitive items given in (1) and (2). The innermost product in (3) 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 by design.

The $\boldsymbol{\beta}_j$ are included in the likelihood term, $f_j(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j)$, to explicitly denote the dependence of (3) on each $\boldsymbol{\beta}_j$. To identify the logistic IRT model, the mean and variance of $\boldsymbol{\theta}$ in the overall population are constrained to be zero and one, respectively. Mislevy and Bock (1982) and Muraki (1992) give details of these models.

2.2. *Distribution of the Latent Proficiency Conditional on the Background Variables*

The $\boldsymbol{\theta}_i$ vectors are assumed to be normally distributed conditional on the $\mathbf{x}_i$. The mean of this conditional distribution is given by the multivariate multiple linear regression, $\boldsymbol{\Gamma}'\mathbf{x}_i$, where $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1 \mid \cdots \mid \boldsymbol{\gamma}_p]$ and $\boldsymbol{\gamma}_j$, $j = 1, \ldots, p$ are unknown regression parameter vectors of length q. The common (unknown) p dimensional conditional variance-covariance matrix is $\boldsymbol{\Sigma}$ with elements Σ_{jk}, (Mislevy, Johnson, and Muraki, 1992). The distribution of $\boldsymbol{\theta}_i$ can be viewed as a normal prior distribution conditional on $\mathbf{x}_i$ and the parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$, before observing the cognitive data $\mathbf{y}_i$: $\phi(\boldsymbol{\theta}_i; \boldsymbol{\Gamma}'\mathbf{x}_i, \boldsymbol{\Sigma})$. As with typical regression models, the distribution of the $\mathbf{x}_i$ is assumed to be ancillary, and it is not explicitly modeled.

2.3. *Combining the Two Components of the Imputation Model*

Applying the independence assumptions, the regression model and the item response model fully specify the distribution of observed data. The likelihood function for the parameters $(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ is the distribution of the data $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, given $(\boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$:

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

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:

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

The $\boldsymbol{\theta}_i$ are regarded as "missing" data that are missing completely at random (Little & Rubin, 1987), because none are observed by design. The $\boldsymbol{\theta}_i$ are imputed for each sampled student as described in the next section. The missing item responses are not directly imputed as part of the multiple imputation procedures.

2.4. *Analyses Using Multiple Imputation*

Multiple imputation replaces each missing $\boldsymbol{\theta}_i$ with several potential values drawn from its posterior distribution producing several completed data sets. Simulations are generated from the posterior distribution in two stages. First, the parameters, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$ are generated from their posterior distribution, and second, the $\boldsymbol{\theta}_i$ are generated from their posterior distribution in (5), substituting the generated values for $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Sigma}$. The imputed $\boldsymbol{\theta}_i$ are linearly transformed to have overall means of zero and variances equal to one for each latent proficiency in each completed data set. The examples in section 4 use the NAEP procedures that employ several approximations to the two stage imputation procedure, which are described in Johnson, Mislevy, and Thomas (1994), and Thomas and Gan (1998).

The common practice is to create $n_{imp} = 5$ imputed data sets based on the recommendation of Rubin (1987) concerning inference for population estimands. For analyses of efficiencies

involving the ratios of standard errors (SE), 40 imputed data sets were formed to reduce the variability of the estimated efficiencies.

A second modification made to the usual imputation procedures was to create separate sequences of random numbers when generating the $(\boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ and the $\boldsymbol{\theta}$. This was done so the same random numbers could be used to generate the $\boldsymbol{\theta}$ from the imputation model that included the full set of covariates, and the imputation model with only the primary variables, thereby eliminating a source of random differences between the two imputation methods. This change could be also be implemented in the operational computing procedures.

An estimator of a summary of the latent proficiency distribution in a subpopulation (e.g., mean, proportion exceeding a cutpoint) computed from the n_{imp} imputed data sets will be denoted generically by $\hat{d}$. It is computed by averaging the corresponding complete data estimators from each imputed data set, $\hat{d}_1, \ldots, \hat{d}_{n_{imp}}$,

$$\hat{d} = n_{imp}^{-1} \sum_{i=1}^{n_{imp}} \hat{d}_i. \tag{6}$$

The standard error of $\hat{d}$ is computed from the usual multiple imputation standard error formula (Rubin, 1987),

$$\operatorname{var}\left(\hat{d}\right) = \operatorname{var}_c\left(\hat{d}\right) + \left(1 + n_{imp}^{-1}\right) \operatorname{var}_b\left(\hat{d}\right), \tag{7}$$

where $\operatorname{var}_c(\hat{d})$ is the average of the complete data variance estimates, $\operatorname{var}_c(\hat{d}_i)$, and

$$\operatorname{var}_b\left(\hat{d}\right) = \left(n_{imp} - 1\right)^{-1} \sum_{i=1}^{n_{imp}} \left(\hat{d}_i - \hat{d}\right)^2.$$

The complete data standard errors were computed using design effects (Johnson & Rust, 1992), as discussed in section 4. Comparisons of the efficiency of different imputation methods use ratios of the variances from the left-hand side of (7).

