Research Report

Evaluation of Methods to Compute Complex Sample Standard Errors in Latent Regression Models

Andreas Oranje

Deping Li

Mathew Kandathil

December 2009

ETS RR-09-49

Evaluation of Methods to Compute Complex Sample Standard Errors in Latent Regression Models

Andreas Oranje, Deping Li, and Mathew Kandathil
ETS, Princeton, New Jersey

December 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

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

Abstract

Several complex sample standard error estimators based on linearization and resampling for the latent regression model of the National Assessment of Educational Progress (NAEP) are studied with respect to design choices such as number of items, number of regressors, and the efficiency of the sample. This paper provides an evaluation of the extent these estimators are appropriate for the models and test lengths often encountered in NAEP and what the effect is on the NAEP imputation model. It is shown that in general the resampling method used in this study provides the most accurate standard errors. However, the differences with the linearization method chosen in this study are relatively small if only small models are used with respect to the independent variables of the latent regression. Illustration is provided through several small simulation studies and NAEP data analysis.

Key words: Linearization, resampling, estimators, NAEP

Ackowledgments

The authors would like to thank John Mazzeo, Jiahe Qian, John Donoghue, Xueli Xu, Matthias von Davier, and Shelby Haberman for many discussions, insights, and suggestions. We also thank the editor and two reviewers for their excellent comments.

1 Introduction

The National Assessment of Educational Progress (NAEP) has two key design characteristics: (a) NAEP assesses probability samples of students within schools and schools within geographic regions, resulting in complex samples and (b) each student answers only a systematic portion of the cognitive item pool so that all students combined provide an approximately equal number of responses to each item in the item pool. The result of this design is that students cannot be compared to each other directly and that individual proficiency estimates cannot be obtained reliably. However, point estimates can be obtained for groups of students based on posterior distributions of proficiency. To estimate these distributions, a latent regression model (Mislevy, 1984, 1985) is used. Subsequently, post-hoc procedures are applied to obtain appropriate standard errors for NAEP's statistics of interest, taking into account the complexity of the sample (Mislevy, 1991; Rust& Johnson, 1992).

Cohen and Jiang (2002) adapted a method developed by Binder (1983) to compute complex sample standard errors through linearization. They implemented this method in AM software (American Institute of Research [AIR], 2003), which can be used to analyze a restricted version of the current NAEP model. This method was also implemented in the current NAEP operational latent regression estimation software (DGROUP), and it was found that the estimator is similar to simple random sample approximations for larger models (Li & Oranje, 2007). In addition, Li and Oranje found that a substantial increase in regression effect standard errors from Cohen and Jiang's method did not affect the distribution of NAEP's imputations (i.e., *plausible values*), regardless of the size of the population model. The size of the population model is the number of student groups used as predictors in the latent regression. Li and Oranje noted in their analyses that relatively many items per student (approximately 80) were available. In this paper, these findings are further investigated by addressing the following two questions.

1. How do various methods for estimating standard errors of latent regression model parameters, including simple random sample approaches, Binder's method, and re-sampling approaches, compare in terms of accuracy and efficiency?

2. How do two design variables, the number of items per student and the population model size, affect these standard errors and the distribution of plausible values?

Plausible values are the result of an imputation step (see section 4.3) to represent measurement variablity due to the fact that proficiency is a latent construct. The connection between the complex sample standard errors and the plausible values is made in this paper because the imputation model is partly based on the standard errors of the model parameters. Hence, if plausible values are not appreciably affected by the populaton model, then a crude approximation of the population parameter standard errors is defensible. However, if plausible values are substantially affected by those standard errors, then approximations have to be carefully evaluated.

1.1 Model

The latent regression model used in NAEP is:

$$\theta_i = \vec{\gamma}' \vec{x_i} + \epsilon_i, \tag{1}$$

where θ_i is the latent proficiency for student $i = 1, 2, \ldots, N$, $\vec{\gamma}$ is a vector of latent regression coefficients, $\vec{x_i}$ is a vector of observed student group indicators, and ϵ_i is a residual term, assumed to be normally distributed. The latent proficiency distribution is characterized by a collection of item response theory (IRT) models (Lord, 1968), and the predictors typically contain information about demographics of a student, his or her school, and his or her teacher(s). Under current NAEP practice, the parameters of this collection of IRT models are estimated in a prior step and considered known when the latent regression is conducted.

Assuming that item responses $\vec{y_i}$ are independently distributed from student group indicators $\vec{x_i}$ conditional on proficiency θ_i, the posterior distribution for a student can be written as

$$P(\theta_i; \vec{y_i}, \mu_i, \vec{\beta}, \sigma^2) \propto L(\vec{y_i}; \theta_i, \vec{\beta}) \phi(\theta_i; \mu_i, \sigma^2), \tag{2}$$

where $\mu_i = \vec{\gamma}' \vec{x_i}$, $\vec{\beta}$ is a vector of known parameters to further specify the item response model, σ^2 is the residual variance, and ϕ represents the normal distribution function. The right side of (2)(2) contains two parts: The first term is the likelihood L of observing a

particular cognitive response pattern for a given θ, and the second term represents the distribution of θ in the population of interest. The first term is a product of probabilities for each of $j = 1, 2, \ldots, n$ items assuming local independence

$$L(\vec{y_i}; \theta_i, \vec{\beta}) = \prod_{j=1}^{n} P(y_{ij}; \theta_i, \beta_j), \tag{3}$$

where each $P(y_{ij}; \theta_i, \beta_j)$ is the probability that examinee i with proficiency θ_i answers item j correctly. By marginalizing with respect to θ and computing maximum likelihood estimates, parameters are estimated in NAEP. A multivariate version of this model follows by assuming that the likelihood can be factorized for each dimension. For T dimensions, the posterior distribution of $\vec{\theta_i}$ is

$$P(\vec{\theta_i}; \vec{y_i}, \vec{\mu_i}, \vec{\beta}, \mathbf{\Sigma}) \propto \left(\prod_{t=1}^{T} L(\vec{y_{it}}; \theta_{it}, \vec{\beta}) \right) \phi_T(\vec{\theta_i}; \vec{\mu_i}, \mathbf{\Sigma}), \tag{4}$$

where ϕ_T is a T-variate normal distribution.

1.2 Estimation

The model parameters in (2) are estimated using an EM algorithm, where the maximization step in the univariate case involves the following two quantities

$$\vec{\hat{\gamma}} = \mathbf{\Xi} \boldsymbol{X}' \mathbf{D}_w \vec{\tilde{\theta}} \tag{5}$$

and

$$\hat{\sigma}^2 = \frac{1}{w_+} \sum_{i=1}^{N} \left(w_i V(\tilde{\theta}_i) + w_i \left(\tilde{\theta}_i - \vec{\hat{\gamma}}' \vec{x_i} \right)^2 \right), \tag{6}$$

where $\boldsymbol{X}$ is an N by g data matrix of N students by g student group indicators, $\mathbf{\Xi}$ is $(\boldsymbol{X}'\boldsymbol{D_w}\boldsymbol{X})^{-1}$, $\boldsymbol{D_w}$ is a diagonal matrix of student sampling weights w_i, and $\vec{\tilde{\theta}}$ is a vector of posterior means, where each element is computed by (7). Furthermore, $V(\tilde{\theta}_i)$ is the posterior variance for a student, computed by (8). Also, w_+ is the sum of student sampling weights.

