Research Report

Estimation of Standard Error of Regression Effects in Latent Regression Models Using Binder's Linearization

Deping Li

Andreas Oranje

Research & Development

March 2007
RR-07-09

Estimation of Standard Error of Regression Effects in Latent Regression Models Using Binder's Linearization

Deping Li and Andreas Oranje
ETS, Princeton, NJ

March 2007

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

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

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

Abstract

Two versions of a general method for approximating standard error of regression effect estimates within an IRT-based latent regression model are compared. The general method is based on Binder's (1983) approach, accounting for complex samples and finite populations by Taylor series linearization. In contrast, the current National Assessment of Educational Progress procedure assumes a simple random sample for standard error of regression effects and applies a jackknife estimator to statistics of interest as a way to account for NAEP's complex sample. In this study, the versions of the general method are formally defined and the general method is extended to multiple dimensions. Furthermore, they are applied in an empirical study to the 2004 NAEP long-term trend data comparing both large, nearly saturated, and small models. Subsequently, the results are compared to the operational-based imputation method. Results show no impact on the imputation-based results, limited impact on large models, and reasonable impact on small models. While it is not readily apparent to what this differential impact can be attributed, several explanations are discussed.

Key words: Variance estimation, latent regression, Taylor series approximation, information matrix, cluster sampling, National Assessment of Educational Progress

1 Introduction

National Assessment of Educational Progress (NAEP) data are analyzed using a latent regression model (Mislevy, 1984, 1985). In this model, various student characteristic variables are regressed onto a latent ability:

$$\theta_{it} = \boldsymbol{\gamma}_t \boldsymbol{x}_i + \varepsilon_i \tag{1}$$

where θ is proficiency for student i on subscale t. Furthermore, the regression coefficients for scale t are represented as γ_t and $\boldsymbol{x}_i$ is a vector of student characteristic variables. Finally, ε_i is a residual term, assumed to be normal distributed. Latent abilities can be inferred from student responses to items using a set of item response models (e.g., Lord & Novick, 1968) with a priori estimated item parameter values. Specifically, a three-parameter logistic (3PL) model is used for multiple-choice items and a generalized partial credit (GPC) model for constructed-response items. Under these constraints, a closed form solution does not exist and an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) is employed to conduct the parameter estimation.

The sampling design of the NAEP assessment follows a multistage stratified scheme. In the first stage, strata are selected, which in most samples are states or regions by metropolitan area status. Each stratum contains a large number of primary sampling units, which are (groups of) counties. Commensurate to size, primary sampling units are drawn and within those units—again commensurate to size—schools are drawn. The final stage contains a simple random sample of students within each drawn school. For some samples, schools are the primary sampling units and one less stage is conducted. Because students within schools are exposed to similar instructional practices and learning experiences and often share demographic characteristics, intraschool correlations are expected to be substantial.

In NAEP, standard error estimates associated with the parameters of this model are less than straightforward to obtain. Despite the sampling design described in the previous paragraph, when approximating standard error of regression effects, the current methodology assumes for estimation purposes that examinees have been selected from a simple random sample and uses the fact that the posterior variance of the regression effects and the variance of regression parameter estimates are essentially equivalent. The posterior variance of the regression effects is further estimated by the sum of the following two parts: (a) the sampling variance, and (b) a component of variation that reflects the uncertainty due to the fact that the examinee ability values have

not been observed directly (Mazzeo, Donoghue, Li, & Johnson, 2006). The resulting standard error and previously obtained point estimates are then used to define a posterior distribution, and imputations for the ability are drawn from this posterior distribution (Mislevy, 1991). More specifically, the imputations are draws from the distribution of the latent ability given all of the model parameters, student characteristics, and item responses.

Finally, NAEP utilizes imputed examinee ability values and the jackknife method to compute the standard errors of the major reporting statistics (e.g., subpopulation means and percentage above a certain level of performance) taking the complex sample into account. For more detailed descriptions on these procedures, interested readers are referred to the NAEP technical reports (e.g., Allen, Donoghue, & Schoeps, 2001). While in theory the standard error estimates do not directly affect the estimates for NAEP major reporting statistics, there are some concerns raised related to to the simple random sample assumption for the imputation model.

This paper explores some approaches to account for the complex sample in the standard errors of the regression parameters. Possibly, regression effects can be reported in addition to the aggregates derived from the imputation model.

1.1 Current Methodology

With NAEP, students answer a small portion of the cognitive tasks to limit testing time. Hence, individual estimates are relatively imprecise. However, NAEP collects student and school background variables and uses this information to interpret the expectation of student abilities. A student record of background variables, denoted as $\boldsymbol{x}_i$ for $i = 1, \cdots, N$, includes Q observations, where N is the total number of students in the assessment. That is, $\boldsymbol{x}_i = (X_{i1}, \cdots, X_{iQ})'$. Also, recall from (1) that the regression coefficients are represented by a Q dimensional vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_Q)'$. The vector of the marginal maximum likelihood estimates (MML) of the regression effects is denoted as $\hat{\boldsymbol{\gamma}}$. The student ability values $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)$' can be inferred and scaled through item responses $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)'$ via item response theory (e.g., Lord & Novick, 1968), where $\boldsymbol{y}_i$ indicates the vector of item responses for student i. Then, using a standard breakdown into two components associated with sampling and measurement, the variance of the regression effects estimates can be expressed as (Mazzeo et al., 2006):

$$\begin{aligned} Var(\hat{\boldsymbol{\gamma}}) &\approx Var(\hat{\boldsymbol{\gamma}}|\boldsymbol{X}, \boldsymbol{Y}) \\ &= E[Var(\hat{\boldsymbol{\gamma}}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})] + Var[E(\hat{\gamma}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})] \end{aligned} \tag{2}$$

The first part of the posterior variance for the regression effects can be estimated by the variance of the regression estimates as if the examinees were selected from a simple random sample and the examinee ability values were observed. Mazzeo et al. (2006) attributed this portion of variation to the uncertainty in sampling. For the univariate case, denoting the residual covariance as σ^2, the first component is evaluated by $(\boldsymbol{X'X})^{-1}\sigma^2$ or $\left(\sum_{i=1}^{N} w_i \boldsymbol{x}_i \boldsymbol{x'}_i\right)^{-1} \sigma^2$ or $\boldsymbol{X'DX}^{-1}\sigma^2$, where w_i indicates the sampling weights and $\boldsymbol{D}$ is a diagonal matrix with the individual sampling weights on the diagonal.

As student abilities are not observed, the second portion for the posterior variance of the regression effects reflects the estimation of the variance due to the latency of $\boldsymbol{\theta}$. For the univariate case, the expectation of $E(\gamma|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}) = (\boldsymbol{X'DX})^{-1}\boldsymbol{X'D\theta}$. Thus the second portion of the standard error depends on the posterior variance of θ, that is,

$$Var(E(\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})) \quad = \quad (\boldsymbol{X'DX})^{-1}\boldsymbol{X'D}Var(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{DX}(\boldsymbol{X'DX})^{-1}. \tag{3}$$

To evaluate this quantity, the examinee's posterior distribution of $\boldsymbol{\theta}$ is assumed to be normal. Mazzeo et al. (2006) summarized this approximate method as follows:

> The more items each examinee receives the more valid this approximation becomes; the posterior distribution of θ has been shown to be normal as the number of items goes to infinity (Chang 1996, Chang & Stout, 1993). In fact, as the number of items becomes very large, this second term begins to vanish, because the (asymptotic) posterior variance of θ goes to zero.'

It should be noted that this relationship has been proven under the assumption of independently and identically distributed (i.i.d.) only.

