Research Report

A Bayesian Hierarchical Model for Large-Scale Educational Surveys: An Application to the National Assessment of Educational Progress

Matthew S. Johnson

Frank Jenkins

Research & Development

January 2005
RR-04-38

A Bayesian Hierarchical Model for Large-Scale Educational Surveys: An Application to the National Assessment of Educational Progress

Matthew S. Johnson
Baruch College, NY

Frank Jenkins
ETS, Princeton, NJ

January 2005

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:

www.ets.org/research/contact.html

Abstract

Large-scale educational assessments such as the National Assessment of Educational Progress (NAEP) sample examinees to whom an exam will be administered. In most situations the sampling design is not a simple random sample and must be accounted for in the estimating model. After reviewing the current operational estimation procedure for NAEP, this paper describes a Bayesian hierarchical model for the analysis of complex large-scale assessments. The model clusters students within schools and schools within primary sampling units. The paper discusses an estimation procedure that utilizes a Markov chain Monte Carlo algorithm to approximate the posterior distribution of the model parameters. Results from two Bayesian models, one treating item parameters as known and one treating them as unknown, are compared to results from the current operational method on a simulated data set and on a subset of data from the 1998 NAEP reading assessment. The point estimates from the Bayesian model and the operational method are quite similar in most cases, but there does seem to be systematic differences in measures of uncertainty (e.g., standard errors, confidence intervals).

Key words: Bayesian analysis, Markov chain Monte Carlo, item response theory (IRT), cluster sampling, large-scale assessments

1 Introduction

Large-scale assessments such as the National Assessment of Educational Progress (NAEP), the National Adult Literacy Survey (NALS), and the Trends in International Mathematics and Science Study (TIMSS) are developed to collect information on the knowledge and skills of the population at hand and report how that knowledge varies across different groups in the population. If knowledge could be observed, then standard survey methodology could be used to analyze the data. However, large-scale assessments, in addition to encountering the difficulties of any large-scale survey, must deal with the fact that knowledge is usually not something that should be considered directly observable.

In an attempt to *measure* the knowledge or skill of interest, large-scale assessments administer a series of test questions that require that skill or knowledge. One of the major difficulties associated with these assessments is trying to extract information about the underlying knowledge from the responses of the survey respondents, or examinees. Early methods for the analysis of assessment data would simply treat the number of items correct or the score on the test as a normal random variable. Modern methods use item response theory (IRT) models (cf. Lord, 1980; Linden & Hambleton, 1997).

Item response models are a class of generalized mixed effects models that describe the dependence structure inherent in the item responses by incorporating a latent variable, denoted θ, into the model. The primary assumptions of item response models are *unidimensionality* (the latent variable θ is unidimensional), *conditional independence* (conditional on θ, the item responses of a single examinee are independent), and *monotonicity* (the probability of receiving a high score on any item is an increasing function of θ). If these assumptions hold, it is at least reasonable to treat θ as equivalent to the skill of interest. This paper uses the term *proficiency* when referring to the latent variable θ.

(Scott & Ip, 2002) proposed a Bayesian hierarchical model for multidimensional assessments; however, their model did not account for the complex clustered structure inherent in the sampling design. This paper proposes a Bayesian hierarchical model that does account for the sampling design. The model expands the superpopulation models (Potthoff, Woodbury, & Manton, 1992; Cassel, Särndal, & Wretman, 1993) proposed by (Longford, 1995) and (Raudenbush, Fotiu, & Fai, 1999), which assumed students' proficiencies were known. Our model extends these superpopulation models by incorporating the scaling level, which relates the observable item

responses to the unobservable student proficiency. The parameters of the model are estimated using Markov chain Monte Carlo (MCMC; Geman & Geman, 1984; Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996). Our approach is an alternative to the approach that has been implemented in some form since 1984 for the official analysis of NAEP data.