3. Reducing Missing Information Due to Matrix Sampling

3.1. Overview

Section 3 contains theoretical results that predict the potential of covariates to improve the precision of the primary reporting. A simplifying normal measurement error model, which approximates the more complex logistic IRT models, is described in section 3.2 and used in the subsequent sections. Section 3.3 contains a summary of the results in Zeger and Thomas (1997) about the use of multidimensional item response data to improve the estimates for each separate proficiency trait. Section 3.4 shows when secondary covariates are included, the situation is similar to multidimensional item data: If a balanced item sampling design is used, there is essentially no improvement in the precision of the proficiency mean estimators, and when a split item sampling design is used, moderate improvement is possible.

3.2. Normal Measurement Error Model

The contribution of the IRT component to the likelihood function in (5), $\prod_{j=1}^{p} f_j(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\beta}_j)$, approaches normality as the number of items increases (Chang & Stout, 1993). A convenient representation of a normal approximation to this component of the likelihood function in terms of a familiar regression model replaces the multivariate item responses, $\mathbf{y}_{ij}$, by a single measurement of the j-th proficiency for the i-th subject,

$$y_{ij} = \theta_{ij} + \delta_{ij}, \qquad \mathbf{y}_i = \left(y_{i1}, \ldots, y_{ip}\right)', \tag{8}$$

where $\delta_{ij} \sim \phi(0, \tau_{ij})$ are independent normally distributed "measurement" errors with known variances, τ_{ij} (which corresponds to fixing the item parameters $\boldsymbol{\beta}_j$ at their maximum likelihood estimates (MLEs) and treating them as known, a common practice in NAEP evaluations). The y_{ij} and τ_{ij} are constructed so that the likelihood function arising from (8) approximates the one in (5). Mislevy (1992) and Thomas (1993) show that useful approximations to the logistic IRT model can be obtained with a suitable choice of the τ_{ij}.

The variance of y_{ij} including the variance of θ_{ij} is

$$\text{var}\left(y_{ij} \mid \mathbf{x}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}\right) = \Sigma_{jj} + \tau_{ij},$$

and the covariance of y_{ij} and $y_{ij'}$ is $\Sigma_{jj'}$ when $j \neq j'$. The mean of $\mathbf{y}_i$ is $\boldsymbol{\Gamma}'\mathbf{x}_i$, the same as that of $\boldsymbol{\theta}_i$.

The model in (8) includes the setting where there are no measurements for the j-th proficiency as a limiting case with $\tau_{ij} \to \infty$, and the setting with exact measurement by $\tau_{ij} \to 0$. Maximum likelihood estimators for the model in (8) are generalized least squares estimators (Johnson, 1984).

3.3. Reducing Missing Information Due to Matrix Sampling Using Multivariate Item Data

This section summarizes the contribution of items measuring one proficiency to the estimation of the other proficiencies. The key results are (a) with balanced designs, there is no information gain when using simultaneous estimation of multivariate mean proficiencies relative to univariate procedures, (b) despite the lack of information gain, the balanced designs are optimal, and (c) with split designs, there is information gained by estimating the mean of each proficiency simultaneously, with the improvement increasing as the correlation between the proficiencies increases.

When the proficiencies are highly correlated, intuition based on common examples suggests that items measuring one proficiency can contribute information about the second proficiency if it is unknown or only partially measured. Consider estimation of the overall means of the two traits (θ_1, θ_2), which will be denoted by (μ_1, μ_2), without background covariates. An example in sections 6.2 and 6.3 of Little and Rubin (1987) illustrates the potential for substantial improvement in estimators of μ_2 using data measuring θ_1. Their example is a limiting case of the model in (8) in which all subjects have θ_1 measured without error ($\tau_{i1} = 0, y_{i1} = \theta_{i1}, i = 1, \ldots, n$), n_{obs} subjects have θ_2 measured without error ($\tau_{i2} = 0, y_{i2} = \theta_{i2}, i = 1, \ldots, n_{obs}$), and $n - n_{obs}$ subjects have no data on θ_2, ($\tau_{i2} = \infty, i = (n_{obs} + 1), \ldots, n$). The subjects without measurements of θ_2 are selected either completely at random and thus the missing θ_2 are missing completely at random(MCAR), or they are selected dependent on the observed measurements so the missing y_2 depend only on the observed data (MAR, i.e., missing at random). The MLE of μ_2 is

$$\hat{\mu}_2 = n^{-1}\left(\sum_{i=1}^{n_{obs}} y_{i2} + \sum_{i=n_{obs}+1}^{n} \hat{y}_{i2}\right) \tag{9}$$

where the predicted values in (9) are based on the regression of the y_{i2} on y_{i1} among the n_{obs} subjects with complete data.

The measurements for the first proficiency contribute to the estimator of μ_2 because the predicted values are a function of y_{i1}. Little and Rubin show that the contribution can produce substantial improvement, depending on the correlation between (y_1, y_2). As the correlation approaches one, the missing information about μ_2 is completely replaced by the data from the first proficiency. In situations where the data are MAR but not MCAR, the bivariate estimator can also reduce bias relative to the complete cases univariate estimator, $n_{obs}^{-1}\sum_{i=1}^{n_{obs}} y_{i2}$.

Zeger and Thomas (1997) show that similar, though more complex results obtain when the bivariate proficiencies are measured with error, that is, when $\tau_{i1} > 0, i = 1, \ldots, n$, $\tau_{i2} > 0, i = 1, \ldots, n_{obs}$, and $\tau_{i2} = \infty, i = n_{obs} + 1, \ldots, n$. The improvement in precision of $\hat{\mu}_2$ due to

the use of the first proficiency is less, however, because the effective correlation determining the improvement is between (y_{i1}, θ_{i2}), and this correlation is attenuated by the measurement error in y_{i1}.