The expectation step involves computation of posterior moments, which are calculated as

$$\tilde{\theta}_i = \frac{\int_\theta \theta L(y_i; \theta, \vec{\beta}) \phi(\theta; \mu_i, \sigma^2) d\theta}{\int_\theta L(y_i; \theta, \vec{\beta}) \phi(\theta; \mu_i, \sigma^2) d\theta} \tag{7}$$

and

$$V(\tilde{\theta}_i) = \frac{\int_\theta (\theta - \tilde{\theta}_i)^2 L(y_i; \theta, \vec{\beta}) \phi(\theta; \mu_i, \sigma^2) d\theta}{\int_\theta L(y_i; \theta, \vec{\beta}) \phi(\theta; \mu_i, \sigma^2) d\theta}, \tag{8}$$

where $\mu_i = \vec{\hat{\gamma}}' \vec{x_i}$.

The variance of $\vec{\hat{\gamma}}$ is estimated by the following equation

$$\mathbf{V}(\vec{\hat{\gamma}}) = E(\mathbf{V}(\vec{\hat{\gamma}})) + \mathbf{V}(E(\vec{\hat{\gamma}})) = \sigma^2 \boldsymbol{\Xi} + \boldsymbol{\Xi} \boldsymbol{X}' \boldsymbol{D}_{\boldsymbol{w}} \boldsymbol{D}_{\boldsymbol{V}(\tilde{\boldsymbol{\theta}})} \boldsymbol{D}_{\boldsymbol{w}} \boldsymbol{X} \boldsymbol{\Xi}, \tag{9}$$

where $\mathbf{D}_{\mathbf{V}(\tilde{\theta})}$ is a diagonal matrix of dimension N with elements equal to the posterior variance of student i. For assumptions about the two terms in (9) see Mardia, Kent, and Bibby (1979, Equations 6.6.5 and 6.6.6, p. 172).

1.3 Student Groups

Students belong to many different groups as defined by variables such as:

- Demographics (e.g., gender, race/ethnicity, parental education)
- Home environment (e.g., number of books in the home)
- School factors (e.g., public or private, location)
- Teacher factors (e.g., professional training)

There are several hundred variables collected and in combination a contingency table can be constructed with a large number of cells. Hence, nearly every student represents a single group and many cells are empty. In current NAEP procedures, predominantly main effects in this contingency table are used. Specifically, most variables appear independently in the model and a small set of two- and three-way tables is used for key reporting variables. This still yields a large number of cells.

Under current NAEP procedure, this large contingency table is reduced by dummy-coding the student group indicators and extracting principal components. The dummy-coding is used to facilitate regression type analysis with categorical variables. The extraction of principal components is to reduce the number of variables. Specifically, a set

of components that explain 90% of the variance is retained. Subsequently, factor scores are computed and used in (1). For the purpose of this paper, it is assumed that the principal component factor scores are an equivalent representation of the dummy codes. Hence, these factor scores are not separately studied. Incidentally, it should be noted that the dummy-coding scheme and subsequent analyses assume that these codes are continuous variables, which might not be entirely appropriate, yet is common practice.

One of the key issues addressed in this study is the effect of a large population model (i.e., many predictors) on the appropriateness of estimates. Although never verified in the context of NAEP, this is of course a well-known issue. Several consistency and asymptotic normality results are available (Haberman, 1977a, 1977b) for exponential response models commonly utilized in NAEP. For the general linear model, results from Portnoy (1984, 1988) indicate that $q^2/N \to 0$ is required for asymptotic normality of $\hat{\gamma}$ where q is the number of parameters. Under that requirement, a usual NAEP state assessment with approximately 2,750 sampled students would be able to support 37 parameters if converging to zero is set equal to .5 and without consideration for design effects. It is of course possible that asymptotic normality can be reached with larger numbers of parameters and fewer students or that this can be obtained for individual coordinates. Verification of those cases is necessary.

NAEP recognizes two primary statistics of interest: student-group means and percentages of students at specific proficiency levels. Both statistics are averaged across plausible values and standard errors are computed as the sum of the between imputation variance and a resampling based variance estimate. The first term of the standard error reflects the variation due to the latency of the construct and the second term represents the variation due to sampling.

1.4 NAEP's Sampling Design

NAEP samples are complex in the sense that students are sampled within schools. Students within a school most likely have similar learning experiences and environments and therefore their responses carry some form of dependency. At the first stage, schools

are sampled from geographic regions. In most cases, these regions are determined by state borders or metropolitan areas and, hence, most schools within a region operate under similar policies, curricula, and funding. Hence, a second level of dependency might exist. In this section, the stratified two-stage sampling design is more precisely described.

NAEP is assessed in two types of samples: national and combined samples. Combined samples are a combination of (all) state samples into one large national sample, while national samples are only representative at the national level. Therefore, the national sample allows for reporting results for student groups at the national level (e.g., female 4th graders), while the combined sample also allows for state-specific results (e.g., female 4th graders in Arizona). Depending on the subject and the grade, samples are either national or combined. For either sample, first the stratification in public schools is determined.

In national samples, stratification is done by region of the country (northeast, south, midwest, and west) and metropolitan statistical area status (yes or no). At the first stage, a systematic sample of primary sampling units (PSU) is drawn with probability proportional to size (PPS) and without replacement in each stratum. A PSU is a county or a group of adjacent counties. Finally, students are sampled following a simple random sample scheme.

In combined samples, the states and Washington, DC serve as primary strata. Instead of PSUs being counties or a group of adjacent counties, in combined samples a systematic sample of schools is directly drawn PPS and without replacement and therefore one less step is conducted. At the second stage, a simple random sample of students is selected from each selected school.

Despite the fact that this is already a complex sampling structure, in practice the sample is even more complex. Private schools are selected separately and follow a different stratification. Also, in most 4^{th} grade schools the option is utilized to assess *all* eligible students to decrease the logistical burden.

2 Theory

Question 1 addresses the extent to which several methods used or proposed to be used in NAEP are effective in estimating complex sample variability in circumstances of interest

to NAEP. To address this question, first these methods are discussed and, subsequently, some empirical work is presented.

2.1 Binder's (1983) Method

Binder's (1983) method, as proposed by Cohen and Jiang (2002) for use in NAEP, is based on a Taylor series expansion of $\mathbf{V}(\vec{\hat{\gamma}})$ at $\hat{\gamma} = \gamma_0$, where γ_0 is the population parameter value. It has the following form

$$\mathbf{V}(\vec{\hat{\gamma}}) = \mathbf{H^{-1}\Omega H^{-1}}, \tag{10}$$

where the Hessian is computed as

$$\mathbf{H} = -\sum_{i=1}^{N} w_i \frac{\left(V(\tilde{\theta})_i + \sigma^2\right) \vec{x}_i \, \vec{x}_i'}{\sigma^4} \tag{11}$$

and the variance of the population value is estimated according to Cohen and Jiang as

$$\mathbf{\Omega} = \sum_{h=1}^{H} \frac{w_+^h}{w_+^h - 1} \sum_{c=1}^{C} (f_{ch} - \bar{f}_h)(f_{ch} - \bar{f}_h)' \tag{12}$$

for $c \in [1, 2, \ldots, C]$ clusters and $h \in [1, 2, \ldots, H]$ strata. Also, w_+^h is the sum of weights in stratum h and f_{chi} is the gradient for student i in cluster c and stratum h, which is computed as

$$f_{chi} = \frac{\vec{x}_{chi} \, \tilde{\theta}_{chi} - \vec{x}_{chi} \vec{x}_{chi}' \vec{\gamma}}{\sigma^2} \tag{13}$$

and $\bar{f}_h$ is the average f_{ch} across all C clusters in stratum h and $f_{ch} = \sum_{i=1}^{N_{ch}} w_i f_{chi}$. Subsequently, the variance of a group mean can be computed by

$$V(\hat{\mu}_\theta) = \vec{x}_.' \, \mathbf{V}(\vec{\hat{\gamma}}) \, \vec{x}_., \tag{14}$$

where $\vec{x}_.$ is a vector of average weighted student indicators across all students in the sample of interest (e.g., this could be all students in the sample or a specific subset such as males or females).