The remainder of the paper is organized as follows: Section 2 summarizes the *reading to perform a task scale* from the NAEP 1998 reading assessment at grade 8; section 3 summarizes the general methodology used in large-scale assessments; section 4 reviews the current operational methodology for the analysis of NAEP data; section 5 introduces the Bayesian hierarchical model and discusses the MCMC algorithm for estimation; section 6 analyzes a simulated data set to study the ability of our model to recover model parameters, and compares the estimates for the NAEP 1998 reading data from our model to those found using the operational methodology; and section 7 discusses the conclusions.

2 NAEP Data

NAEP is a survey of the academic achievement of students in the United States. NAEP is administered by the National Center for Educational Statistics, a division of the U.S. Department of Education, to a sample of students in 4th, 8th, and 12th grades. The results are reported for a number of academic subjects on a regular basis in a document called *The Nation's Report Card*.

The assumption central to the NAEP methodology is that each student in the population has some unobservable underlying proficiency for each subject area, and that the items administered on the test are discrete measures of this latent proficiency. Let θ_i denote the latent proficiency of student i. One of the major goals of NAEP is to summarize the distribution of abilities (θ) for the entire population of students in the United States and for various subgroups within that population (e.g., female students whose parents finished college).

NAEP presents results for each assessment and compares these results to previous assessments in the subject area. The academic achievement is described in terms of the average student proficiency for all students in the United States and in terms of the percentage of students attaining fixed levels of proficiency. These *achievement levels*, defined by the National Assessment Government Board, are judgments of what students should know at each grade level. In addition to producing these numbers for the nation as a whole, NAEP reports the same results for

subgroups in the population (e.g., racial groups, male, female).

Because the goal of NAEP is to measure the academic achievement of the United States as a whole rather than the achievement of individual students, it is not necessary that each student receives a large number of items. To reduce the number of items a given examinee is presented, NAEP uses a matrix sampling of items. Items are split into several blocks of items, and each examinee is presented with three blocks.

The NAEP 1998 in reading assessment at grade 8 contained 110 items split into 10 blocks; each item was developed to measure one dimension of a three-dimensional reading proficiency. Forty-eight of these items were developed to measure *reading to gain information*, 33 items were developed to measure *reading to perform a task*, and 29 items were developed to measure *reading for literary experience*. This paper performs a unidimensional analysis of the *reading to perform a task* scale.

There are two types of cognitive items used in NAEP assessments— constructed-response and multiple-choice items. For reading, constructed-response items are designed to provide an in-depth view of students' proficiency to read thoughtfully and to respond appropriately to what was read. Constructed responses can be as short as a sentence or two for simple questions, or they can be used for items where more thoughtful consideration is required. Multiple-choice items are used whenever reading comprehension can be measured using such items. For the NAEP 1998 grade 8 *reading to perform a task*, 11 items were multiple choice, and 22 items were constructed response.

As part of the NAEP assessments, a series of questionnaires was also administered to the students, teachers, and school administrators. Students were asked to provide the answers to three sets of multiple-choice background questions: general background (e.g., gender, parents' educations, race, etc.), reading background, and motivation. In total there were 46 questions on the students' reading background questionnaire. The teacher and administrator questionnaires were used to supplement the students' responses. This paper focuses on three background questions from the students' general background set.

Eleven-thousand-fifty-one (11,051) eighth grade students participated in the NAEP 1998 reading assessment. The students were sampled according to a three-stage sampling design. The primary sampling units (PSU) were made up of cities, counties, or groups of counties. In the second stage, schools were sampled from within each PSU. Finally, students were sampled within each school. The three stages are summarized below; for a more detailed discussion of the

procedure, see *The NAEP 1998 Technical Report* (Allen, Donoghue, & Schoeps, 2001).