In summary, the standard errors of the regression effects in the univariate case are approximated by

$$\begin{aligned} Var(\hat{\boldsymbol{\gamma}}) &\approx Var(\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Y}) \\ &= E[Var(\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})] + Var[E(\boldsymbol{\gamma}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta})] \\ &= (\boldsymbol{X'DX})^{-1}\sigma^2 + (\boldsymbol{X'DX})^{-1}\boldsymbol{X'D}Var(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{DX}(\boldsymbol{X'DX})^{-1}. \end{aligned} \tag{4}$$

For the p-variate case, the student ability vector $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{ip})$ for $i = 1, \cdots, N$ is assumed to have a common residual variance matrix $\boldsymbol{\Sigma}$. The variation due to sampling is $Cov(\hat{\boldsymbol{\gamma}}_s, \hat{\boldsymbol{\gamma}}_t)$,

which can be expressed as

$$Cov(\hat{\boldsymbol{\gamma}}_s, \hat{\boldsymbol{\gamma}}_t) \quad = \quad E(\hat{\boldsymbol{\gamma}}_s - \boldsymbol{\gamma}_s)(\hat{\boldsymbol{\gamma}}_t - \boldsymbol{\gamma}_t)', \tag{5}$$

for $s, t = 1, \cdots, p$. If the ability values were observed,

$$Cov(\hat{\boldsymbol{\gamma}}_s, \hat{\boldsymbol{\gamma}}_t) \quad = \quad \sigma_{st}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{6}$$

where σ_{st} is an element of the covariance matrix $\boldsymbol{\Sigma}$, which is

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_{pp} \end{pmatrix}.$$

For example, the variance of the estimates for the regression effects due to sampling for subscale t can be estimated as $var(\hat{\gamma}_t) = \sigma_{tt}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Hence, the standard error for the regression effects is the square root of the diagonal elements of the matrix $\sigma_{tt}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ for $t = 1, \cdots, p$.

1.2 Alternative Approaches

In NAEP and similar large-scale educational assessments, the abilities of examinees from the same sampling units (e.g., schools) are almost certainly positively correlated. Hence, a consequence of ignoring the complex sampling design is that the magnitude of the standard errors of $\hat{\boldsymbol{\gamma}}$ will be underestimated. It has been argued (e.g., Mazzeo et al., 2005) that the effect of ignoring the sampling design to calcuate the standard error of the regression effects is likely to be small relative to the size of the standard error of the target statistics (e.g., subgroup means and percentages above achievement levels). For target statistics, the sampling design is accounted for by a leave-out-group jackknife method and therefore the underestimation only pertains to the variation due to the latency of the construct of interest. This usually accounts for approximately 5% to 10% of the variability. However, it is important to examine how severe the underestimation is and to what extent a different approach would facilitate the reporting of regression effects and their standard errors in addition to NAEP target statistics.

Several alternatives for the estimation of the standard error of regression parameters have been suggested. A comprehensive discussion in relation to NAEP is provided by von Davier, Sinharay, Oranje, and Beaton (2007). Their work is concentrated on White's (1980) robust

method, Taylor series linearization (also referred to as Rao's delta method), and a method based on importance sampling using a Monte Carlo EM (MCEM) algorithm to estimate parameters. Taylor series linearization, following Binder's (1983) method, will be discussed below as it is seemingly the only one of the three approaches that takes the complex sample design into account. A computationally intensive approach based on the jackknife or alternative replication type methods (e.g., bootstrap, balanced repeated replications; see Kovar, 1985; Wolter, 1985) is conceivable as well. In that case, the estimation of the model parameters is carried out many times. This approach is left for future work.

1.3 Binder's (1983) Method

Binder's method has been advocated by Cohen and Jiang (2002) in order to use the regression effects directly for reporting while obtaining appropriate standard errors that take the complex sample into account. This is a deviation from the current methodology where draws from an imputation model form the basis of the report and statistical inference. Additionally, the method can be used to improve the variability under the imputation model and therefore the estimation of the variability due to measurement.

Specifically, Binder's method could provide a consistent variance estimate using a between-cluster estimator in combination with a Taylor series linearization approach. To approximate the variance of the marginal maximum likelihood (MML) $\hat{\boldsymbol{\gamma}}$ in the univariate case, Binder suggested using a first-order Taylor series expansion of a Q-dimensional function $\boldsymbol{W}(\hat{\boldsymbol{\gamma}})$ around the true unknown parameter $\boldsymbol{\gamma}$

$$\boldsymbol{W}(\hat{\boldsymbol{\gamma}}) \approx \boldsymbol{W}(\boldsymbol{\gamma}) + \boldsymbol{H}(\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}). \tag{7}$$

$\boldsymbol{W}(\boldsymbol{\gamma})$ in (7) is the partial derivative of the log-likelihood function with respect to $\boldsymbol{\gamma}$,

$$\boldsymbol{W}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} w_i \boldsymbol{g}_i(\boldsymbol{\gamma}), \tag{8}$$

with $\boldsymbol{g}_i(\boldsymbol{\gamma}) = (g_{i1}(\boldsymbol{\gamma}), \cdots, (g_{ij}(\boldsymbol{\gamma}), \cdots, g_{iQ}(\boldsymbol{\gamma}))$ and

$$g_{ij}(\boldsymbol{\gamma}) = \frac{\partial log L_i(\boldsymbol{\gamma}, \sigma^2)}{\partial \gamma_j}. \tag{9}$$

Because $\hat{\boldsymbol{\gamma}}$ is the vector of MML estimates, $\boldsymbol{W}(\hat{\boldsymbol{\gamma}}) = 0$. $\boldsymbol{H}(\boldsymbol{\gamma})$ in (7) is a Hessian matrix and defined as the partial derivative of $\boldsymbol{W}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ (i.e., $\boldsymbol{H}(\boldsymbol{\gamma}) = \frac{\partial \boldsymbol{W}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$). $\boldsymbol{H}(\boldsymbol{\gamma})$ is a

$Q \times Q$ matrix of partial derivatives with elements $\left[\frac{\partial W_k(\boldsymbol{\gamma})}{\partial \gamma_j}\right]$ or $\left[\sum_{i=1}^{N} \frac{\partial^2 log L_i(\boldsymbol{\gamma},\sigma^2)}{\partial \gamma_k \partial \gamma_j}\right]$ of $W_j(\boldsymbol{\gamma})$ with respect to γ_k for $j,k = 1, \cdots, Q$. Thus,

$$\begin{aligned} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} &\approx -\boldsymbol{H}(\boldsymbol{\gamma})^{-1}\boldsymbol{W}(\boldsymbol{\gamma}) \\ Var(\hat{\boldsymbol{\gamma}}) &= \boldsymbol{H}(\boldsymbol{\gamma})^{-1} Var(\boldsymbol{W}(\boldsymbol{\gamma}))\boldsymbol{H}(\boldsymbol{\gamma})'^{-1}. \end{aligned} \tag{10}$$

Binder (1983) suggested approximating the matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ by evaluating the Hessian matrix $\boldsymbol{H}(\hat{\boldsymbol{\gamma}})$ at MML estimates of $\hat{\boldsymbol{\gamma}}$. Clearly, the essential elements for the variance estimation are the gradient function $\boldsymbol{g}_i(\boldsymbol{\gamma})$, for $i = 1, \cdots, N$, the Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$, and the variance matrix $Var(\boldsymbol{W}(\boldsymbol{\gamma}))$. In the univariate case, the analytic expression for $\boldsymbol{W}(\boldsymbol{\gamma})$, the partial derivative of the log likelihood function with respect to $\boldsymbol{\gamma}$, can be written as

$$\boldsymbol{W}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} w_i \left(\frac{\boldsymbol{x}_i \tilde{\theta}_i - \boldsymbol{x}_i \boldsymbol{x}'_i \boldsymbol{\gamma}}{\sigma^2} \right). \tag{11}$$

Obviously, from (8), it follows that

$$\boldsymbol{g}_i(\boldsymbol{\gamma}) = \frac{\boldsymbol{x}_i \tilde{\theta}_i - \boldsymbol{x}_i \boldsymbol{x}'_i \boldsymbol{\gamma}}{\sigma^2}, \tag{12}$$

where $\tilde{\theta}_i$ indicates the posterior mean of the ability for student i. Therefore, the Hessian matrix can be obtained through the second order derivative of the log likelihood function with respect to $\boldsymbol{\gamma}$, or the first-order derivative of $\boldsymbol{W}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. That is, in the univariate case (see also the appendix)

$$\boldsymbol{H}(\boldsymbol{\gamma}) = \frac{\partial \boldsymbol{W}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = -\frac{1}{\sigma^4} \sum_{i=1}^{N} w_i \boldsymbol{x}_i \boldsymbol{x}'_i (\tilde{\sigma}_i^2 + \sigma^2). \tag{13}$$

Cohen and Jiang (2002) claimed that the Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ can be approximated by

$$\boldsymbol{H}(\boldsymbol{\gamma}) \approx \sum_{i=1}^{N} w_i \boldsymbol{g}_i(\boldsymbol{\gamma}) \boldsymbol{g}'_i(\boldsymbol{\gamma}), \tag{14}$$

and the variance of $\boldsymbol{W}(\boldsymbol{\gamma})$ [denoted as $\boldsymbol{\Omega}(\boldsymbol{\gamma})$] is the variance of $\boldsymbol{W}(\boldsymbol{\gamma})$ across observations. In addition, Cohen and Jiang used the stratified, between-primary sampling unit (PSU) weighted estimator to obtain the estimate of $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ [denoted as $\widehat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\gamma}})$]. That is,

$$\widehat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\gamma}}) = \sum_{h=1}^{H} \left(\frac{n_h}{n_h - 1} \right) \sum_{i=1}^{n_h} (\boldsymbol{g}_{hi} - \bar{\boldsymbol{g}}_h)(\boldsymbol{g}_{hi} - \bar{\boldsymbol{g}}_h)', \tag{15}$$

where $\boldsymbol{g}_{hi} = \sum_{k=1}^{m_{hi}} w_{hik} \frac{\partial log(L_k)}{\partial \boldsymbol{\gamma}}$, and $\bar{\boldsymbol{g}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \boldsymbol{g}_{hi}$, in which h indexes the stratum, i indexes the primary sampling unit, k indexes individuals, and m_{hi} is the number of students in hth strata and ith PSU unit.