2.2 Approximation of the Hessian Matrix

According to Cohen and Jiang (2002), (11) can be approximated by

$$\mathbf{H} = -\sum_{i=1}^{N} w_i f_i' f_i. \tag{15}$$

Subsequently, this implies that

$$-\mathbf{H} = \sum_{i=1}^{N} w_i \frac{\left(V(\tilde{\theta})_i + \sigma^2\right) \vec{x_i}\vec{x_i}'}{\sigma^4} \approx \sum_{i=1}^{N} w_i \left(\frac{\vec{x_i}\ \tilde{\theta}_i - \vec{x_i}\vec{x_i}'\vec{\gamma}}{\sigma^2}\right)^2, \tag{16}$$

which, essentially, is only true if $V(\tilde{\theta})_i$ is very small as $\sigma^2 \approx E[V(\tilde{\theta})_i] + E(\tilde{\theta}_i - \vec{x_i}'\vec{\gamma})^2$. It can reasonably be expected that in an assessment with many items per student the approximation is quite accurate. However, in an assessment with few items per student, the approximation is likely to be poor. In concrete terms, the Hessian matrix in (15) might be underestimated and, therefore, the standard error might be overestimated as the inverse of the Hessian is pre- and postmultiplied. In addition, it should be noted that there is a striking difference in the level of complexity of the approximations presented here and more standard procedures presented in Cochran (1977).

2.3 Resampling

The current NAEP methodology uses resampling to estimate standard errors, taking the complex nature of the sample into account. Specifically, a leave-out-group jackknife procedure is used based on the PSUs. One or more of those units are removed from the sample and the statistic of interest is re-estimated. This process is repeated a reasonably large number of times to quantify the variability in the sample. Details about this procedure can be found in Allen, Donoghue, and Schoeps (2001). This method and related methods, such as balanced repeated replications and variations thereof, are predominantly used in the practice of survey research. The argument for using empirical approximations is that the complexity of the sampling practice implies that appropriate analytic formulae are exceedingly complex to derive. This issue can be circumvented by using empirical methods, although the question surfaces whether it provides a satisfactory result under all relevant

circumstances. A proposal from Qian (2005) and Qian and Haberman (2006) is to apply the jackknife not only to NAEP's reporting statistics, but also to the latent regression model itself, essentially re-estimating the model with resamples. This approach has been followed in this study.

The usual jackknife variance estimator is (Wolter, 1985)

$$V(\hat{\mu}_\theta) = \frac{C-1}{C} \sum_{c=1}^{C} \left(\hat{\mu}_\theta^{(c)} - \hat{\mu}_\theta \right)^2, \tag{17}$$

where C is the total number of clusters and $\hat{\mu}_\theta^{(c)}$ is the statistic based on the total sample except cluster c. This estimator is slightly different from NAEP's jackknife repeated replications (JRR) approach. In NAEP's JRR, 62 pairs of equivalent clusters (i.e., primary sampling units or schools) are formed based on auxiliary income or proficiency data. Subsequently, for one pair at a time, one of the two units is removed, the other unit is doubled in weight, and the statistic of interest is computed based on this modified sample, serving as $\hat{\mu}_\theta^{(c+)}$ in the following equation

$$V^{JRR}(\hat{\mu}_\theta) = \sum_{c=1}^{C} \left(\hat{\mu}_\theta^{(c+)} - \hat{\mu}_\theta \right)^2. \tag{18}$$

However, for the comparisons in this paper, the usual jackknife variance estimator in (17) is used.

3 Method

To investigate the research questions, a simulation study was conducted followed by some real data analysis. The goals of the simulation were to look at relatively simple models and small sample sizes to evaluate a large number of conditions. While that strategy can illustrate important mechanisms underlying each of the approaches, the disadvantage is that the simulation is somewhat disconnected from operational NAEP procedures. Hence, real data analyses were conducted to relate the simulation results to NAEP applications.

There were several steps involved in the simulation:

1. For each of 100 replications and for each condition described below, parameters of a

two-parameter item response model were generated, specifically $a \sim U(0.5, 1.4)$ and $b \sim N(0, 1)$.

2. Population parameters were generated following $\gamma \sim N(0, 0.04)$, to reflect a typical range of regression effects found in operational NAEP.

3. Ability parameters were drawn for each cluster $c \in [1, 2, \ldots, 60]$ from a multivariate normal distribution with dimension 50 (i.e., the number of students per cluster, where the first 50 students are in Cluster 1, the second 50 in Cluster 2 and so on), $\theta_c \sim MVN(\mu_c, \mathbf{S})$, where $\mu_c = (\gamma' x_{c1}, \gamma' x_{c2}, \ldots, \gamma' x_{c50})$ and

$$\mathbf{S} = \begin{vmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ r_{ci,ci'} & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{vmatrix}$$

4. Three variables were manipulated in this study: (a) the level of clustering, $r_{ci,ci'} \in [0, 0.2, 0.5, 0.8]$; (b) the number of items per student, $k \in [10, 30, 70, 90]$; and (c) the level of saturation of the population model, reflected in the number of students per group g, $N_g \in [3000, 600, 300, 100, 50]$, or, conversely, the number of regression coefficients needed to represent all G groups, $G \in [1, 5, 10, 30, 60]$

   The matrix $\mathbf{x} = I_G \otimes i'_{N_g}$ was used to indicate which student belongs to what group via dummy-coding, where i'_{N_g} is a N_g-dimensional unit column vector. In an unsaturated model, there is only one group of 3000 students and $\mathbf{x}$ is a 3000-dimensional unit vector. In a saturated population model, there would be 3000 groups with each a single student and $\mathbf{x}$ would be a 3000×3000 unit diagonal matrix. This model is not identified. In this study, there were always 50 students in 60 clusters and, therefore, the condition with $N_g = 50, G = 60$ represented a complete fixed effect hierarchical model.

5. Based on θ, item responses were generated according to the model. Different than in NAEP, items were not sampled across simulees.

6. Population model parameters and posterior moments were estimated following standard NAEP methodology for marginal maximum likelihood estimation of one-dimensional problems. Hence, the method of integration was a simple rectangular rule with 41 equally spaced points in the -5 to 5 interval for a normal prior. A solution was deemed converged when none of the parameters changed more than $1 \cdot 10^{-6}$ between two consecutive iterations.

7. The following variance estimates were computed: (a) the empirical variance based on 1,000 additional replications, taken as the true variance; (b) a simple random sample approach, following (9); (c) Binder's method, following (10); (d) Binder's method with the approximated Hessian, following (15); and (e) a jackknife approach, following (17).

   For Methods c through e, the cluster designation used to generate the data was also used as the cluster variable for computing variances.

4 Results

4.1 Simulation Study

In this section, the model parameters for the simple, nonclustering case are first inspected, followed by standard errors evaluated across several levels of clustering. With respect to the results for model parameters, first residual variances and then regression coefficients are discussed. It should be noted that convergence of the estimation process was carefully monitored and no issues (e.g., nonconvergence) were encountered.