Sampling of Geographic Areas

The basic PSU sampling design is a stratified probability sample with a single PSU drawn in each stratum. A single PSU consists of a consolidated metropolitan statistical area (CMSA), a metropolitan statistical area (MSA), a New England county metropolitan area (NECMA), a county, or group of contiguous counties in the United States. The PSU sampling frame follows the rules below:

- Each CMSA and MSA not contained in a CMSA is a separate PSU. In New England, NECMAs were the metropolitan PSUs. We use the term MSA for all three metropolitan area PSUs (MSA, NECMA, CMSA).
- Non-MSA PSUs consist of non-MSA counties. In most cases, non-MSA PSUs contain only contiguous counties with a minimum 1990 population of 60,000 people in the northeastern and southwestern regions of the country and 45,000 in the other regions of the country.

The sampling frame consists of 1,027 PSUs; 290 of the PSUs are MSAs.

The largest 22 PSUs were included in the sample with certainty. The remaining PSUs were first stratified according to region of the country and MSA status of the PSU. Crossing the region of the country (northeast, southeast, central, west) with MSA status (MSA, non-MSA) defined the 8 major strata. Within each of the major stratum, further stratification was achieved by ordering PSUs on additional socioeconomic characteristics, yielding 72 strata with approximately equal populations according to the 1990 U.S. Census. Within each of the 72 strata, a single PSU was drawn with probability proportional to the population of the PSU. Combining the 22 certainty PSUs with the 72 noncertainty PSUs gave a total of 94 from which schools were sampled.

Sampling of Schools

Schools were sampled within PSUs with a sampling probability proportional to a measure of school size with two adjustments. In order to increase the precision of estimates for the group means of minority and private school students, schools designated high minority (HM; the Black and Hispanic students account for more than 10% of the total student population) and private schools were oversampled. The sampling probability for high minority schools was double that of

nonhigh minority schools; and the sampling probability was tripled for private schools. That is,

$$Pr\{\text{School } s \text{ is in the sample}\} \propto \begin{cases} K_s & \text{if school } s \text{ is public, non-HM} \\ 2K_s & \text{if school } s \text{ is public, HM} \\ 3K_s & \text{if school } s \text{ is private} \end{cases}$$

where K_s is a measure of the size of school s. A total of 483 schools participated in the NAEP 1998 reading assessment at grade 8.

Sampling of Students

Students were sampled randomly from a list of all enrolled students in the specified grade. In public schools with low minority enrollment, Black and Hispanic students were oversampled. This was accomplished by creating a list of all of the nonselected Black and Hispanic students and sampling from it the same number of Black and Hispanic students as in the original student sample from the school. A similar method was used to oversample students with disabilities and students of limited English proficiency.

To account for the differential probabilities of selection and to allow for adjustments for nonresponse, each student was assigned a sampling weight. Let w_i denote the sampling weight for examinee i. The interested reader can find more information on the derivation of the sampling weights in (Allen et al., 2001).

3 The General Methodology

One of the goals of NAEP is to report how students in the United States are performing in any given subject. NAEP utilizes the cognitive responses and background information of each examinee to report various characteristics of the entire population and for numerous subgroups of this population (e.g., female students whose parents graduated from college) in terms of the unobservable proficiency or proficiency θ.

This paper examines two types of summaries commonly examined in NAEP: (a) group means and (b) proportion of students within achievement levels for each group. Both summaries can be treated as the average value of some function of the latent proficiencies $[g(\theta)]$ for individuals in a given group. Let $\mathcal{G}$ denote such a group. For group means $g(\theta) = \theta$, and for achievement level

results

$$g(\theta) = I_{(a,b)}(\theta) = \begin{cases} 1 & \theta \in (a,b) \\ 0 & \theta \notin (a,b) \end{cases}$$

where a and b are cutoffs defining regions of the ability scale.