From the discussion of Binder's method (1983) above, it can be seen that the computation for standard errors depends on the evaluation of two matrices: the Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ defined in (13) and the variance matrix across PSUs $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ defined in (15). In the following section two alternatives for implementation of Binder's general approach will be discussed, where in the second method an approximation to the Hessian matrix is used. Also, a method based on Fisher's information matrix will be used as benchmark representing a simple random sample approach in addition to the current NAEP methodology described in section 1.1.

Method 1. For Method, 1 the Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ defined in (13) and $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ defined in (15) are used. In practice, the variances of the regression effects $Var(\boldsymbol{\gamma})$ are estimated by substituting $\boldsymbol{H}(\boldsymbol{\gamma})$ in (13) and $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ in (15) into (10).

Method 2. For Method 2, the Hessian matrix, suggested by Cohen and Jiang (2002) and defined in (14), and $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ are used in (15) to estimate the variances of the regression effects. The difference between Methods 1 and 2 is the computation of the Hessian matrix. This matrix is easier to compute than the matrix in (13). However, additional estimation errors are introduced by this approximation.

Method 3. The matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ defined in (13) is related to the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\gamma})$ (e.g., Lord & Novick, 1968, p. 418), which follows

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\gamma}) &= E\left(\frac{\partial}{\partial \boldsymbol{\gamma}} logL(\boldsymbol{\gamma}, \boldsymbol{\Sigma})\right)^2 \\ &= -E\left(\frac{\partial^2}{\partial \boldsymbol{\gamma}^2} logL(\boldsymbol{\gamma}, \boldsymbol{\Sigma})\right) \\ &= -E\left[\boldsymbol{H}(\boldsymbol{\gamma})\right], \end{aligned} \tag{16}$$

where $logL(\boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is the log likelihood function. Thus, the variance of regression effects estimates can be evaluated by

$$Var(\boldsymbol{\gamma}) = \boldsymbol{I}(\boldsymbol{\gamma})^{-1} \approx \left[-\boldsymbol{H}(\boldsymbol{\gamma})\right]^{-1}. \tag{17}$$

Method 3 directly uses the information matrix to compute estimates for $Var(\boldsymbol{\gamma})$. One advantage of Method 3 is that this method is relatively simple to implement and compute. Moreover, the

estimates of the standard errors are also MML estimates and this method provides a baseline for comparison with the other methods in terms of assessing the underestimation of the variance. However, it should be noted that Method 3 does not account for the variation due to cluster sampling.

1.4 Summary

In summary, two potential methods that claim to take the complex sample characteristics into account are defined to estimate the standard error of regression effects of the NAEP latent regression model. There are two important applications of these methods. First, a more accurate notion of the variability of regression effects could possibly be obtained, possibly for use with reporting regression effects. Second, the multiple imputations model can be improved by acknowledging the complex sample at all stages of the imputation process.

In the following section, Binder's method will be extended to multiple dimensions. Following that, an empirical study will be presented to compare these methods as well as their applications. The current NAEP methodology is used for comparison. In addition, a method based on Fisher's information matrix is added, which is a computational simplification of the current NAEP methodology.

2 Extending Binder's (1983) Method to Multiple Dimensions

Mazzeo et al. (2006) suggested that an improvement to the current NAEP methodolgy would be to generalize Binder's (1983) method to multivariate models. Let $\boldsymbol{\gamma}_1 = (\gamma_{11}, \gamma_{12}, ..., \gamma_{1Q})'$ be the vector of regression effects for the first subscale. Collect all these regression effects for each subscale in a column $\boldsymbol{\gamma} = (\boldsymbol{\gamma}'_1, \cdots, \boldsymbol{\gamma}'_p)'$, then $\boldsymbol{\gamma}$ is a pQ-dimensional vector containing the regression effects for all p subscales. Define $\boldsymbol{g}_{ij}(\gamma_t) = \frac{\partial L_i(\boldsymbol{\gamma},\Sigma)}{\partial \gamma_{tj}}$, $\boldsymbol{W}_j(\boldsymbol{\gamma}_t) = \sum_{i=1}^N w_i \boldsymbol{g}_{ij}(\gamma_{tj})$ for $j = 1, 2, \cdots, Q$ and $t = 1, \cdots, p$. The analytic expression for the gradient vector can be written as

$$\boldsymbol{g}_i(\boldsymbol{\gamma}) = \frac{\partial L_i(\boldsymbol{\gamma}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\gamma}} = \boldsymbol{x}_i \otimes \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\gamma}' \boldsymbol{x}_i). \tag{18}$$

Subsequently, collect the gradients $\boldsymbol{g}_i(\boldsymbol{\gamma}_t)$ in a column for p subscales, that is, $\boldsymbol{g}_i(\boldsymbol{\gamma}) = (\boldsymbol{g}'_i(\boldsymbol{\gamma}_1), \cdots, \boldsymbol{g}'_i(\boldsymbol{\gamma}_p))'$, as is defined in (18). From the previous section, note that $\boldsymbol{W}(\boldsymbol{\gamma})$ is a vector of the first-order partial derivatives of the likelihood function for p subscales with respect to $\boldsymbol{\gamma}$, and $\boldsymbol{W}(\boldsymbol{\gamma}) = \sum_{i=1}^N w_i \boldsymbol{g}_i(\boldsymbol{\gamma})$. Hence, the difference from the univariate case is that the

regression parameter vector $\boldsymbol{\gamma}$ contains all p subscale regression parameters with dimension pQ. The first-order Taylor expansion of the pQ-dimension function of $W(\hat{\boldsymbol{\gamma}})$ around the true parameter vector $\boldsymbol{\gamma}$ is given by

$$\boldsymbol{W}(\hat{\boldsymbol{\gamma}}) \approx \boldsymbol{W}(\boldsymbol{\gamma}) + \boldsymbol{H}(\boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}). \tag{19}$$

Because $\hat{\boldsymbol{\gamma}}$ maximizes the likelihood function, it also solves the likelihood equation system $\boldsymbol{W}(\hat{\boldsymbol{\gamma}}) = \mathbf{0}$. From that it follows

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \approx -\boldsymbol{H}(\boldsymbol{\gamma})^{-1}\boldsymbol{W}(\boldsymbol{\gamma}) \tag{20}$$

$$Var(\hat{\boldsymbol{\gamma}}) = \boldsymbol{H}(\boldsymbol{\gamma})^{-1}Var(W(\boldsymbol{\gamma}))[\boldsymbol{H}(\boldsymbol{\gamma})^{-1}]'. \tag{21}$$

$\boldsymbol{H}(\boldsymbol{\gamma})$ is a Hessian matrix with dimension $pQ \times pQ$ of partial derivatives, that is, $\boldsymbol{H}(\boldsymbol{\gamma}) = ((h_{kj}))$, for $h_{kj} = [\sum_{i=1}^{N} \frac{w_i \partial^2 log L_i(\boldsymbol{\gamma}, \boldsymbol{\Sigma})}{\partial \gamma_k \partial \gamma_j}] = \frac{\partial W_k(\boldsymbol{\gamma})}{\partial \gamma_j}$, for $k, j = 1, 2, \cdots, pQ$. Simply put, $\boldsymbol{H}(\boldsymbol{\gamma}) = \frac{\partial \boldsymbol{W}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ can be expressed as a block matrix, where each diagonal block matrix is occupied by the Hessian matrix for each subscale. Each block component matrix in $\boldsymbol{H}(\boldsymbol{\gamma})$ has dimension $Q \times Q$. The analytic expression for computing the diagonal block matrix in $\boldsymbol{H}(\boldsymbol{\gamma})$ is given by

$$\boldsymbol{H}(\boldsymbol{\gamma}) = -\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i' \otimes \left(\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\Sigma}}_i \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1} \right). \tag{22}$$