4.1.1 Residual Variances

Table 1 shows residual variance estimates $\hat{\sigma}^2$ and Table 2 shows squared prediction errors $\hat{\epsilon}_i^2 = \left(\tilde{\theta}_i - \hat{\gamma} x_i\right)^2$ averaged across 100 replications. Note that σ^2 is the sum of the average posterior variance and $E\left(\epsilon^2\right)$, as shown in (6). Table 1 shows that the residual variance decreases as the number of groups G increases. The number of items have no

particular influence on the size of the residual variance, except for the case where the number of items are small (i.e., $n = 10$). In contrast, Table 2 shows that the mean squared regression error does increase as the number of items, k, increases. Also, this error becomes smaller as the number of groups, G increases. In other words, as k increases, the information about an individual student becomes more precise and, subsequently, the regression error becomes larger as the population model fits the data relatively poorly. One interpretation would be that the posterior distributions are becoming peaked and, hence, the influence of the population model decreases. Another way to express this is that there is less opportunity for regression towards the mean.

The observation that the residual variance is largely constant across the number of items, k, can be explained from the variance decomposition of the residual variance into posterior variance and ϵ^2. As the number of items increases, the success of prediction by the population model decreases, but the certainty about students' ability increases commensurately, shifting variability from one term to the other.

Table 1
Average Residual Variance Estimates, $\hat{\sigma}^2$***, for Varying Number of Items,*** $k \in [10, 30, 70, 90]$***, and Number of Groups,*** $G \in [1, 5, 10, 30, 60]$***, Across*** 100 ***Replications***

	k			
G	10	30	70	90
1	0.999	1.001	0.999	0.999
5	0.969	0.962	0.965	0.969
10	0.961	0.961	0.964	0.960
30	0.950	0.953	0.956	0.954
60	0.934	0.940	0.944	0.944

4.1.2 Regression Effects

As far as the regression effects are concerned, Table 3 shows that the absolute bias of regression effect estimates seems largely unaffected by the number of items. However, the saturation of the model does affect the bias. This is to be expected because the saturation

Table 2
Average Squared Regression Error, $\hat{\epsilon}^2$***, for Varying Number of Items,*** $k \in [10, 30, 70, 90]$***, and Number of Groups,*** $G \in [1, 5, 10, 30, 60]$***, Across*** 100 ***Replications***

	k			
G	10	30	70	90
1	0.776	0.913	0.958	0.967
5	0.749	0.874	0.925	0.937
10	0.747	0.874	0.924	0.928
30	0.735	0.867	0.916	0.923
60	0.719	0.854	0.904	0.913

Table 3
Average Absolute Difference Between γ ***and*** $\hat{\gamma}$ ***Across*** G ***Parameters, for Varying Number of Items,*** $k \in [10, 30, 70, 90]$***, and Number of Groups,*** $G \in [1, 5, 10, 30, 60]$***, Across*** 100 ***Replications***

	k			
G	10	30	70	90
1	0.016	0.015	0.014	0.015
5	0.036	0.032	0.034	0.032
10	0.052	0.048	0.047	0.047
30	0.090	0.083	0.079	0.078
60	0.127	0.117	0.113	0.112

determines the number of students a coefficient estimate is based upon.

4.1.3 Variances

One of the findings in Li and Oranje (2007) was that when a large nearly saturated model was used, the Binder (1983) estimator appeared to severely underestimate the standard error of the latent regression coefficients and was in some cases close to a simple random sample estimate. In the following tables, several approaches to the estimation of variances will be presented, averaged across regression effects and across repetitions. These are compared to the empirical variance (i.e., the variance across 1,000 additional replications). The four approaches are the simple random sample approach in (9), Binder's

formula in (10), Binder's formula with an approximated Hessian matrix from (15), and a leave-out group jackknife following (17).

Samples without clustering. The left side of Table 4 shows the ratio of the variance estimates and the empirical variance (mean-squared error, or MSE) without clustering. The top set of rows provides the MSE, which behaves mostly according to expectation: The variability increases if there are fewer simulees per group. Also, the variability tend to decrease if more items are assessed per simulee. However, this decrease is minor because the success of prediction with the regression model reduces as well, even though more items means less measurement variance. As far as the variance computation approaches under investigation, the simple random sample approach appears to mostly underestimate the empirical variance unless the number of items is substantial and the model not too large. Because the true proficiency for each simulee is in part based on group membership, even the nonclustering condition is in fact not a simple random sample, but is sampled following a fixed effect multilevel model. Even though the model explains more variability as the number of groups increases, the number of simulees available to estimate each regression effect negatively impacts the variability (i.e., a smaller sample size increases the variability).

Binder's approach appears to yield underestimates of the empirical variance by at least 60% but improves markedly as the number of items per student increases. Also, Cohen and Jiang's (2002) approximation generally provides reasonable variance estimates as long as the model is not too large. In a large model, this method breaks down completely as all the cluster variance is explained by the model. In contrast, the jackknife approach is generally on target, but tends to show some overestimation for larger models. For the condition $G = 30$, the variances are twice as large as the empirical variances regardless of the number of cognitive items assessed per student. In this case, for each jackknife replicate, half of the sample of a particular group is removed. Subsequently, a possible explanation for this finding could be that the regression effect estimation for that group is relatively unstable. In the condition $G = 60$, the model represents the clustering in the sense that every cluster mean is estimated by a separate parameter. Binder's (1983)method, Cohen and Jiang's approximation, and the jackknife approach display severe underestimation. For

the jackknife approach, this can be expected as the regression effects for all other groups are relatively indifferent to the removal of one cluster and no regression effect is estimated for the removed cluster to assure identification. For the other two estimators, the model explains all variability between clusters and, hence, the quantity Ω in (12) reduces to zero.

Clustered samples. In the right side of Table 4 and 5, average variance estimates can be found for samples with clustering effects $r_{ci,ci'} = 0.2$, 0.5, and 0.8. The results are quite similar to the results for the condition where the within-cluster correlation is zero. The variance does increase with increasing clustering for all methods except for the simple random sampling approach as the number of groups increase. This is due to the fact that models with many student groups explain the variability associated with clustering better and, hence, the residual variance decreases.

4.2 Real Data Analysis

In this section real data analyses will be presented. NAEP 2005 reading data in Grade 12 and NAEP 2004 mathematics long-term trend data for Age 17 was analyzed and standard error computations were conducted using the methods described in the theory section. For detailed information about sample sizes, predictors, and the instrument designs, Perie, Grigg, and Donahue (2005) can be consulted for reading and Perie and Moran (2005) for long-term trend. Because the accuracy of our implementation of Binder's (1983) method has to be established, parallel runs with AM beta version 0.06.03 (AIR, 2003) were conducted alongside the current NAEP operational methodology. AM is publicly available software for the analysis of proficiency data for large-scale assessments and has several specialized NAEP analysis modules. For standard errors, AM computes the Binder's method standard errors with approximation, which can be compared only to those in our implementation. For the standard errors in Tables 6-8, Jackknife is computed following (17) and Binder's approach following (10) through (13). Approx. represents the approximation as implemented in AM, where (11) is approximated by (15). The standard error from the AM is also provided as is the simple random sample (SRS) estimator used in NAEP for the imputation model. The data sets chosen are two assessments with relatively few