The missing data design considered by Little and Rubin is closely related to the "split" design in which $m/2$ students are assigned items measuring the first proficiency only, another $m/2$ students receive items measuring the second proficiency only, and the remaining $n - m$ students receive an equal number of items measuring each of the two proficiencies. The variance of the measurements focused on a single proficiency will be denoted by τ, and assuming that items have equal measurement error variances independent of the proficiency levels being measured, the variance of the measurements among students receiving items divided equally between the two proficiencies will be 2τ. The 1992 4th grade Reading assessment was a split design with $m = 0.75n$. Zeger and Thomas showed that the improvement in $\hat{\mu}_2$ due to the use of items measuring the first proficiency was approximately 10%.

In a balanced design, the number of items measuring each proficiency may differ (e.g., 15 algebra, 5 geometry), but is the same for each student. The distribution of the (y_{i1}, y_{i2}) in (8) reduces to the standard form of bivariate multiple regression with homogeneous variances. The MLE of the multivariate normal mean is the same as the univariate MLE for each mean in this setting (Rao, 1965), and thus there is no improvement in the estimation of the second proficiency due to the items measuring the first proficiency. The lack of improvement occurs even when there is high correlation between the proficiencies. Zeger and Thomas showed, for example, the gain in precision in the 1990 NAEP Mathematics assessment from the use of the multivariate MLE relative to the univariate MLE was approximately 1% even though the correlation between the proficiencies was approximately 0.9. Despite the lack of gain from multivariate estimation, the balanced design is optimal among a large class of matrix sampling designs that includes many split designs.

3.4. Reducing Missing Information Due to Matrix Sampling Using Covariates

Calculations in this section show that the contribution of the covariates to the estimation of the mean proficiencies within primary reporting subpopulations depends on the design of the matrix sampling of items in a manner similar to that of the simultaneous multivariate estimation of several correlated proficiencies. To simplify the analytic results, only two binary covariates will be considered,

$$\begin{aligned} x_{i1} = 0, 1 \qquad & \mathbf{x}_1' = (x_{i1}, \ldots, x_{n1}), \\ x_{i2} = 0, 1 \qquad & \mathbf{x}_2' = (x_{12}, \ldots, x_{n2}), \end{aligned}$$

where $\mathbf{x}_1$ is the primary reporting variable, and $\mathbf{x}_2$ is the covariate. To further simplify notation and formulas, estimation will be restricted to the mean of a single proficiency. The subscript indicating the proficiency dimension is suppressed throughout, so the proficiency for the i-th student is denoted by θ_i, and it is measured by y_i with error variance τ_i and population variance Σ.

The measurements for the proficiency will be modeled by a saturated model for the means,

$$\begin{aligned} \mathrm{E}(\theta_i \mid x_{i1}, x_{i2}) &= \mathrm{E}(y_i \mid x_{i1}, x_{i2}) = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_{12} x_{i1} x_{i2}, \\ \mathrm{var}(y_i \mid x_{i1}, x_{i2}) &= \Sigma + \tau_i. \end{aligned} \tag{10}$$

The estimand is the mean of θ among students with $x_1 = 1$, denoted by $\mu_{(1)}$:

$$\mathrm{E}(\theta \mid x_1 = 1) = \mathrm{E}(y \mid x_1 = 1) = \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_{12})\,\lambda \equiv \mu_{(1)}.$$

The additional parameter, λ, determining $\mu_{(1)}$ is the conditional probability that $x_2 = 1$ given $x_1 = 1$. Analogous results obtain for the estimators of the mean of θ with $x_1 = 0$ and are thus not presented.

The first design considered corresponds to the balanced design in which $\tau_i \equiv 2\tau, i = 1, \ldots, n$. The estimator of $\mu_{(1)}$ if the covariate data are ignored is the mean of y among subjects with $x_1 = 1$,

$$\tilde{\mu}_{(1)} \equiv n_1^{-1} \sum_{i=1}^{n} y_i I(x_{i1} = 1),$$

where $I(\ )$ is the indicator function, and $n_1 = \sum_{i=1}^{n} I(x_{i1} = 1)$. When x_2 is utilized, the weighted average across the two values of x_2, which is the MLE when the proficiency and measurement errors are normally distributed, is given by

$$\hat{\mu}_{(1)} \equiv \overline{y}_{11}\hat{\lambda} + \overline{y}_{10}\left(1 - \hat{\lambda}\right),$$

where

$$n_{1k} = \sum_{i=1}^{n} I(x_{i1} = 1)\, I(x_{i2} = k) \qquad \overline{y}_{1k} = n_{1k}^{-1} \sum_{i=1}^{n} y_i I(x_{i1} = 1)\, I(x_{i2} = k),\ k = 0, 1$$

and $\hat{\lambda} = n_{11}/n_1$. Using $n_{10} = n_1 - n_{11}$, it is easy to check that $\tilde{\mu}_{(1)} = \hat{\mu}_{(1)}$, so there is no gain for the primary reporting subpopulations as a consequence of including the covariate when the data are collected from a balanced matrix sample design.