From (22), it can be seen that the Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ is a block diagonal matrix and can be expressed as

$$\boldsymbol{H}(\boldsymbol{\gamma}) = \begin{pmatrix} \boldsymbol{H}_{11}(\boldsymbol{\gamma}) & \boldsymbol{H}_{12}(\boldsymbol{\gamma}) & \boldsymbol{H}_{13}(\boldsymbol{\gamma}) & \cdots & \boldsymbol{H}_{1p}(\boldsymbol{\gamma}) \\ \boldsymbol{H}_{21}(\boldsymbol{\gamma}) & \boldsymbol{H}_{22}(\boldsymbol{\gamma}) & \boldsymbol{H}_{23}(\boldsymbol{\gamma}) & \cdots & \boldsymbol{H}_{2p}(\boldsymbol{\gamma}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{H}_{p1}(\boldsymbol{\gamma}) & \boldsymbol{H}_{p2}(\boldsymbol{\gamma}) & \boldsymbol{H}_{p3}(\boldsymbol{\gamma}) & \cdots & \boldsymbol{H}_{pp}(\boldsymbol{\gamma}) \end{pmatrix}.$$

The Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ can be approximated by $\sum_{i=1}^{N} w_i \boldsymbol{g}_i(\boldsymbol{\gamma}) \boldsymbol{g}_i'(\boldsymbol{\gamma})$, similar to the univariate case.

The variance matrix in (10), $Var[\boldsymbol{W}(\boldsymbol{\gamma})]$ or $\boldsymbol{\Omega}(\boldsymbol{\gamma})$, is the variance of $\boldsymbol{W}(\boldsymbol{\gamma})$ across observations. By extending (15) to a multivariate setting, the estimates of $\boldsymbol{\Omega}(\boldsymbol{\gamma})$ are obtained with dimension $pQ \times pQ$.

3 Empirical Results

An empirical study was conducted to compare the methods described in the previous section. The study was limited to a univariate problem. Data was from the 2004 NAEP long-term trend mathematics assessment, which contained almost 7,600 students at the age of 17 sampled from the population of U.S. students. The cognitive part of the assessment had 162 items combining multiple-choice and constructed-response answering formats. All constructed-responses were dichotomously scored. Each student responded to approximately half of the item pool. A large set of student characteristic variables was reduced by principal component analysis to 156 components, accounting for 90% of the total variance. Both this large saturated or nearly saturated model as well as a smaller model including only student characteristics such as gender and race/ethnicity were evaluated in this study. It should be noted that NAEP's current approach is to use the large principal-components-based model.

In addition to the proposed methods, the current NAEP methodology were added to the comparison. Also, NAEP's operational imputation methodology (Mislevy, 1991) of proficiency was applied to all methods to make a comparison between estimates of student group effects based on the regression coefficients and estimates of these effects based on imputations of ability. In addition, it can be assessed what the impact of accounting for the complex sample with respect to measurement error is on NAEP's target statistics.

3.1 A Large Model With Principal Components

First, the estimates of the standard errors of NAEP's current approach are compared with the estimates from the proposed methods. In theory, Methods 1 and 2 would yield a larger estimate of the standard errors than those from the the current NAEP approach, since these two methods take the variation across clusters into account. The third method would yield similar estimates of standard errors to those from the current NAEP approach, since both methods ignore the complex sampling designs during the estimation process. This method is included predominantly for its appeal with respect to computational simplicity of the Hessian matrix in (13) and the asymptotic properties of MML estimates.

Table 1 lists the estimates of the standard errors for the regression parameters $\boldsymbol{\gamma}$ from all methods along with results from the current NAEP approach (denoted as NAEP SE). The first column is the number of the conditioning variables or principal components from the latent

regression model (the principal component factor scores, or PCFS). The second column presents the estimates of the regression parameters for these 156 principal components. The third column shows standard error estimates via the current NAEP procedure. The fourth through sixth columns are standard error estimates using the three alternative methods. Note that only the first 15 and the last 5 regression parameters are listed for brevity.

Table 1
Standard Error Estimates for Regression Coefficients γ

PCFS	γ	NAEP S.E.	Method 1	Method 2	Method 3
Intercept	.0046	.009	.0125	.0155	.0083
2	– .0374	.001	.0011	.0014	.0009
3	.0049	.0013	.0016	.002	.0012
4	.0142	.0016	.0018	.0022	.0015
5	– .0171	.0018	.0021	.0025	.0017
6	– .0461	.0019	.0027	.0032	.0017
7	– .0017	.0019	.0015	.002	.0018
8	– .0235	.0021	.0026	.0031	.0019
9	– .001	.0021	.0024	.003	.002
10	.0013	.0023	.0029	.0034	.0021
11	.0223	.0023	.0022	.0026	.0021
12	.009	.0023	.002	.0026	.0022
13	– .0091	.0025	.0031	.0033	.0022
14	.0149	.0025	.0029	.0034	.0024
15	– .0071	.0027	.0024	.0029	.0025
⋮	⋮	⋮	⋮	⋮	⋮
153	.0002	.0101	.0097	.0112	.0093
154	.0098	.0101	.0098	.011	.0093
155	– .0184	.0101	.0111	.0136	.0094
156	– .028	.0101	.0095	.0108	.0094
157	– .0155	.0102	.0114	.0122	.0094

Note. S.E. = standard error, PCFS = principal component factor scores.

From Table 1, it can be seen that the standard error estimates from Methods 1 and 2 are very close to each other and also close to the results from the current NAEP operational approach. Specifically, most of the estimates from Method 1 are slightly greater than those from the current NAEP estimates. However, there are a few estimates that are smaller. Results from Method 2 are uniformly greater than the current NAEP approach. Results from Method 3 are generally smaller than the current NAEP approach, which is also expected since this method assumes a

simple random sample and does not account for the variation due to the latency, as is done under the current NAEP approach.

The difference in magnitude between NAEP's approach and Methods 1 and 2 is certainly surprising as the sample is generally believed to have a design effect between 2 and 3, based on studies employing resampling methods (see Allen et al., 2001). A design effect is the ratio between a complex sample variance estimate and a variance estimate using a simple random sample estimator. A possible explanation can be that the saturated model contains a large number of school variables and, therefore, in some sense a fixed effects hierarchical model is estimated. In other words, the hierarchical structure with respect to the measurement model is largely accounted for by the model. This will be further discussed below.

3.2 Direct Estimates of Subpopulation Characteristics With a Large Model

An important question is how these different estimators affect the standard error estimates of student group abilities. The MML estimate of the mean proficiency vector for group G can be obtained from the estimated regression parameters (Mazzeo et al., 2006) as:

$$\hat{\boldsymbol{\mu}}_G = \hat{\boldsymbol{\Gamma}}' \bar{\boldsymbol{x}}'_G, \tag{23}$$

where $\bar{\boldsymbol{x}}_G$ is the sample mean vector of the background variables from examinees in group G,

$$\bar{\boldsymbol{x}}'_G = \frac{\sum_{i \in G} w_i \boldsymbol{x}_i}{\sum_{i \in G} w_i}. \tag{24}$$

The variance of the group mean estimate is computed by

$$Var(\hat{\boldsymbol{\mu}}_G) = \bar{\boldsymbol{x}}_G \, Var(\hat{\boldsymbol{\Gamma}}) \bar{\boldsymbol{x}}'_G. \tag{25}$$

The empirical Bayesian estimate of the same quantity is given by

$$\tilde{\boldsymbol{\mu}}_G = \sum_{i \in G} \frac{w_i \tilde{\mu}_i}{\sum_{i \in G} w_i}, \tag{26}$$

where $\tilde{\mu}_i$ is the mean for the posterior distribution for examinee i evaluated at $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$. The difference between (23) and (26) is that the first computation is the mean implied by the regression model and the second is the mean with respect to the complete model. Hence, misspecification of the regression model will yield some differences. These estimates are generally described as direct

estimates in contrast to imputation based estimates. Hence, direct estimation of subgroup means requires the estimates for regression effects $\hat{\gamma}$ and the means of the student group variables. Tables 2 to 5 show the student group mean estimates for all methods, both untransformed (the third column) and transformed (the fifth column) to the current NAEP reporting scales. Following (25), the standard errors for the direct estimates of subgroup means depend on the covariance matrix of regression effects estimates. The effect of the four different approaches on the standard error estimates for these subgroup mean estimates will be described below.