Table 4

Mean Squared Error (MSE) Based on 1000 Repetitions and Average Estimated Ratios of the Variance of the Regression Coefficients and the MSE for Several Methods, for Varying Number of Items, $k \in [10, 30, 70, 90]$***, and Groups,*** $G \in [1, 5, 10, 30, 60]$***, Across*** 100 ***Repetitions for Samples With No and Mild Clustering,*** $r_{ci,ci'} \in [0, .2]$

r_c	0.0				0.2			
k	10	30	70	90	10	30	70	90
G	MSE							
1	4.05 d^{-4}	3.59 d^{-4}	3.71 d^{-4}	3.27 d^{-4}	3.46 d^{-3}	3.48 d^{-3}	3.64 d^{-3}	3.79 d^{-3}
5	2.13 d^{-3}	1.87 d^{-3}	1.70 d^{-3}	1.69 d^{-3}	1.79 d^{-2}	1.77 d^{-2}	1.73 d^{-2}	1.74 d^{-2}
10	4.24 d^{-3}	3.56 d^{-3}	3.32 d^{-3}	3.32 d^{-3}	3.60 d^{-2}	3.52 d^{-2}	3.54 d^{-2}	3.38 d^{-2}
30	1.22 d^{-2}	1.06 d^{-2}	1.01 d^{-2}	9.99 d^{-3}	1.07 d^{-1}	1.04 d^{-1}	1.05 d^{-1}	1.04 d^{-1}
60	2.51 d^{-2}	2.11 d^{-2}	1.98 d^{-2}	1.98 d^{-2}	2.12 d^{-1}	2.07 d^{-1}	2.09 d^{-1}	2.09 d^{-1}
	Simple random sample approach/MSE							
1	8.21 d^{-1}	9.29 d^{-1}	8.98 d^{-1}	9.60 d^{-1}	9.58 d^{-2}	9.54 d^{-2}	9.15 d^{-2}	8.80 d^{-2}
5	7.59 d^{-1}	8.61 d^{-1}	9.48 d^{-1}	9.56 d^{-1}	8.90 d^{-2}	9.02 d^{-2}	9.29 d^{-2}	9.19 d^{-2}
10	7.57 d^{-1}	9.02 d^{-1}	9.68 d^{-1}	9.63 d^{-1}	8.62 d^{-2}	8.90 d^{-2}	8.80 d^{-2}	9.15 d^{-2}
30	7.81 d^{-1}	9.01 d^{-1}	9.46 d^{-1}	9.56 d^{-1}	8.10 d^{-2}	8.31 d^{-2}	8.26 d^{-2}	8.25 d^{-2}
60	7.47 d^{-1}	8.95 d^{-1}	9.56 d^{-1}	9.57 d^{-1}	7.12 d^{-2}	7.29 d^{-2}	7.23 d^{-2}	7.21 d^{-2}
	Binder's approach/MSE							
1	4.18 d^{-1}	7.12 d^{-1}	8.04 d^{-1}	8.20 d^{-1}	4.31 d^{-1}	7.09 d^{-1}	8.11 d^{-1}	8.15 d^{-1}
5	3.65 d^{-1}	6.08 d^{-1}	7.63 d^{-1}	8.19 d^{-1}	3.58 d^{-1}	6.45 d^{-1}	7.94 d^{-1}	8.12 d^{-1}
10	3.34 d^{-1}	5.83 d^{-1}	7.27 d^{-1}	7.33 d^{-1}	3.25 d^{-1}	5.70 d^{-1}	6.94 d^{-1}	7.21 d^{-1}
30	2.01 d^{-1}	3.31 d^{-1}	4.32 d^{-1}	4.37 d^{-1}	1.77 d^{-1}	3.37 d^{-1}	4.12 d^{-1}	4.27 d^{-1}
60	9.58 d^{-32}	1.19 d^{-31}	1.33 d^{-31}	1.32 d^{-31}	4.16 d^{-32}	4.34 d^{-32}	4.49 d^{-32}	4.60 d^{-32}
	Cohen et al.'s approximation/MSE							
1	1.03 d^{+0}	1.01 d^{+0}	9.45 d^{-1}	9.31 d^{-1}	1.06 d^{+0}	1.01 d^{+0}	9.53 d^{-1}	9.25 d^{-1}
5	9.16 d^{-1}	8.74 d^{-1}	9.04 d^{-1}	9.36 d^{-1}	8.84 d^{-1}	8.99 d^{-1}	9.09 d^{-1}	8.98 d^{-1}
10	8.28 d^{-1}	8.45 d^{-1}	8.61 d^{-1}	8.40 d^{-1}	7.98 d^{-1}	7.78 d^{-1}	7.67 d^{-1}	7.73 d^{-1}
30	5.13 d^{-1}	4.84 d^{-1}	5.20 d^{-1}	5.07 d^{-1}	4.12 d^{-1}	4.05 d^{-1}	3.96 d^{-1}	3.97 d^{-1}
60	2.67 d^{-31}	1.85 d^{-31}	1.72 d^{-31}	1.65 d^{-31}	1.67 d^{-31}	8.28 d^{-32}	6.45 d^{-32}	6.13 d^{-32}
	Jackknife/MSE							
1	1.06 d^{+0}	1.03 d^{+0}	9.61 d^{-1}	9.47 d^{-1}	1.07 d^{+0}	1.03 d^{+0}	9.68 d^{-1}	9.38 d^{-1}
5	1.07 d^{+0}	1.02 d^{+0}	1.05 d^{+0}	1.09 d^{+0}	1.06 d^{+0}	1.08 d^{+0}	1.10 d^{+0}	1.08 d^{+0}
10	1.17 d^{+0}	1.19 d^{+0}	1.21 d^{+0}	1.18 d^{+0}	1.19 d^{+0}	1.17 d^{+0}	1.17 d^{+0}	1.17 d^{+0}
30	2.00 d^{+0}	1.87 d^{+0}	2.01 d^{+0}	1.96 d^{+0}	1.97 d^{+0}	2.02 d^{+0}	1.97 d^{+0}	1.96 d^{+0}
60	1.18 d^{-4}	1.35 d^{-5}	3.54 d^{-6}	2.40 d^{-6}	2.56 d^{-5}	3.92 d^{-6}	8.83 d^{-7}	5.89 d^{-7}

Note. $d = 10$. MSE = mean-squared error.

Table 5

Mean-Squared Error (MSE) and Average Estimated Ratio of the Variance of the Regression Coefficients and the MSE for Several Methods, for Varying Number of Items, $k \in [10, 30, 70, 90]$***, and Groups,*** $G \in [1, 5, 10, 30, 60]$***, Across*** 100 ***Repetitions for Clustered Samples With*** $r_{ci,ci'} \in [0.5, 0.8]$