In a simplified split design, the covariate data are observed for all students, but the proficiencies are measured only for a subset of n_{obs} students selected completely at random. The measurements for the n_{obs} students are recorded first and assumed to have the same variability, $\tau_i \equiv \tau$. The sample moments and proportions in the subpopulation with $x_{i1} = 1$ among the n_{obs} students with measured y values are denoted by

$$n_{obs_1} = \sum_{i=1}^{n_{obs}} I(x_{i1} = 1) \qquad n_{obs_{1k}} = \sum_{i=1}^{n_{obs}} I(x_{i1} = 1)\, I(x_{i2} = k),\ k = 0, 1$$

$$\overline{y}_{nobs_1} = n_{obs_1}^{-1} \left\{ \sum_{i=1}^{n_{obs}} y_i I(x_{i1} = 1) \right\},$$

$$\overline{y}_{nobs_{1k}} = n_{obs_{1k}}^{-1} \left\{ \sum_{i=1}^{n_{obs}} y_i I(x_{i1} = 1)\, I(x_{i2} = k) \right\},\ k = 0, 1$$

$$\hat{\lambda}_{nobs} = n_{obs_{11}}/n_{obs_1}.$$

The estimator of $\mu_{(1)}$ when the $\mathbf{x}_2$ covariate is ignored is the mean of the observed responses in the subpopulation, $\tilde{\mu}_{(1)} \equiv \overline{y}_{nobs_1}$, and the MLE when the $\mathbf{x}_2$ covariate is utilized is a weighted average of the means of the observed responses within the more refined subpopulations formed by $\mathbf{x}_2$, $\hat{\mu}_{(1)} \equiv \overline{y}_{nobs_{11}}\hat{\lambda} + \overline{y}_{nobs_{10}}(1 - \hat{\lambda})$. Note that the weights in $\hat{\mu}_{(1)}$ are based on complete data for $(\mathbf{x}_1, \mathbf{x}_2)$, $\hat{\lambda}$, so $\tilde{\mu}_{(1)} \neq \hat{\mu}_{(1)}$, unlike the situation with the balanced design.

The variances of $\tilde{\mu}_{(1)}$ and $\hat{\mu}_{(1)}$ conditional on $\mathbf{x}_1$ are obtained by first computing the expectations of the variances conditional on $(\mathbf{x}_1, \mathbf{x}_2)$, and then the corresponding variances of the conditional expectations. The variance of the estimator $\tilde{\mu}_{(1)}$ conditional on $(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\mathrm{var}\left(\tilde{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right) = \frac{\Sigma + \tau}{n_{obs_1}}.$$

The corresponding variance of $\hat{\mu}_{(1)}$ is

$$\mathrm{var}\left(\hat{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right) = \hat{\lambda}^2 \left(\frac{\Sigma + \tau}{n_{obs_{11}}}\right) + \left(1 - \hat{\lambda}\right)^2 \left(\frac{\Sigma + \tau}{n_{obs_{10}}}\right).$$

Because the subjects without measurements on θ are MCAR,

$$\hat{\lambda} = \frac{n_{11}}{n_1} \approx \frac{n_{obs_{11}}}{n_{obs_1}} = \hat{\lambda}_{nobs}, \qquad 1 - \hat{\lambda} = \frac{n_{10}}{n_1} \approx \frac{n_{obs_{10}}}{n_{obs_1}} = 1 - \hat{\lambda}_{nobs},$$

so

$$\begin{aligned} \mathrm{var}\left(\hat{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right) &\approx \hat{\lambda}_{nobs} \frac{(\Sigma + \tau)}{n_{obs_1}} + (1 - \hat{\lambda}_{nobs}) \frac{(\Sigma + \tau)}{n_{obs_1}} \\ &= \frac{\Sigma + \tau}{n_{obs_1}} \\ &= \mathrm{var}\left(\tilde{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right). \end{aligned}$$

The contributions of the variances conditional on $(\mathbf{x}_1, \mathbf{x}_2)$ to the variances of the estimators are thus very similar for $\tilde{\mu}_{(1)}$ and $\hat{\mu}_{(1)}$.

The conditional expectation of $\tilde{\mu}_{(1)}$ given $(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\mathrm{E}\left(\tilde{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right) = \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_{12}) \left(\frac{n_{obs_{11}}}{n_{obs_1}}\right) = \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_{12})\, \hat{\lambda}_{nobs},$$

so the variance of this conditional expectation computed with respect to $\mathbf{x}_2$ is

$$\mathrm{var}\left(\mathrm{E}\left(\tilde{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right)\right) = (\gamma_2 + \gamma_{12})^2 \frac{\lambda(1 - \lambda)}{n_{obs_1}}.$$

The conditional expectation of $\hat{\mu}_{(1)}$ given $(\mathbf{x}_1, \mathbf{x}_2)$ is

$$\begin{aligned} \mathrm{E}\left(\hat{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right) &= \hat{\lambda}\, (\gamma_0 + \gamma_1 + \gamma_2 + \gamma_{12}) + \left(1 - \hat{\lambda}\right) (\gamma_0 + \gamma_1) \\ &= \gamma_0 + \gamma_1 + (\gamma_2 + \gamma_{12})\, \hat{\lambda}, \end{aligned}$$

so the corresponding variance is

$$\mathrm{var}\left(\mathrm{E}\left(\hat{\mu}_{(1)} \mid \mathbf{x}_1, \mathbf{x}_2\right)\right) = (\gamma_2 + \gamma_{12})^2 \frac{\lambda(1 - \lambda)}{n_1}.$$

The variances of $\tilde{\mu}_{(1)}$ and $\hat{\mu}_{(1)}$ thus differ by the variances of the estimators of the proportion λ. The improved precision due to the use of the covariate in a split design is determined by this difference and the strength of the dependence of y on $\mathbf{x}_2$, that is, $[\gamma_2 + \gamma_{12}]$. Gains can be substantial when there are many subjects without measurements and the covariates are highly predictive of y. The adjustment to the estimator, $\hat{\mu}_{(1)}$, obtained from the use of the $\mathbf{x}_2$ covariate, is a form of post-stratification (Cochran, 1977), which uses the improved estimate of the distribution of the $\mathbf{x}_2$ covariate based on the entire sample.