Table 2

Direct Estimates for the Subgroup Means and Standard Errors—NAEP Before Application of Jackknife

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0572	0.0127	306.8993	0.3932	306.1287	307.6700	1.5413
Female	– 0.0473	0.0127	303.6739	0.3904	302.9087	304.4392	1.5305
White	0.1936	0.0108	311.1059	0.3334	310.4524	311.7594	1.3069
Black	– 0.6599	0.0251	284.7762	0.7757	283.2558	286.2966	3.0407
Hispanic	– 0.4453	0.0239	291.3955	0.7359	289.9533	292.8378	2.8845
A./P.I.A.	0.3296	0.0450	315.3021	1.3885	312.5807	318.0235	5.4428

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 2 shows the direct estimates for gender and race student group means and the corresponding standard errors following the latent regression coefficients standard errors before application of the jackknife estimator. NAEP does not publish these estimates because the complex sampling design is ignored. However, it is informative to assess the impact of using complex sample variance estimators. Essentially, the same set of regression effects is used but with different estimates for the variance of these regression effects. Surprisingly, the standard errors are quite similar. Between Methods 1 and 2, the approximation used in the second method appears to inflate that variance between 10% and 20% for this particular sample. Also, Method 3 seems to underestimate the variance even when compared to the standard errors following the NAEP

Table 3

Direct Estimates for the Subgroup Means and Standard Errors—Method 1

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0572	0.0133	306.8993	0.4096	306.0966	307.7020	1.6055
Female	– 0.0473	0.0154	303.6739	0.4751	302.7428	304.6051	1.8623
White	0.1936	0.0141	311.1059	0.4339	310.2554	311.9564	1.7010
Black	– 0.6599	0.0320	284.7762	0.9869	282.8418	286.7106	3.8688
Hispanic	– 0.4453	0.0241	291.3955	0.7425	289.9402	292.8509	2.9107
A./P.I.A.	0.3296	0.0452	315.3021	1.3950	312.5679	318.0364	5.4684

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 4

Direct Estimates for the Subgroup Means and Standard Errors—Method 2

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0572	0.0143	306.8993	0.4415	306.0340	307.7647	1.7307
Female	– 0.0473	0.0211	303.6739	0.6501	302.3997	304.9481	2.5484
White	0.1936	0.0179	311.1059	0.5530	310.0221	312.1897	2.1677
Black	– 0.6599	0.0388	284.7762	1.1977	282.4288	287.1236	4.6948
Hispanic	– 0.4453	0.0296	291.3955	0.9146	289.6029	293.1882	3.5853
A./P.I.A.	0.3296	0.0505	315.3021	1.5592	312.2460	318.3582	6.1122

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

formulas before the complex sample is accounted for. This could be due to the fact that Method 3 does not take the variability due to the latency of the construct into account as NAEP does.

In Table 6, the NAEP *published* results are presented, based on the multiple imputation

Table 5

Direct Estimates for the Subgroup Means and Standard Errors—Method 3

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0572	0.0117	306.8993	0.3615	306.1908	307.6078	1.4170
Female	– 0.0473	0.0117	303.6739	0.3602	302.9679	304.3799	1.4120
White	0.1936	0.0100	311.1059	0.3081	310.5020	311.7098	1.2078
Black	– 0.6599	0.0230	284.7762	0.7103	283.3839	286.1684	2.7845
Hispanic	– 0.4453	0.0218	291.3955	0.6720	290.0784	292.7126	2.6342
A./P.I.A.	0.3296	0.0410	315.3021	1.2639	312.8250	317.7793	4.9544

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 6

NAEP Plausible Value-Based Estimates for the Subgroup Means and Standard Errors

Subgroup	Means trans	S.E trans.	L (95%)	U (95%)	Width
Male	307.04	0.894	305.288	308.792	3.504
Female	303.58	0.795	302.022	305.138	3.116
White	311.15	0.685	309.807	312.493	2.685
Black	284.23	1.397	281.492	286.968	5.476
Hispanic	291.84	1.208	289.472	294.208	4.735
A./P.I.A.	315.15	3.000	309.270	321.030	11.760

Note. Trans. = transformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

methodology and with application of the jackknife variance estimator. In comparison, Methods 1 and 2 underestimate the variance for most groups except Hispanic students. This is surprising

as all three approaches in theory are expected to yield similar results. This is a function of the modest increase of standard errors of the regression effects in the (nearly) saturated model and will be further explored in the discussion section of this report. Yet, it is clear that direct estimates based on a (nearly) saturated model are inappropriate. There are also some small differences between the mean estimates of some racial groups comparing the published results and the direct estimates. This is expected as the multiple imputations methodology has a random element.

3.3 Multiple Imputation Based Estimates of Subpopulation Characteristics With a Large Model

An interesting application of Binder's method to NAEP could be to improve the standard error estimates of the regression coefficients in order to more accurately depict the width of the distribution used to draw multiple imputations. In Tables 7 to 10 plausible value-based results are shown including jackknife based standard errors. It can be argued that Methods 1 and 2 already take the complex sample into account in the regression variance estimates and, therefore, that the multiple imputations should depict an appropriately wide distribution. However, the standard errors of the regression effects only inform the imputation model with respect to the measurement variance. If during the assessment each student is exposed to a reasonably large number of items, then the shape of the item likelihood can be expected to be peaked and, subsequently, it will dominate the model. The influence of the measurement variation will be rather limited as far as the posterior moments are concerned. These moments are in turn used to draw multiple imputations. Specifically, the regression parameter estimates $\hat{\gamma}$ and the associated variance matrix are used as parameters of a multivariate normal distribution and a set of regression parameters is drawn from this distribution. Subsequently, posterior moments are computed and a set of multiple imputations is generated from normal distributions with parameters equal to those moments. These two steps are repeated for the number of desired imputations. In the current example, students answered approximately half of the 162 questions, which can be considered a substantial number of items.

Hence, results are reported on how standard errors estimates would affect NAEP student group mean estimates and standard errors following the multiple imputations. Please note that the imputation method is approximated for Methods 1 and 2 using univariate distributions instead of multivariate normal distributions. One concern of the Binder method under this large

Table 7

Plausible Value-Based Estimates for the Subgroup Means and Standard Errors—NAEP Methodology

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0728	0.0313077	308.039	0.965843	306.146	309.932	3.7861
Female	– 0.0357	0.0269593	304.514	0.831693	302.884	306.144	3.26024
White	0.2064	0.0255627	312.293	0.788608	310.748	313.839	3.09134
Black	– 0.6553	0.077085	284.611	2.37807	279.95	289.272	9.32204
Hispanic	– 0.4195	0.058173	292.259	1.79464	288.741	295.776	7.03497
A./P.I.A.	0.3533	0.114849	317.465	3.5431	310.52	324.409	13.889

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 8

Plausible Value-Based Estimates for the Subgroup Means and Standard Errors—Method 1

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0728	0.0312491	308.141	0.964036	306.252	310.031	3.77902
Female	– 0.0357	0.0279076	304.675	0.860948	302.988	306.363	3.37492
White	0.2064	0.0247893	312.413	0.764749	310.914	313.912	2.99782
Black	– 0.6553	0.0780948	284.549	2.40922	279.827	289.271	9.44416
Hispanic	– 0.4195	0.060444	292.468	1.8647	288.813	296.123	7.30961
A./P.I.A.	0.3533	0.125613	317.952	3.87517	310.357	325.548	15.1907

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 9

Plausible Value-Based Estimates for the Subgroup Means and Standard Errors—Method 2

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0728	0.0307026	308.17	0.947174	306.314	310.027	3.71292
Female	– 0.0357	0.0261356	304.518	0.806282	302.937	306.098	3.16063
White	0.2064	0.0247	312.308	0.761994	310.815	313.802	2.98702
Black	– 0.6553	0.0796551	284.644	2.45736	279.827	289.46	9.63285
Hispanic	– 0.4195	0.0618095	292.42	1.90682	288.683	296.157	7.47474
A./P.I.A.	0.3533	0.112409	317.85	3.46783	311.053	324.647	13.5939

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 10

Plausible Value-Based Estimates for the Subgroup Means and Standard Errors—Method 3

Group	Means	S.E. untrans.	Scale	S.E. trans.	L (95%)	U (95%)	Width
Male	0.0728	0.0321202	308.099	0.990909	306.157	310.041	3.88436
Female	– 0.0357	0.0276155	304.58	0.851937	302.91	306.25	3.33959
White	0.2604	0.0257705	312.357	0.79502	310.798	313.915	3.11648
Black	– 0.6553	0.077136	284.516	2.37965	279.852	289.181	9.32821
Hispanic	– 0.4195	0.0641211	292.296	1.97814	288.419	296.173	7.75429
A./P.I.A.	0.3533	0.115223	317.821	3.55463	310.854	324.788	13.9341

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

model and the relatively small sample size is whether the associated covariance matrices are positive definite. However, since the off-diagonal elements carry near-zero values, the impact of this approximation is expected to be minimal. The purpose of this comparison is to show how the various estimates of the variance of regression effects affect NAEP student group means and standard errors and how biased these estimates may be if the complex sampling design is ignored.