r_c	0.5				0.8			
$k =$	10	30	70	90	10	30	70	90
G	MSE							
1	8.85 d^{-3}	8.97 d^{-3}	9.19 d^{-3}	9.52 d^{-3}	1.36 d^{-2}	1.53 d^{-2}	1.35 d^{-2}	1.36 d^{-2}
5	4.10 d^{-2}	4.17 d^{-2}	4.10 d^{-2}	4.25 d^{-2}	6.51 d^{-2}	6.47 d^{-2}	6.50 d^{-2}	6.43 d^{-2}
10	8.27 d^{-2}	8.19 d^{-2}	8.31 d^{-2}	8.18 d^{-2}	1.32 d^{-1}	1.32 d^{-1}	1.29 d^{-1}	1.29 d^{-1}
30	2.49 d^{-1}	2.47 d^{-1}	2.43 d^{-1}	2.46 d^{-1}	3.95 d^{-1}	3.91 d^{-1}	3.90 d^{-1}	3.91 d^{-1}
60	4.96 d^{-1}	4.92 d^{-1}	4.90 d^{-1}	4.92 d^{-1}	7.89 d^{-1}	7.82 d^{-1}	7.79 d^{-1}	7.82 d^{-1}
	Simple random sample approach/MSE							
1	3.77 d^{-2}	3.73 d^{-2}	3.63 d^{-2}	3.50 d^{-2}	2.44 d^{-2}	2.19 d^{-2}	2.48 d^{-2}	2.45 d^{-2}
5	3.82 d^{-2}	3.73 d^{-2}	3.82 d^{-2}	3.67 d^{-2}	2.35 d^{-2}	2.38 d^{-2}	2.34 d^{-2}	2.39 d^{-2}
10	3.61 d^{-2}	3.61 d^{-2}	3.60 d^{-2}	3.68 d^{-2}	2.17 d^{-2}	2.14 d^{-2}	2.21 d^{-2}	2.24 d^{-2}
30	2.98 d^{-2}	2.95 d^{-2}	3.03 d^{-2}	2.95 d^{-2}	1.48 d^{-2}	1.51 d^{-2}	1.47 d^{-2}	1.51 d^{-2}
60	1.93 d^{-2}	1.95 d^{-2}	1.97 d^{-2}	1.94 d^{-2}	4.90 d^{-3}	5.01 d^{-3}	4.97 d^{-3}	5.23 d^{-3}
	Binder's approach/MSE							
1	3.99 d^{-1}	6.63 d^{-1}	7.80 d^{-1}	7.63 d^{-1}	4.09 d^{-1}	6.20 d^{-1}	8.48 d^{-1}	8.64 d^{-1}
5	3.73 d^{-1}	6.12 d^{-1}	7.75 d^{-1}	7.77 d^{-1}	3.55 d^{-1}	6.39 d^{-1}	7.67 d^{-1}	8.20 d^{-1}
10	3.27 d^{-1}	5.65 d^{-1}	7.09 d^{-1}	7.36 d^{-1}	3.11 d^{-1}	5.45 d^{-1}	7.00 d^{-1}	7.47 d^{-1}
30	1.68 d^{-1}	3.16 d^{-1}	4.14 d^{-1}	4.18 d^{-1}	1.32 d^{-1}	2.84 d^{-1}	3.66 d^{-1}	4.05 d^{-1}
60	3.45 d^{-32}	3.12 d^{-32}	3.65 d^{-32}	3.70 d^{-32}	3.62 d^{-32}	2.38 d^{-32}	2.88 d^{-32}	2.98 d^{-32}
	Cohen et al.'s approximation/MSE							
1	9.77 d^{-1}	9.40 d^{-1}	9.16 d^{-1}	8.68 d^{-1}	9.97 d^{-1}	8.81 d^{-1}	9.96 d^{-1}	9.81 d^{-1}
5	9.14 d^{-1}	8.59 d^{-1}	8.89 d^{-1}	8.56 d^{-1}	9.17 d^{-1}	9.37 d^{-1}	9.31 d^{-1}	9.64 d^{-1}
10	7.82 d^{-1}	7.72 d^{-1}	7.69 d^{-1}	7.72 d^{-1}	7.98 d^{-1}	8.14 d^{-1}	8.54 d^{-1}	8.84 d^{-1}
30	3.65 d^{-1}	3.37 d^{-1}	3.38 d^{-1}	3.26 d^{-1}	3.29 d^{-1}	3.48 d^{-1}	3.44 d^{-1}	3.57 d^{-1}
60	3.60 d^{-31}	9.39 d^{-32}	6.47 d^{-32}	5.92 d^{-32}	5.47 d^{-30}	2.42 d^{-31}	9.88 d^{-32}	7.92 d^{-32}
	Jackknife/MSE							
1	9.89 d^{-1}	9.53 d^{-1}	9.30 d^{-1}	8.82 d^{-1}	1.00 d^{+0}	8.93 d^{-1}	1.01 d^{+0}	9.96 d^{-1}
5	1.11 d^{+0}	1.04 d^{+0}	1.08 d^{+0}	1.04 d^{+0}	1.09 d^{+0}	1.09 d^{+0}	1.07 d^{+0}	1.10 d^{+0}
10	1.22 d^{+0}	1.19 d^{+0}	1.20 d^{+0}	1.20 d^{+0}	1.20 d^{+0}	1.17 d^{+0}	1.20 d^{+0}	1.23 d^{+0}
30	2.18 d^{+0}	2.08 d^{+0}	2.08 d^{+0}	1.99 d^{+0}	2.22 d^{+0}	2.13 d^{+0}	1.98 d^{+0}	2.06 d^{+0}
60	1.67 d^{-5}	2.87 d^{-6}	6.70 d^{-7}	4.27 d^{-7}	1.13 d^{-5}	1.69 d^{-6}	4.98 d^{-7}	3.33 d^{-7}

Note. $d = 10$. MSE = mean-squared error.

(reading) and relatively many (mathematics) items per student, providing an interesting comparison. Because AM allows only analysis of a single-variable, a limited set of models was investigated. Specifically, intercept-only, gender, and school-reported race/ethnicity models were used. For this study, the reading subscales literary experience, information, and perform-a-task subscales were each analyzed as separate, univariate scales.

Table 6 shows that the means are similar between NAEP and AM, commensurate with previous findings (von Davier, 2003). Also, the approximation made by AM and the implementation thereof for the purpose of this paper are largely the same, suggesting that our implementation is consistent. With respect to Binder's (1983) method, the approximation of the Hessian is to some extent defensible for long-term trend, which has many items per student, but not defensible for reading, which has few items per student. Nevertheless, the approximation is relatively close to the jackknife method. Table 7 shows that the same is true for the gender model. Table 8 illustrates these findings further for a race/ethnicity model.

It should be noted that the comparison with the simple random sample estimator is somewhat misleading. The dominant part of this estimator is usually the sampling part, which is $\sigma^2\mathbf{\Xi}$. This part is overestimated because the model does not take into account hierarchical relations and, instead, attributes this variability to residual error. Subsequently, the standard error is inflated. In combination with the fixed effect model, it could be argued that this simple random sample estimator is not entirely inappropriate to account for the hierarchical relations in the sample. However, the major issue is that the assessment of the complexity of the sample through the residual variance is probably not very precise.

4.3 Impact on Plausible Values

The plausible value methodology was developed to provide secondary users with a complete dataset and a straightforward procedure to take variability due to the latency of proficiency into account in subsequent analyses. The following steps are executed:

1. A draw of a multivariate normal distribution with mean vector $\hat{\gamma}$ and covariance matrix $V(\hat{\gamma})$ is taken to yield a vector of provisional estimates $\hat{\gamma}_p$.

Table 6

Intercept-Only Model With 2005 Reading Strands, Grade 12, and 2004 Long-Term Trend Mathematics, Age 17, Means and Standard Errors Using Various Estimation Methods

	Method	Literary	Reading information	Task	LTT math
Means	NAEP	0.120	0.120	0.068	0.040
	AM	0.119	0.119	0.067	0.037
SE	Jackknife	0.022	0.018	0.021	0.039
	Binder	0.012	0.009	0.012	0.034
	Approx.	0.022	0.017	0.020	0.038
	AM	0.022	0.017	0.022	0.039
	SRS	0.012	0.011	0.012	0.012

Note. LTT = long term trend, SRS = simple random sample.