Note that in actual applications, the improved precision will be less because the expressions for the variances will represent residual variances after adjustment for the measurements of other proficiencies. Although simple analytic expressions are available only for estimators of means, results in section 4 for estimators of proportions exceeding cutpoints are very similar to those for means.

4. Two Examples from NAEP

4.1. Description of Data Sets

4.1.1. 1992 8th Grade National Mathematics Assessment

A sample of 10,291 U.S. students who were in the 8th grade or who were 13 years old was collected for the 1992 8th Grade National Mathematics Assessment. There were five proficien-

TABLE 1.
Latent proficiencies

8th Grade Mathematics Assessment	
Label	Description
NUMOP	Numbers and Operations
MEAS	Measurements
GEOM	Geometry
DAST	Data analysis and statistics
ALGE	Algebra
COMP	Sum of the other proficiencies
4th Grade Reading Assessment	
LIT	Reading for literary experience
INFO	Reading to gain information
COMP	Sum of the two proficiencies

Note: The proficiencies are from the 1992 National Assessment.

cies reported, as described in Table 1, which were measured by a total of 368 different cognitive items. Each form administered had approximately the same distribution of items measuring the different proficiencies (i.e., a balanced design). The item parameter estimates, $\hat{\boldsymbol{\beta}}$, discussed in section 2.4, were copied from the actual assessment and were used for all of the imputation models. A detailed description of the sample and its contents are in Johnson and Carlson (1994).

4.1.2. 1992 4th Grade Reading Assessment

A sample of 8,416 U.S. students who were either in the 4th grade or who were 9 years old was collected for the 1992 4th Grade National Reading Assessment. There were two proficiencies reported, Reading for Literary Experience (LIT) and Reading to Gain Information (INFO), which were measured by a total of 85 items. There were three types of matrix sampling forms for the items: (a) forms with items measuring the LIT proficiency, given to 38% of the students; (b) forms with items measuring the INFO proficiency, given to 37% of the students; and (c) forms with items measuring both proficiencies, given to 25% of the students (i.e., a split design). The other characteristics of the Reading assessment were similar to the Mathematics assessment.

4.2. Covariates for the Regression Models

The primary reporting variables are described in Table 2; they are the same in both examples. The covariates included are a large subset of all possible covariates. A representative sample of 35 questions of different types represented by 104 indicator variables were included in the Mathematics example, and there were 39 questions represented by 119 indicator variables for the Reading example; the specific covariates are in Thomas (2000). Not all covariates were included to reduce the dimensionality and avoid colinearities in the complete set of covariates. NAEP often utilizes more complex principal component methods to reduce dimensionality while representing almost all of the variation in the covariates, which is necessary when a secondary analysis involving any subset of the covariates is possible. The imputation model based on the more complete set of the covariates will be called the full model, and the imputation model based only on the primary variables will be call the primary-only model.

TABLE 2.
Latent proficiencies

Symbol	Description
DSEX1	Male
DSEX2	Female
DRACE1	White/Other
DRACE2	African American
DRACE3	Hispanic
DRACE4	Asian American
STOC1	Low metropolitan
STOC2	High metropoitan
STOC3	Other
REGION1	Northeast
REGION2	Southeast
REGION3	Central
REGION4	West
PARED1	Parent education less than high school
PARED2	Parent education equals high school
PARED3	Parent education includes post high school
PARED4	Parent education equals college graduate
PARED5	Parent education unknown or missing
MODGRAG1	Modal grade/less than modal age
MODGRAG2	Less than modal grade/modal age
MODGRAG3	Modal grade/modal age
MODGRAG4	Greater than modal grade/modal age
MODGRAG5	Modal grade/greater than modal age
SCHTYPE1	Public
SCHTYPE2	Private

Note: The variables represent the coding of the primary reporting categories in all of the imputation models.

4.3. Evaluation of Efficiency

4.3.1. Mathematics Example

The mean and the proportion of students exceeding the overall 75th percentile estimates were computed using (6) and the standard errors using (7) for each subpopulation defined by the primary variables in Table 2. Efficiency was measured by the ratio of the SE (squared) based on the full imputation model divided by the SE (squared) from the primary-only model.

A design effect of 2.0 was used to compute the complete-data component of the SE, which is typical of NAEP experience, and the resulting contribution of the sampling variability to the total variation in the Mathematics example closely matches the operational results based on a jackknife estimator in Table 13 through 34 of Johnson and Carlson (1994); the sampling variance is approximately 80–90% of the total variance for these proficiencies.