Table 7 lists the results using the current NAEP estimates of variance matrix for regression parameters. The results should in theory be consistent with those in Table 6, except that the imputation is slightly different. Hence, slight differences are observed in particular for the standard error estimates of Black students. Since the imputation is similar for all four methods, the results from Table 7 will provide the baseline for comparison. Method 1 in Table 8 appears quite similar to the baseline, where smaller groups are somewhat overestimated. This effect is similar if not larger in Method 2 in Table 9. The exception is Asian/Pacific Island American students, which comprise a nonreportable group for NAEP following statistical standards for minimum cell sample sizes. Method 3 generally provides somewhat larger standard errors except for the last group. Note that the effect of using alternative variance estimators for regression effects only affects a small part of the variance.

3.4 Subpopulation Estimates Based on Small Models

As far as the authors are aware, comparisons of direct estimates and multiple imputations based on large regression models have not been published. However, several comparisons have been made between direct estimates from a small model and multiple imputation based results (e.g., Cohen & Jiang, 2002; von Davier, 2003). In addition, some concerns about the large model have been voiced (e.g., Aitkin, 2003). Therefore, in this section three small models have been estimated. The models are a gender-only, a race-only, and a combined gender and race model. Dummy codes have been used to distinguish the variable categories and every model has an intercept. For example, the gender model has an intercept and a dummy variable that is equal to 1 for the female category and 0 otherwise.

The MML estimates for the regression effects parameters for these three models are given in the second column in Tables 11, 12, and 13, respectively. We expect that the residual variance σ^2 will be larger than that in the large operational model. The residual variance estimates corresponding to the these three small models are .978, .8674, and .8653, which are all larger than

Table 11

Standard Error Estimates for Regression Coefficients γ on a Small Model With Gender-Only

PCFS	$\hat{\gamma}$	NAEP/SRS S.E.	Method 1	Method 2	Method 3
Intercept	0.0750	0.0166	0.0358	0.0368	0.0157
Female	– 0.1104	0.0234	0.0224	0.0262	0.0222

Note. $\sigma^2 = .978$. SRS S.E. = simple random sample standard error, PCFS = principal component factor scores.

Table 12

Standard Error Estimates for Regression Coefficients γ on a Small Model With Race/Ethnicity-Only

PCFS	$\hat{\gamma}$	NAEP/SRS S.E.	Method 1	Method 2	Method 3
Intercept	0.2080	0.0131	0.0318	0.0352	0.0124
Black	– 0.8680	0.0343	0.0514	0.0615	0.0322
Hispanic	– 0.6300	0.0328	0.0454	0.0503	0.0308
A./P.I.A.	0.0690	0.0593	0.0801	0.0812	0.0553

Note. $\sigma^2 = .8674$. SRS S.E. = simple random sample standard error, PCFS = principal component factor scores, A./P.I.A. = Asian/Pacific Island American.

the residual variance estimates for NAEP operational models of .5673.

The standard error of the regression effects estimates are given in column 3 through column 6 in Tables 11, 12, and 13. Column 3 gives the standard error for the regression parameter estimates if (4) was used and column 4 through 6 provide the standard errors for the regression parameters estimates following Methods 1 through 3. The results follow a similar pattern as for the large model. That is, Method 3 yields the smallest estimates as this method does not account for the cluster sample design and measurement errors. Methods 1 and 2 are very similar to each other and also provide larger standard errors estimates than those based on (4), which is to be expected as this equation does not take the complex sample into account.

Table 13

Standard Error Estimates for Regression Coefficients $\hat{\gamma}$ on a Small Model With Gender + Race/Ethnicity-Only

PCFS	$\hat{\gamma}$	NAEP/SRS S.E.	Method 1	Method 2	Method 3
Intercept	0.2514	0.0170	0.0341	0.0354	0.0161
Female	– 0.0882	0.0221	0.0195	0.0221	0.0208
Black	– 0.8640	0.0343	0.0514	0.0614	0.0322
Hispanic	– 0.6270	0.0328	0.0454	0.0505	0.0308
A./P.I.A.	0.0729	0.0593	0.0795	0.0807	0.0552

Note. $\sigma^2 = .8674$. SRS S.E. = simple random sample standard error, PCFS = principal component factor scores, A./P.I.A. = Asian/Pacific Island American.

Table 14

Direct Estimates for the Subgroup Means and Standard Errors on Gender-Only Model

Method	Group	Means	S.E. untrans.	Scale	S.E trans.	L (95%)	U (95%)	Width
1	Male	0.0750	0.0358	307.4478	1.1052	305.2816	309.6139	4.3323
	Female	– 0.0354	0.0382	304.0419	1.1791	301.7308	306.3530	4.6221
2	Male	0.0750	0.0368	307.4478	1.1340	305.2251	309.6704	4.4452
	Female	– 0.0354	0.0457	304.0419	1.4085	301.2812	306.8027	5.5215
3	Male	0.0750	0.0157	307.4478	0.4855	306.4962	308.3993	1.9030
	Female	– 0.0354	0.0157	304.0419	0.4832	303.0949	304.9890	1.8941
	Male	0.0750	0.1803	307.4478	5.5616	296.5470	318.3485	21.8016

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

Table 15

Direct Estimates for the Subgroup Means

and Standard Errors on Race/Ethnicity Only-Model

Method	Group	Means	S.E. untrans.	Scale	S.E trans.	L (95%)	U (95%)	Width
1	White	0.2080	0.0318	311.5508	0.9806	309.6287	313.4729	3.8441
	Black	– 0.6600	0.0455	284.7730	1.4036	282.0219	287.5241	5.5022
	Hispanic	– 0.4220	0.0380	292.1153	1.1736	289.8151	294.4155	4.6003
	A./P.I.A.	0.2770	0.0711	13.6795	2.1937	309.3799	317.9790	8.5991
2	White	0.2080	0.0352	311.5508	1.0874	309.4195	313.6821	4.2626
	Black	– 0.6600	0.0559	284.7730	1.7256	281.3908	288.1552	6.7644
	Hispanic	– 0.4220	0.0422	292.1153	1.3014	289.5645	294.6661	5.1016
	A./P.I.A.	0.2770	0.0705	313.6795	2.1738	309.4189	317.9400	8.5212
3	White	0.2080	0.0124	311.5508	0.3837	310.7987	312.3029	1.5042
	Black	– 0.6600	0.0297	284.7730	0.9170	282.9757	286.5703	3.5946
	Hispanic	– 0.4220	0.0282	292.1153	0.8695	290.4112	293.8194	3.4083
	A./P.I.A.	0.2770	0.0539	313.6795	1.6622	310.4214	316.9375	6.5160

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

The student group estimates based on these three small models are given in Tables 14, 15, and 16, respectively. First note that, as is the case for the large model, the direct estimates of the transformed student group means (column 3 in Tables 14, 15, and 16) are close to NAEP's operational results (Table 6). Second, the transformed standard errors estimates (column 4) for student group means for the three small models under Methods 1 and 2 are somewhat larger than those from the NAEP jackknife standard errors in Table 6, column 3. One possible explanation is that the residual variance is relatively large for the three small models. Similar to the larger model, the approximation of Method 2 inflates the standard error estimates between 10% and 20%.