Table 7

Gender Model With 2005 Reading Strands, Grade 12, and 2004 Long-Term Trend Mathematics, Age 17, Means and Standard Errors Using Various Estimation Methods

		Male			Female			LTT	
	Method	Lit.	Info.	Task	Lit.	Info.	Task	Male	Female
Means	NAEP	-0.031	0.003	-0.148	0.258	0.229	0.266	0.097	-0.016
	AM	-0.032	0.001	-0.149	0.257	0.229	0.265	0.094	-0.019
SE	Jackknife	0.028	0.018	0.025	0.029	0.024	0.031	0.040	0.042
	Binder	0.014	0.010	0.014	0.016	0.013	0.017	0.035	0.037
	Approx.	0.026	0.017	0.024	0.030	0.025	0.031	0.036	0.044
	AM	0.026	0.017	0.025	0.030	0.025	0.032	0.037	0.045
	SRS	0.017	0.015	0.017	0.016	0.015	0.017	0.016	0.016

Note. Lit. = literary, LTT = long-term trend, info. = information, SRS = simple random sample.

Table 8

Race/Ethnicity Model With 2005 Reading Strands, Grade 12, and 2004 Long-Term Trend Mathematics, Age 17, Means and Standard Errors Using Various Estimation Methods

	Method	White	Black	Hispanic	Asian	Native	Other
				Literary			
Means	NAEP	0.280	-0.353	-0.241	0.120	0.110	-0.360
	AM	0.279	-0.357	-0.244	0.121	0.113	-0.381
SE	Jackknife	0.028	0.055	0.046	0.069	0.192	0.206
	Binder	0.015	0.026	0.022	0.036	0.100	0.098
	Approx.	0.027	0.056	0.045	0.065	0.228	0.295
	AM	0.027	0.056	0.045	0.066	0.230	0.296
	SRS	0.015	0.026	0.031	0.053	0.145	0.145
				Information			
Means	NAEP	0.263	-0.292	-0.209	0.179	-0.102	0.163
	AM	0.262	-0.293	-0.210	0.179	-0.106	0.167
SE	Jackknife	0.020	0.034	0.036	0.072	0.265	0.195
	Binder	0.010	0.017	0.018	0.037	0.116	0.098
	Approx.	0.019	0.037	0.038	0.065	0.221	0.189
	AM	0.019	0.037	0.039	0.066	0.223	0.190
	SRS	0.013	0.023	0.027	0.046	0.142	0.123
				Task			
Means	NAEP	0.176	-0.284	-0.094	0.039	-0.317	-0.066
	AM	0.175	-0.288	-0.095	0.038	-0.329	-0.069
SE	Jackknife	0.026	0.048	0.050	0.069	0.153	0.270
	Binder	0.014	0.027	0.028	0.038	0.085	0.143
	Approx.	0.025	0.049	0.050	0.070	0.258	0.204
	AM	0.027	0.049	0.050	0.070	0.260	0.205
	SRS	0.016	0.027	0.032	0.054	0.157	0.147
				Long-term trend			
Means	NAEP	0.226	-0.627	-0.395	0.389	-0.068	0.171
	AM	0.223	-0.635	-0.400	0.385	-0.071	0.176
SE	Jackknife	0.036	0.053	0.044	0.094	0.156	0.158
	Binder	0.031	0.045	0.037	0.075	0.114	0.135
	Approx.	0.035	0.055	0.042	0.076	0.106	0.201
	AM	0.036	0.056	0.042	0.076	0.107	0.203
	SRS	0.013	0.032	0.032	0.052	0.130	0.152

Note. Asian = Asian American, Native = Native American, SRS = simple random sample.

Table 9

Example of Posterior Means with $k \in [10, 20, 80]$ ***and Last Term Means***

$\mu \in [1, 2, 3, 4]$

μ	1	2	3	4
$k = 10$	$9.48 \cdot 10^{-3}$	$2.88 \cdot 10^{-2}$	$7.57 \cdot 10^{-2}$	$1.82 \cdot 10^{-1}$
$k = 20$	$6.45 \cdot 10^{-5}$	$2.00 \cdot 10^{-4}$	$5.51 \cdot 10^{-4}$	$1.50 \cdot 10^{-3}$
$k = 80$	$6.06 \cdot 10^{-18}$	$1.87 \cdot 10^{-17}$	$5.16 \cdot 10^{-17}$	$1.41 \cdot 10^{-16}$

2. Provisional moments $\tilde{\theta}_p$ and $V(\tilde{\theta})_p$ are obtained for each student using $\hat{\gamma}_p$ in (7) and (8).

3. A draw from a normal distribution with mean $\tilde{\theta}_p$ and variance $V(\tilde{\theta})_p$ is taken to obtain the p^{th} plausible value for each student.

4. Steps 1 through 3 are conducted P times to obtain P plausible values.

In (2), a product containing $k+1$ terms is used, where each of the terms is a probability distribution. The product of these distributions has a posterior mean, which is estimated by (7), and a dispersion, which is estimated by the posterior variance in (8). Suppose k distributions are approximately normal with a similar mean and similar, substantially large dispersion. Furthermore, assume that the $(k+1)^{th}$ distribution also has a substantially large dispersion. In that case, the larger n is, the less contribution the final term will make to the posterior. To illustrate this point, posterior means were computed for $k \in [10, 20, 80]$ and all k-distributions are standard normal. Furthermore, the last term was also normal with mean $\mu \in [1, 2, 3, 4]$ and unit standard deviation. Integration was conducted using a rectangular rule over 11 points in the interval of -5 to 5. The results in Table 9 show that with many items, the mean of the last term has very little influence on the posterior mean even if the last term is deviant from the mean of the first k terms.

To provide an answer to the second part of Question 2 from the first section of this report, several deductions can be made regarding the effect of the number of items and

model saturation on plausible values. As shown in Table 9, in an assessment with relatively many items per student, the posterior mean and variance are largely unaffected by the population model. Hence, the expectation of the results in Step 3 in the imputation process is similar regardless of the provisional values for $\hat{\gamma}$ obtained in Step 1. Therefore, although the result of most complex sample estimators is an increase in the standard error of the regression parameters, the effect of using these estimators on the distribution of plausible values will be small. If the number of items per student is small, these estimators will of course have more impact. Yet, full account of the clustering can be accomplished only if the posterior means are also drawn following a complex sample. In other words, for each cluster, multivariate draws would have to be taken, where off-diagonal elements of a distribution covariance matrix represent the intraclass correlation.

5 Discussion and Conclusion

In this study, the impact of several assessment designs on the success of estimating latent regression model parameters and their standard errors, taking into account the complex nature of the assessment, was evaluated. For these designs, the number of items, the number of predictors, and the rate of clustering were manipulated. In this study, a jackknife-based method was compared with Binder's (1983) method and an approximation thereof proposed by Cohen and Jiang (2002). The initial investigation of this approximation (Li & Oranje, 2007) showed several surprising results related to the number of predictors and the amount of information (i.e., number of items per student) used to define the latent part of the model. Simulation and real data analysis were used to determine to what extent these competing methods could be fruitful for designs typically found in NAEP. Conclusions and limitations of the study are presented below.

5.1 Conclusions

There are several preliminary conclusions that can be drawn, provided that the limited nature of the study is well understood.

- If a nearly saturated student grouping model is used, then essentially a fixed effect hierarchical model is estimated and methods such as Binder's (1983) will severely underestimate sampling variability. Estimation using a straightforward jackknife method has similar problems for nearly saturated models. This is associated with identification of the model parameters. However, in all other cases, this method provides a good approximation of the true variability. The results are somewhat conservative for larger models.

- The approximation to the Hessian matrix in Binder's (1983) method is only appropriate if the posterior variance is small. This is usually not the case for subjects such as main NAEP reading or mathematics, where the number of items is relatively small. Nevertheless, in practical terms this method provides reasonable results as long as the number of predictors is very small.