The results for the estimated means are in Table 3, and the results for the proportions are in Table 4. A value less than one indicates that the primary-only imputation model was less efficient, and a value greater than one favors the primary-only imputation model. As predicted by the theoretical calculations in section 3, there is no indication that the inclusion of the covariates increased the precision of estimators of summaries of the subpopulations defined by the primary

TABLE 3.
Efficiency of subpopulation mean estimates: Mathematics example

Subpopulation	NUMOP	MEAS	GEOM	DAST	ALGE	COMP
DSEX1	1.00	1.01	1.00	1.00	1.02	1.00
DSEX2	0.98	1.01	0.99	0.97	1.01	0.99
DRACE1	0.97	0.98	0.95	1.05	0.94	0.99
DRACE2	0.97	1.06	0.84	1.00	0.88	1.03
DRACE3	0.96	0.98	0.94	1.02	0.92	1.00
DRACE4	1.05	0.95	0.87	1.19	0.99	1.02
STOC1	0.99	0.91	0.94	0.93	0.90	1.01
STOC2	0.83	1.09	1.09	0.94	1.07	0.98
STOC3	0.97	1.00	1.00	0.99	0.97	1.00
REGION1	1.01	0.99	1.02	1.18	1.05	1.06
REGION2	1.00	0.93	0.91	1.03	0.95	0.99
REGION3	0.93	1.07	0.91	1.06	1.06	1.00
REGION4	1.02	1.04	1.04	1.07	0.92	0.99
PARED1	1.04	0.99	1.13	0.88	1.02	1.01
PARED2	0.98	0.84	0.97	1.03	0.97	0.99
PARED3	0.92	1.01	1.04	1.05	0.89	0.98
PARED4	0.98	0.99	1.04	1.00	0.96	1.03
PARED5	1.00	0.84	0.83	0.93	0.89	0.90
MODGRAG1	0.98	1.03	1.08	1.01	1.04	1.03
MODGRAG2	1.05	1.00	1.00	0.95	0.98	0.99
MODGRAG3	1.07	0.95	1.01	1.06	1.08	1.00
MODGRAG4	1.10	1.03	1.04	1.08	1.14	1.09
MODGRAG5	0.99	0.98	1.13	1.06	1.04	1.00
SCHTYPE1	0.98	0.99	1.00	1.03	0.99	1.00
SCHTYPE2	0.89	0.99	1.01	1.15	0.96	0.99

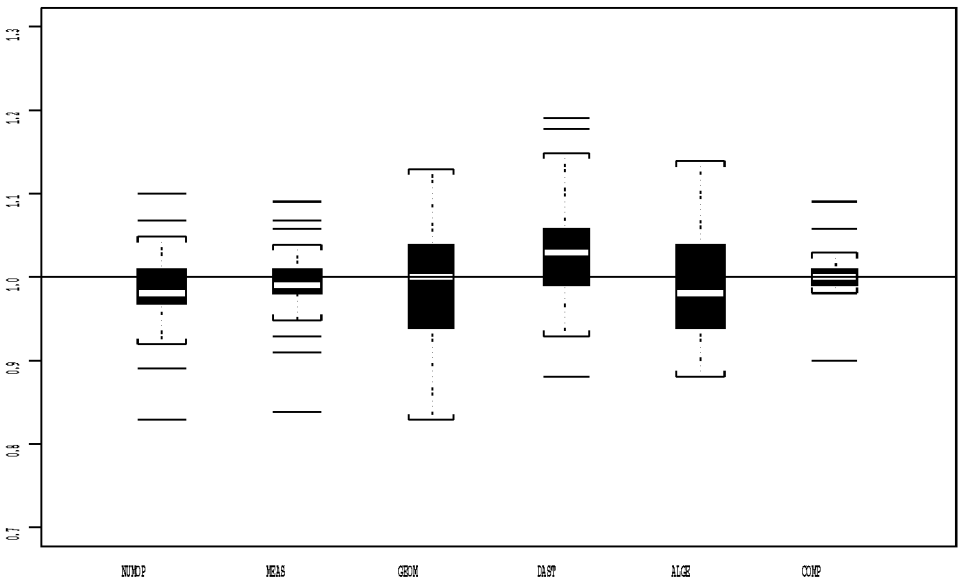
Note: Efficiency is the SE (squared) based on the full imputation model divided by the SE (squared) based on the primary-only model. Values < 1 favor the full model.

reporting variables with the balanced matrix sampling design. Even though the standard errors are based on 40 imputed data sets, there is still considerable variability in the standard error estimates. A boxplot in Figure 1(a) summarizes the efficiencies of the estimates of the proportion exceeding the 75th percentile cutpoint in Table 4. The distribution of the efficiency estimates is centered very close to one, the value representing no difference in efficiency for the methods.