Table 16

Direct Estimates for the Subgroup Means

and Standard Errors on Gender + Race/Ethnicity Only-Model

Method	Group	Means	S.E. untrans.	Scale	S.E trans.	L (95%)	U (95%)	Width
1	Male	0.0741	0.0282	307.4212	0.8692	305.7175	309.1248	3.4073
	Female	– 0.0371	0.0261	303.9907	0.8054	302.4121	305.5692	3.1571
	White	0.2082	0.0319	311.5566	0.9850	309.6261	313.4871	3.8610
	Black	– 0.6606	0.0456	284.7553	1.4078	281.9959	287.5146	5.5187
	Hispanic	– 0.4215	0.0378	292.1311	1.1656	289.8465	294.4157	4.5692
	A./P.I.A.	0.2769	0.0705	313.6765	2.1762	309.4111	317.9418	8.5307
2	Male	0.0741	0.0289	307.4212	0.8920	305.6728	309.1696	3.4968
	Female	– 0.0371	0.0321	303.9907	0.9895	302.0513	305.9300	3.8788
	White	0.2082	0.0356	311.5566	1.0970	309.4066	313.7066	4.3000
	Black	– 0.6606	0.0561	284.7553	1.7309	281.3627	288.1479	6.7852
	Hispanic	– 0.4215	0.0421	292.1311	1.2981	289.5868	294.6754	5.0887
	A./P.I.A.	0.2769	0.0699	313.6765	2.1577	309.4474	317.9055	8.4581
3	Male	0.0741	0.0148	307.4212	0.4554	306.5286	308.3137	1.7851
	Female	– 0.0371	0.0147	303.9907	0.4533	303.1022	304.8791	1.7769
	White	0.2082	0.0124	311.5566	0.3833	310.8054	312.3078	1.5023
	Black	– 0.6606	0.0297	284.7553	0.9158	282.9603	286.5503	3.5900
	Hispanic	– 0.4215	0.0281	292.1311	0.8684	290.4291	293.8331	3.4039
	A./P.I.A.	0.2769	0.0538	313.6765	1.6602	310.4225	316.9304	6.5079

Note. S.E. untrans. = standard error untransformed to the NAEP reporting scales, S.E. trans. = standard error transformed to the NAEP reporting scales, L = lower, U = upper, A./P.I.A. = Asian/Pacific Island American.

4 Discussion and Conclusions

In this study, several alternative methods have been explored for the computation of standard error of regression effects in NAEP's latent regression model. Currently, a simple random sample

assumption is made at the estimation stage and a post hoc complex sample estimator is used to appropriately account for the design. Hence, the standard error of intermediate statistics derived at the modeling stage is possibly underestimated and as such, while using model parameters directly, are deemed inappropriate from which to draw conclusions about the population of interest. For example, the regression effect for a variable indicating membership to the class of females cannot be used directly without further complex sample variance estimation procedures.

Two methods were compared mostly based on Binder's general methodology. Both large and small models were compared using both imputation and regression coefficient-based aggregation. The results indicate that Binder's method does provide larger standard error estimates, but in a very limited way for a large model. The impact of using Binder's method on final imputation based results is minimal. However, if a small model is estimated, Binder's method does increase standard error estimates quite substantially. However, it is not advisable to use the approximation to the covariance matrix of Method 2, because it inappropriately inflates results by 10% to 20%. This inflation might even be stronger if less items per student are administered.

The differences between the large and the small models are quite noticeable and require further discussion. One of the most important differences between the models is the residual variance, which is high in the small model and moderate in the larger model. Whether the increased variation is due to the complex sample or to the model is not entirely clear. It is, for example, possible that the variables in a large model distinguish between schools and therefore in some sense a fixed effects (i.e., heterogeneous) model is estimated. To provide further insight, the large model was used to estimate the standard errors for male and female student proficiency means following Method 1. However, instead of using all components, only the first few were used, from 2 through 20, leaving out an intercept-only model. Table 17 shows that the standard errors increase substantially as the number of principal components used decreases. This requires further careful investigation.

The reason that Binder's method does not impact the final imputation based results for student groups means and standard errors is likely due to the fact that a relatively large number of items was assessed for each student. The result is that the item likelihood under the NAEP model is quite peaked and—relative to the group model—will be the determining factor for the calculation of the student posterior mean. Simply put, regardless of the population mean, each student's ability is reasonably well-estimated. Obviously, it would be interesting to also study

Table 17

Standard Errors of Direct Estimates Based on a Large Model Using Subsets of Principal Components and the Intercept

No. of PCs	Male	Female
2	.022	.022
3	.021	.021
4	.021	.020
5	.020	.020
6	.016	.016
7	.015	.018
8	.015	.018
9	.015	.018
10	.015	.018
11	.014	.018
12	.014	.018
13	.014	.018
14	.013	.018
15	.013	.018
16	.013	.017
17	.013	.017
18	.013	.016
19	.013	.016
20	.013	.016

Note. PCs = principal components.

samples where the number of items per student is much smaller. In that case, a multi-subscale subject would be of interest as the number of items per student and subscale is relatively sparse.

Finally, it should be noted that while much of the methodology of NAEP was used, the only results that exactly follow the operational procedures are in Table 6. Also, for Binder's method

the replicate strata have been used as cluster variable, which may or may not be appropriate. Lastly, a simulation study to further examine the merits of this method is advisable given the surprising results between saturated and small latent regression models.

References

Aitkin, M. (2003). [Review of the paper *Marginal estimation in NAEP: Current operational procedures and AM by John Mazzeo, John R. Donoghue, and Matthew Johnson*]. Washington, DC: National Center for Education Statistics.

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report* (NCES-2001-509). Washington, DC: National Center for Education Statistics.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51,* 279-292.

Chang, H. -H. (1996). The asymptotic posterior normality of the latent trait for polytomous IRT models. *Psychometrika, 61,* 445-463.

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

Cohen, J. & Jiang, T. (2002). *Direct estimation of statistics for the National Assessment of Educational Progress.* Washington, DC: American Institutes for Research.

von Davier, M. (2003). *Comparing conditional and marginal direct estimation of subgroup distributions* (ETS Research Rep. No. RR-03-02). Princeton, NJ: ETS.

von Davier, M., Sinharay, S, Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent development and future directions. In C.R. Rao & S. Sinharay (Eds.) *Handbook of statistics, volume 26.* Amsterdam: Elsevier.

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

Kovar, J. (1985). *Variance estimation of nonlinear statistics in stratified samples* (Working paper No. BSMD 85-052E). Ottawa: Statistics Canada.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mazzeo, J., Donoghue, J. R., Li, D., & Johnson, M. (2006). *Marginal estimation in NAEP: Current operational procedures and AM.* (Available from the National Center of Education Statistics, 1990 K Street NW, Washington, DC 20006).

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993-997.

Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177-196.

von Davier, M. (2003). *Comparing conditional and marginal direct estimation of subgroup distributions* (ETS Research Rep. No. RR-03-02). Princeton, NJ: ETS.

von Davier, M., Sinharay, S, Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent development and future directions. In C.R. Rao & S. Sinharay (Eds.) *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1036–1056). New York: Elsevier.

White, H. (1980). A heteroscedasticity-consistent covarance matrix estimator and a direct test for heteroscedasticity. *Econometrica, 48,* 817-838.

Wolter, K. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.