- A simple random sample formula for computing regression effect standard errors might, in practice, not be entirely inappropriate. Complex sample variability may to some extent be represented in an overestimate of the residual variance, which increases the standard error comensurately.

- In real data analysis applications, Cohen and Jiang's (2002) approximation seems to be relatively close to a jackknife calculation when the statistic of interest is based on a relatively large number of students. This is in line with the expectation of first-order approximations. For statistics based on small samples, this is much less the case.

- The concept of model saturation has surfaced several times and can be viewed in different ways. As described in the introduction, in operational NAEP a large number of variables is transformed into dummy variables and, subsequently, this matrix is reduced via principal component analysis. It can be said that this model is saturated with respect to the available student group information that was collected in the first place. However, it is uncertain to what extent this model is also saturated with respect to the clustering and the cognitive model as neither cluster variables nor direct

indicators of proficiency (e.g., raw scores, normit scores) are entered into the model. Hence, the degree of saturation with respect to proficiency and clustering is mostly limited to how well the student group information predicts proficiency and clustering. Most NAEP subjects have residual variance terms between 40% and 60% of the total variance, indicating a decidedly nonsaturated model at least with respect to proficiency.

- Individual proficiencies can be determined relative precisely if many items are assessed per student. Subsequently, the population part of the model does not provide a substantial contribution to the estimation of posterior distributions of proficiency and plausible values are relatively indifferent to the accuracy of regression effects and their standard errors.

5.2 Limitations of the Study

In NAEP, a balanced incomplete block design is employed to assess a broad framework represented by a large item pool while only a limited amount of time per student is demanded. Under this design, students receive a systematic portion of the assessment and are, as such, not directly comparable. However, as a group they receive the full range of content and through common-item IRT equating student proficiencies are defined in terms of a common latent scale. NAEP does not report individual proficiency data and only posterior distributions for student groups are calculated. The simulation study assumed that all students were responding to the same set of items.

As mentioned in the introduction, the number of covariates in the NAEP operational population models is far larger and far more complex than that used in the simulation study. Several hundred variables are collected and used in these latent regression models. The operational models are large enough to say that for each cluster on average a fixed effect is estimated through some linear combination of this large set of variables. However, this might not be uniformly the case across clusters and within a cluster there is substantial variability accounted for by the model (i.e., fixed-effect linear model). In the simulation study, all within-cluster variability was due to proficiency.

The complex-sample standard errors studied here require determination of a cluster

variable. For the real data analysis, pseudo-strata were used that serve as relatively large, but homogeneous sampling units. However, many other choices could have been made, such as schools, primary sampling units based on geographic location, or some combination of those.

In estimating group effects in the real data analysis, prior means are presented as essentially group effects, $\gamma' x$. However, this estimate is based on the probability of observing proficiency given the student group indicators. Alternatively, a Bayesian estimate can be used, which is based on the probability of observing proficiency given both the student group indicators and item responses. This point has also been made by Mazzeo et al. (2006, p. 22).

Finally, the jackknife in the real data analysis addressed only the sampling part of the variance. In operational NAEP, a component due to the latency of proficiency is also added. This component usually accounts for 5% to 10% of the variance unless the number of items assessed per student is very small. For relatively many items per student assessed, this term is negligible. That being said, in this study the complete model was subject to the jackknife. This is unlike operational procedures in the sense that the jackknife is applied to plausible values.

References

Allen, N., Donoghue, J., & Schoeps, T. *NAEP 1998 technical documentation*. Washington, DC: National Center of Education Statistics.

American Institutes for Research. (2003). AM beta 0.06.03 [software]. Washington, DC: Author.

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

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

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

Haberman, S. J. (1977a). Log-linear models and frequency tables with small expected cell counts. *The Annals of Statistics, 5*(6), 1148-1169.

Haberman, S. J. (1977b). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5*(5), 815-841.

Li, D., & Oranje, A. (2007). *Estimation of standard errors of regression effects in latent regression models* (ETS Research Rep. No. RR- 07-09). Princeton, NJ: ETS.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic Press.

Mazzeo, J., Donoghue, J. R., Li, D., & Johnson, M. S. (2006). *Marginal estimation in NAEP: current operational procedures and AM*. Paper prepared for National Center of Education Statistics.

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.

Perie, M., Grigg, W. S., & Donahue, P. L. (2005). *The nation's report card: Reading 2005*. Washington, DC: National Center for Education Statisics, Institute for Education Sciences.

Perie, M., & Moran, R. (2005). *NAEP 2004 trends: Three decades of student performance in reading and mathematics*. Washington, DC: National Center for Education Statistics, Institute for Education Sciences.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. consistency. *The Annals of Statistics, 12*(4), 1298-1309.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics, 16*(1), 356-366.

Qian, J. (2005). *Estimating variance in highly clustered survey samples.* Paper presented at the annual Joint Statistical Meetings, Minneapolis, MN.

Qian, J., & Haberman, S. (2006). *Improve variance estimation for the assessments based on the plausible values approach*. Paper presented at the annual Joint Statistical Meetings, Seattle, WA.

Rust, K., & Johnson, E. (1992). Sampling and weighting in the national assessment. *Journal of Education Statistics, 17*(2), 131-154.

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

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

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*(298), 348-368.

Appendix

Cochran's Formulae

While the following formulae were not used in the empirical study, they are presented here to illustrate differences in complexity between the approximations following the solutions proposed by Cohen et al. (2002). Cochran (1977) provides a host of formulae that are specifically geared to the sampling procudures of surveys such as NAEP as described in section 1.4. NAEP's statistics of interest are usually referred to as Horvitz-Thompson estimators (Qian, 2005) assuming the form

$$w_+\hat{\mu}_\theta = \sum_{i=1}^{N} \frac{\hat{\theta}_i}{\pi_i} \tag{A1}$$

for the sum of proficiencies where π_i is the probability that student i is selected in the sample, $\frac{1}{\pi_i} = w_i$, and $\hat{\theta}_i$ is the proficiency estimate for student i. The associated variance estimate (Cochran, 1977, Equation 9A.38) for single-stage sampling is:

$$V_I(w_+\hat{\mu}_\theta) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}\hat{\theta}_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\hat{\theta}_i\hat{\theta}_j, \tag{A2}$$

where π_{ij} is the probability that student i and j are both in the sample, which is also referred to as the inverse of the secondary order weight. This is currently not calculated in NAEP. Cochran also provides two-stage sampling formulas when sampling is commensurate the size of the sampling units. However, the units within a cluster are considered independent and unweighted. The formula, decomposing the variance into within and between cluster variance, is (Cochran, 1977, Equation 11.3)

$$V_{II}(\hat{\mu}_\theta) = \frac{1}{\nu}\left(\sum_{c=1}^{C} \frac{\nu_c - N_c}{N_c} s_c^2 + \sum_{c=1}^{C} \nu_c \left(\mu_{\theta_c} - \mu_\theta\right)\right), \tag{A3}$$

where $s_c^2 = \frac{1}{\nu_c - 1}\sum_{i=1}^{\nu_c}\left(\hat{\theta}_{c_i} - \mu_{\theta_c}\right)^2$, ν is the total number of units (e.g., students) in the population, ν_c the total number of units in cluster c in the population, N_c the number of units in the sample, μ_θ the average proficiency in the population and μ_{θ_c} the average proficiency in the population for cluster c. Cochran also combines (A2) and (A3) in Equation 11.14 for selection of units with equal size and there are several options to

obtain sample estimates for population quantities. This illustrates that an exact analytic formula is exceedingly complex, given that even these formulae do not take into account all characteristics of an unequal probability multi-stage design.