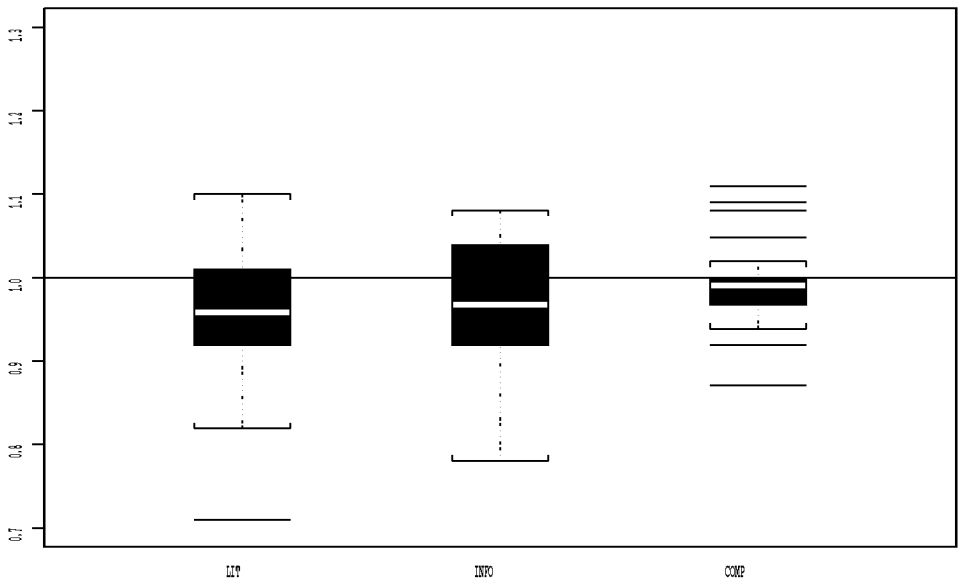
Note that the failure to gain efficiency from the inclusion of the covariates occurred despite the fact that they are predictive of cognitive performance. The r^2 of the primary reporting variables ranged from 0.28 to 0.37 for the five latent proficiencies, and the r^2 when the covariates were added to the primary variables ranged from 0.49 to 0.67. The moderate improvement in r^2 was also evident in the posterior variances for each student's latent proficiencies. The average posterior variance (across all students in the sample) was 4–11% smaller for the five latent proficiencies when the full imputation model was used compared to the primary-only model.

4.3.2. Reading Example

Analyses similar to the Mathematics example were performed for the Reading example. A design effect of 2.0 was also used to compute the complete data component of the SE, resulting in good agreement with the contribution of the sampling variability to the total variability



(a) Mathematics assessment



(b) Reading assessment

FIGURE 1.
Boxplot of efficiencies of proportion estimators. Efficiency is the SE (squared) based on the full imputation model divided by the SE (squared) based on the primary-only model. Values < 1 favor the full model.

in Table 12 through 25 of Johnson and Carlson (1994); the sampling variability is 85–90% of the total variance. The results for the estimated means and proportions are in Table 5, and the efficiencies for the proportion estimator are displayed in Figure 1(b). The results indicate that the expected gain in efficiency using the covariates for the Reading design is roughly 5% for both the mean and proportion estimators. A gain of 5% corresponds to a large reduction in the measurement error variance, but only a moderate reduction in the total variance.

TABLE 4.
Efficiency of subpopulation proportion estimates: Mathematics example

Subpopulation	NUMOP	MEAS	GEOM	DAST	ALGE	COMP
DSEX1	1.03	0.99	1.06	0.87	1.11	1.01
DSEX2	1.03	0.89	1.04	1.07	0.98	0.99
DRACE1	0.91	0.93	1.03	1.03	0.95	1.03
DRACE2	1.03	0.96	0.93	1.00	0.83	0.93
DRACE3	1.01	1.03	0.93	0.91	1.05	1.08
DRACE4	1.05	0.90	0.89	1.16	0.95	0.95
STOC1	0.99	0.93	1.11	1.00	1.04	1.03
STOC2	0.93	1.01	1.00	1.01	0.96	1.01
STOC3	0.97	1.00	1.07	1.03	1.08	1.01
REGION1	0.92	1.02	0.98	1.11	0.93	0.97
REGION2	1.05	0.99	1.03	1.02	0.90	1.00
REGION3	0.94	1.01	0.93	1.10	1.00	0.97
REGION4	0.95	1.08	0.97	1.12	0.90	1.00
PARED1	0.96	1.02	1.11	0.90	1.03	0.98
PARED2	0.95	0.98	1.00	1.03	0.88	1.01
PARED3	1.04	1.05	1.03	1.06	0.95	0.97
PARED4	1.00	0.97	1.10	0.95	1.00	0.99
PARED5	0.96	0.91	1.02	0.86	1.02	0.99
MODGRAG1	1.03	1.10	1.02	0.98	1.04	0.98
MODGRAG2	0.98	0.92	1.03	1.00	0.94	1.00
MODGRAG3	1.06	0.92	1.07	1.13	1.09	1.03
MODGRAG4	1.01	1.04	1.00	0.94	0.91	0.98
MODGRAG5	1.03	1.07	1.05	1.06	0.99	1.04
SCHTYPE1	1.00	0.94	1.06	0.99	1.04	1.00
SCHTYPE2	0.95	0.94	1.05	0.98	0.97	0.96

Note: Efficiency is the SE (squared) based on the full imputation model divided by the SE (squared) based on the primary-only model. Values < 1 favor the full model.

The r^2 for the primary-only model is 0.27 and 0.30 for the LIT and INFO proficiencies respectively, and 0.49 and 0.55 for the full model, which is comparable to the r^2 for the Mathematics assessment. The reductions in the average posterior variances were 7% and 11% for the LIT and INFO proficiencies, respectively. The improved efficiency due to the covariates in the Reading assessment is attributable to the use of the split matrix sampling design, and not to more predictive covariates.