Appendix

In the multivariate case, regression effects $\boldsymbol{\gamma}$ will become a long vector including the regression effects for each subscale, that is, $\boldsymbol{\gamma}$ is a pQ-dim vector $(\boldsymbol{\gamma}_1', \cdots, \boldsymbol{\gamma}_p')'$ with each subscale regression effects $\boldsymbol{\gamma}_t$ of Q components for $t = 1, \cdots, p$. Let $\boldsymbol{Z}_i$ be the collection of background variables for each student in the p-scale assessment, then

$$\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{x}_i' & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}_i' & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{x}_i' \end{pmatrix}_{p \times pQ}.$$

The Hessian matrix used in (13) and (28) to compute the standard errors for regression effects is defined as

$$\boldsymbol{H}(\boldsymbol{\gamma}) = \frac{\partial^2 L}{\partial \boldsymbol{\gamma}^2}, \tag{27}$$

where L is the total marginal likelihood function for N students' response to the test items, which is expressed as

$$\begin{aligned} L &= log\left[\prod_{i=1}^{N} P(\boldsymbol{y}_i|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})^{w_i}\right] \\ &= \sum_{i=1}^{N} w_i log\left[P(\boldsymbol{y}_i|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})\right] \\ &= \sum_{i=1}^{N} w_i log\left[\int P(\boldsymbol{y}_i|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})d\boldsymbol{\theta}\right]. \end{aligned} \tag{28}$$

$\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})$ represents the multivariate norm density with mean vector $\boldsymbol{Z}_i\boldsymbol{\gamma}$ and covariance matrix $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})$, and the density function is

$$\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{Z}_i'\boldsymbol{\gamma})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{Z}_i'\boldsymbol{\gamma})\right]. \tag{29}$$

The partial derivative of $\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\gamma}$ is

$$\frac{\partial \phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\gamma}} = \phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma}, \boldsymbol{\Sigma})\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{Z}_i\boldsymbol{\gamma}). \tag{30}$$

Therefore,

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{N} w_i \int \frac{P(\boldsymbol{y}_i|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})}{P(\boldsymbol{y}_i)} \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma})d\boldsymbol{\theta} \\ &= \sum_{i=1}^{N} w_i \int P(\boldsymbol{\theta}|\boldsymbol{y}_i)\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma})d\boldsymbol{\theta} \\ &= \sum_{i=1}^{N} w_i \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{Z}_i\boldsymbol{\gamma}). \end{aligned} \tag{31}$$

The gradient defined in (15) and (25) is given by

$$\begin{aligned}\boldsymbol{g}_i(\boldsymbol{\gamma}) &= \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{Z}_i\boldsymbol{\gamma}) \\ &= \boldsymbol{x}_i \otimes \boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\theta}}-\boldsymbol{Z}_i\boldsymbol{\gamma}). \end{aligned} \tag{32}$$

For one subscale case, $\boldsymbol{Z}_i$ becomes $\boldsymbol{x}_i$ and $\boldsymbol{\Sigma}^{-1}$ becomes σ^{-2}, and the gradient $\boldsymbol{g}_i(\boldsymbol{\gamma})$ for $i = 1, \cdots, N$ becomes

$$\boldsymbol{g}_i(\boldsymbol{\gamma}) = \frac{\boldsymbol{x}_i(\tilde{\theta}-\boldsymbol{x}_i\boldsymbol{\gamma})}{\sigma^2}. \tag{33}$$

continue (34) to do the second derivative of L with respect to $\boldsymbol{\gamma}$, then the Hessian matrix $\boldsymbol{H}(\boldsymbol{\gamma})$ can be further written as

$$\boldsymbol{H}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} w_i \frac{\partial}{\partial \boldsymbol{\gamma}} \left[\int \frac{P(\boldsymbol{y}_i|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})}{P(\boldsymbol{y}_i)} \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma})d\boldsymbol{\theta} \right]. \tag{34}$$

Denote the term that includes $\boldsymbol{\gamma}$ in the integral as $\boldsymbol{f}(\boldsymbol{\gamma})$, that is,

$$\boldsymbol{f}(\boldsymbol{\gamma}) = \left[\frac{\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})}{P(\boldsymbol{y}_i)} \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma}) \right]. \tag{35}$$

Then

$$\frac{\partial \boldsymbol{f}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \frac{\partial P(\boldsymbol{y}_i)\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma})\right] - P(\boldsymbol{y}_i)\partial\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma})\right]}{P(\boldsymbol{y}_i)^2}. \tag{36}$$

Let $\boldsymbol{z}_i$ be denoted by the term $\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{Z}_i\boldsymbol{\gamma})$, then (37) can be simplified as (40).

$$\boldsymbol{H}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} w_i \left[\int P(\boldsymbol{y}_i|\boldsymbol{\theta}) \frac{\partial \boldsymbol{f}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} d\boldsymbol{\theta} \right]. \tag{37}$$

The partial derivative of $\boldsymbol{f}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ is given by

$$\begin{aligned}
\frac{\partial \boldsymbol{f}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} &= \frac{\partial P(\boldsymbol{y}_i)\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i'\right] - P(\boldsymbol{y}_i)\partial\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i\right]}{P(\boldsymbol{y}_i)^2} \\
&= \frac{\int P(\boldsymbol{y}_i|\boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i d\boldsymbol{\theta}\left[\phi(\theta|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i'\right] - P(\boldsymbol{y}_i)\partial\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i\right]}{P(\boldsymbol{y}_i)^2} \\
&= \frac{\tilde{\boldsymbol{z}}_i\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i'\right] - \partial\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i\right]}{P(\boldsymbol{y}_i)} \\
&= \frac{\tilde{\boldsymbol{z}}_i\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i'\right] - \left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})(\boldsymbol{z}_i\boldsymbol{z}_i' - \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i)\right]}{P(\boldsymbol{y}_i)}. \qquad (38)
\end{aligned}$$

Substitute (41) back to (40), then

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{\gamma}) &= \sum_{i=1}^{N} w_i \int \left[P(\boldsymbol{y}_i|\boldsymbol{\theta})\frac{\tilde{\boldsymbol{z}}_i\left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})\boldsymbol{z}_i'\right] - \left[\phi(\boldsymbol{\theta}|\boldsymbol{Z}_i\boldsymbol{\gamma},\boldsymbol{\Sigma})(\boldsymbol{z}_i\boldsymbol{z}_i' - \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i)\right]}{P(\boldsymbol{y}_i)}\right] d\boldsymbol{\theta} \\
&= \sum_{i=1}^{N} w_i \left[\tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}'_i - \int (\boldsymbol{z}_i\boldsymbol{z}'_i)P(\boldsymbol{\theta}|\boldsymbol{y}_i)d\boldsymbol{\theta} - \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i\right], \qquad (39)
\end{aligned}$$

where the expression of $\tilde{\boldsymbol{z}}_i$ is given by

$$\begin{aligned}
\tilde{\boldsymbol{z}}_i &= \int \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{Z}_i\boldsymbol{\gamma})P(\boldsymbol{\theta}|\boldsymbol{y}_i)d\boldsymbol{\theta} \\
&= \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{\theta}} - \boldsymbol{Z}_i\boldsymbol{\gamma}). \qquad (40)
\end{aligned}$$

The integration in (42) can be expressed as

$$\begin{aligned}
\int (\boldsymbol{z}_i\boldsymbol{z}'_i)P(\boldsymbol{\theta}|\boldsymbol{y}_i)d\boldsymbol{\theta} &= \int \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{Z}_i\boldsymbol{\gamma})(\boldsymbol{\theta} - \boldsymbol{Z}_i\boldsymbol{\gamma})'\boldsymbol{Z}_i\boldsymbol{\Sigma}^{-1}P(\boldsymbol{\theta}|\boldsymbol{y}_i)d\boldsymbol{\theta} \\
&= \int \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\theta}} - \boldsymbol{Z}_i\boldsymbol{\gamma})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} + \tilde{\boldsymbol{\theta}} - \boldsymbol{Z}_i\boldsymbol{\gamma})'\boldsymbol{Z}_i\boldsymbol{\Sigma}^{-1}P(\boldsymbol{\theta}|\boldsymbol{y}_i)d\boldsymbol{\theta} \\
&= \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i + \tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}'_i. \qquad (41)
\end{aligned}$$

Finally,

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{\gamma}) &= \sum_{i=1}^{N} w_i \left[\tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}'_i - \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i - \tilde{\boldsymbol{z}}_i\tilde{\boldsymbol{z}}'_i - \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i\right] \\
&= -\sum_{i=1}^{N} w_i \left[\boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i + \boldsymbol{Z}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_i\right] \\
&= -\sum_{i=1}^{N} w_i \left[(\boldsymbol{x}_i\boldsymbol{x}_i') \otimes \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1} + (\boldsymbol{x}_i\boldsymbol{x}_i') \otimes \boldsymbol{\Sigma}^{-1}\right] \\
&= -\sum_{i=1}^{N} w_i \left[(\boldsymbol{x}_i\boldsymbol{x}_i') \otimes (\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1})\right] \qquad (42)
\end{aligned}$$

For univariate case,

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\gamma}) &= -\sum_{i=1}^{N} w_i \left[(\boldsymbol{x}_i \boldsymbol{x}_i')(\sigma^{-2}\tilde{\sigma}_i^2\sigma^{-2} + \sigma^{-2}) \right] \\ &= -\frac{1}{\sigma^4} \sum_{i=1}^{N} w_i \boldsymbol{x}_i \boldsymbol{x}_i' (\tilde{\sigma}_i^2 + \sigma^2). \end{aligned} \tag{43}$$