There are two ways to think about the quantity of interest. The first method is to treat the problem as a finite population problem, in which case the quantity we wish to estimate for group $\mathcal{G}$ is

$$\overline{g}_{\mathcal{G}} = \frac{1}{N_{\mathcal{G}}} \sum_{i \in \mathcal{G}} g(\theta_i), \tag{1}$$

where $N_{\mathcal{G}}$ is the number of individuals in group $\mathcal{G}$, a subset of the entire population Ω. The second method is to treat the problem as a superpopulation problem where the goal is to estimate:

$$E[g(\theta) \mid \mathcal{G}] = \int_t g(t) dF(t \mid \mathcal{G}),$$

where $F(\theta \mid \mathcal{G})$ is the distribution of proficiencies in subgroup $\mathcal{G}$. The estimation of either quantity would be *relatively* straightforward if the latent proficiencies were observed for each examinee in the sample. However, we only observe the repeated discrete measures of the proficiency from the test items and the background data.

The NAEP methodology applies item response models (Linden & Hambleton, 1997) to recover information about the latent proficiency (θ) from these multiple item responses. Item response models are a class of generalized mixed effects models that account for the dependency among the item responses of a single individual by modeling the probability of the responses as a function of the latent variable θ. The basic assumption in item response models is the conditional independence assumption; examinee's responses $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})^t$ to items $j = 1, \ldots, J$ are independent conditional on the examinee's proficiency θ_i.

The NAEP methodology models multiple-choice items with the three-parameter logistic (3PL) model (Lord, 1980). The 3PL defines the probability of a correct response with the following function of examinee's unobservable proficiency θ_i:

$$Pr\{X_{ij} = 1 \mid \theta_i, \boldsymbol{\psi}_j\} \equiv P_j(\theta_i) = \pi_j + (1 - \pi_j) R_j(\theta_i) \tag{2}$$

where

$$R_j(\theta_i) = \frac{1}{1 + \exp\{\alpha_j(\beta_j - \theta_i)\}}, \tag{3}$$

α_j is the *discrimination* of item j, β_j is the *difficulty* of item j, and π_j is the *asymptote* of item j, and $\boldsymbol{\psi}_j = (\alpha_j, \beta_j, \pi_j)^t$.

In addition to multiple-choice items, NAEP assessments contain a number of constructed-response items, which can be dichotomous (correct/incorrect) or polytomous (degrees of correctness). Let $X_{ij} \in \{0, \ldots, M_j\}$ be the response of examinee i to item j where $X_{ij} = 0$ indicates poor performance and $X_{ij} = M_j$ indicates excellent performance of examinee i on item j. For the NAEP 1998 grade 8 *reading to perform a task scale* $M_j \in \{1, \ldots, 5\}$. NAEP posits the generalized partial credit model (GPCM; Muraki, 1992) for constructed-response items. The GPCM defines the response function by assuming the adjacent category logits are linear in the latent proficiency θ_i:

$$\text{logit}\left(Pr\{X_{ij} = k \mid X_{ij} \in \{k, k-1\}, \theta_i, \boldsymbol{\psi}_j\}\right) = \alpha_j(\theta_i - \beta_j + \delta_{jk}) \tag{4}$$

for $k = 1, \ldots, M_j$, where the *item-category* parameters $\boldsymbol{\delta}_j$ satisfy $\sum_{k=1}^{M_j} \delta_{jk} = 0$ for all j. Notice when $M_j = 1$, the GPCM in (4) is equivalent to $R_j(\theta)$ in (3). This typically is called the two-parameter logistic (2PL) model (Birnbaum, 1968).

The information contained in the cognitive responses is combined with background information to estimate the quantities of interest, namely subgroup means and subgroup achievement level results. NAEP samples students from a complex, two-stage clustered sample, so care must be taken when calculating the variances of these estimators.

Section 4 provides a review of the current three-stage operational procedures used for the analysis of NAEP data. Section 5 formulates the problem as a hierarchical Bayesian model and suggest an MCMC procedure for the estimation of the quantities of interest.