5. Conclusions

The improvement in the efficiency of the estimated means and proportions exceeding cutpoints, which are featured in the primary reporting, due to the use of covariates during the creation of imputed values depends on the item sampling design utilized. When balanced item sampling designs are used, little or no gain is anticipated due to the inclusion of covariates. When split designs are used, small to moderate gains are anticipated. There is less improvement for the composite scores commonly reported compared to estimates for the more refined proficiencies.

The results show that for the purpose of primary reporting, the collection of covariates could be reduced or eliminated when utilizing the common balanced design. Such a reduction might

TABLE 5.
Efficiency of subpopulation mean and proportion estimates: Reading example

	MEAN			PROPORTION		
Subpopulation	LIT	INFO	COMP	LIT	INFO	COMP
DSEX1	0.92	1.05	0.98	0.93	1.01	0.99
DSEX2	0.92	1.06	0.99	0.93	1.10	0.98
DRACE1	0.96	0.98	0.99	0.94	0.96	0.96
DRACE2	0.71	1.03	0.92	0.88	0.95	0.99
DRACE3	1.01	0.94	1.00	0.98	0.93	0.97
DRACE4	1.07	0.86	1.08	1.04	0.92	1.00
STOC1	1.05	1.03	0.97	1.07	0.99	0.97
STOC2	0.86	1.06	0.97	0.96	0.91	1.01
STOC3	1.01	1.02	1.00	1.00	1.02	0.99
REGION1	1.01	1.08	1.09	0.97	1.07	1.06
REGION2	0.96	0.89	0.96	0.99	0.90	0.98
REGION3	0.97	0.94	1.00	1.00	0.96	0.98
REGION4	1.02	1.04	1.05	1.01	1.03	1.01
PARED1	0.87	0.93	1.00	0.84	1.02	0.94
PARED2	0.89	0.94	0.99	0.85	0.96	0.95
PARED3	1.04	0.91	0.99	0.96	0.91	0.97
PARED4	0.99	0.92	1.00	1.04	0.92	1.02
PARED5	1.10	0.88	0.98	0.95	0.83	0.97
MODGRAG1	1.00	1.07	1.11	1.00	1.28	0.94
MODGRAG2	0.82	0.88	0.96	0.86	0.82	0.95
MODGRAG3	0.95	0.96	0.98	1.03	0.88	0.97
MODGRAG4	0.94	0.78	0.87	1.07	0.94	1.02
MODGRAG5	0.95	0.97	0.94	1.03	0.99	0.97
SCHTYPE1	0.99	1.01	1.00	0.96	0.95	1.05
SCHTYPE2	0.95	1.05	1.02	0.86	1.11	1.01

Note: Efficiency is the SE (squared) based on the full imputation model divided by the SE (squared) based on the primary-only model. Values < 1 favor the full model.

slightly increase the precision of primary reporting by allowing increased time for cognitive testing. There are many alternative strategies that involve only a partial reduction in the collection of covariates, for example, reducing the frequency with which they are collected. The reduced complexity of the analyses that could be achieved by reducing the number of background variables must be weighed against the intrinsic interest in obtaining these variables in nationally representative samples.

Another strategy to reduce the number of covariates that could simplify and shorten the effort required to prepare primary reports (Forsyth, Hambleton, Linn, & Mislevy, 1996) would exclude most of the covariates during the creation of imputations for the primary reporting. Summaries of the distribution of covariates, which are the most common type of reporting of covariates, would be unaffected by the change. The covariates could then be used in the generation of imputed values for a follow-up data file created for secondary analyses that correlate covariates with cognitive performance. A problem with this approach is that primary analyses re-derived on the secondary files will differ slightly from the primary reports. Evaluations in Thomas (2000) show that most of the differences between estimates based on the two different sets of imputations will be less than 1/4 standard error, although an occasional difference as large as one

standard error can occur. The seriousness of such discrepancies is debatable. It can be argued that they reduce the credibility of the reporting, but similar revisions of reports are common in other government data systems.

References

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

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

Cochran, W.G. (1977). *Sampling techniques*. New York, NY: Wiley.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). 1996 Design/Feasibility Team: Report to the National Governing Board. Washington, D.C.: National Governing Board.

Johnson, E.G., & Carlson, J. (1994). *The NAEP 1992 Technical Report*. Washington D.C.: National Center for Educational Statistics.

Johnson, E.G., Mislevy, R., & Thomas, N. (1994), Scaling procedures. In E.G. Johnson & N.L. Allen (Eds.), *The NAEP 1992 Technical Report* (Report 23-TR-20, pp. 241–256). Washington D.C.: National Center for Educational Statistics.

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

Johnson, J.G. (1984). *Econometric Methods*. New York, NY: McGraw–Hill.

Little, R., & Rubin, D. (1987) *Statistical analysis with missing data*. New York, NY: Wiley.

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

Mislevy, R. (1992). Scaling procedures. In E.G. Johnson & N.L. Allen (Eds.), *The NAEP 1990 Technical Report* (Report 21-TR-20, pp. 199–214). Washington D.C.: National Center for Educational Statistics.

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

Mislevy, R., Johnson, E.G., & Muraki, E. (1992), Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Statistics*, *17*, 131–154.

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

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

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

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

Thomas, N. (2000). The role of secondary covariates in NAEP primary reporting (NCES Technical Report). Washington D.C.: National Center for Educational Statistics.

Thomas, N., & Gan, N. (1998). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, *23*, 425–445.

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*, 416–438.

Manuscript received 28 FEB 2000
Final version received 8 NOV 2000