4 Operational Procedure for the Analysis of NAEP

The current operational method for the analysis of NAEP data is a three-stage procedure. The first stage called *scaling* uses item responses to estimate parameters for each item in the NAEP assessment. The *conditioning* stage uses the estimates in conjunction with item responses and background characteristics to estimate the effects of background variables and to impute the value of the latent variable θ for each student in the assessment. Section 4.2 describes these imputed or *plausible values* (PV; Mislevy, 1991) in more detail. The proper estimation of sampling variability for the statistics reported in NAEP is a complicated task. For this reason NAEP uses a

jackknife estimate of the standard errors for the reported statistics. Each of the three estimation procedures are summarized below. For more detailed information about the current operational procedures see *The NAEP 1998 Technical Report* (Allen et al., 2001) or the special issue of the *Journal of Educational Statistics* on NAEP (Zwick, 1992).

Although this section discusses the estimation procedure in terms of the NAEP analysis, similar methodologies have been employed for the analysis of the TIMSS and the NALS.

Scaling

The first step in the operational analysis of NAEP data is to estimate the item parameters of the item response models in (2) and (4). Under the assumption of conditional independence of the item responses given the latent proficiency score θ, the joint probability of the item response vector $\boldsymbol{x}_i$ conditional on θ_i is

$$L_i(\theta \mid \boldsymbol{x}_i, \boldsymbol{\psi}) = Pr\{\boldsymbol{x}_i \mid \theta_i, \boldsymbol{\psi}\} = \prod_{j=1}^{J} Pr\{X_{ij} = x_{ij} \mid \theta_i, \boldsymbol{\psi}_j\}. \tag{5}$$

The joint maximum likelihood (JML) estimates of the item parameters ($\boldsymbol{\psi}$) and examinee abilities are found by maximizing $L(\boldsymbol{\psi}, \boldsymbol{\theta}; \boldsymbol{X}) = \prod_i L_i(\theta \mid \boldsymbol{x}_i, \boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ simultaneously. However, as (Andersen, 1970) noted, JML estimates are inconsistent. The NAEP operational analysis estimates item parameters with the marginal likelihood found by integrating the joint likelihood function over a distribution of student abilities, denoted F_Θ. That is

$$L(\boldsymbol{\psi};\ \boldsymbol{X}) = \int_\Theta L(\boldsymbol{\psi}, \boldsymbol{\theta};\ \boldsymbol{X}) dF_\Theta(\theta), \tag{6}$$

is maximized with respect to the item parameters $\boldsymbol{\psi}$.

The NAEP analysis assumes examinee proficiencies θ_i are independently and identically distributed and that the distribution F_Θ is a discrete distribution on 41 equally spaced points from -4 to 4 with unknown mass. That is, the distribution for θ is

$$f_\Theta(t) = \begin{cases} p_t & \text{if } t \in \{-4, -3.8, \ldots, 4\} \\ 0 & \text{otherwise} \end{cases}$$

where $\sum_t p_t = 1$. For this distribution of proficiencies the marginal likelihood in (6) becomes

$$L(\boldsymbol{\psi};\ \boldsymbol{X}) = \prod_{i=1}^{N} \sum_{t \in \{-4, \ldots, 4\}} Pr\{\boldsymbol{x}_i \mid t, \boldsymbol{\psi}\} p_t.$$

The NAEP BILOG/PARSCALE program estimates item parameter vectors $\boldsymbol{\psi}_j$ and the probability masses of the proficiency distribution p_t. The interested reader is directed to (Mislevy & Bock, 1982) and (Muraki & Bock, 1997) for more information on the BILOG/PARSCALE software.

Conditioning

The conditioning stage uses background information from examinees, teachers, schools, and PSUs to estimate the effects of these background variables on the student proficiency θ. In the conditioning stage of estimation, the operational analysis fixes item parameter vectors $\boldsymbol{\psi}_j$ at the estimates $\hat{\boldsymbol{\psi}}_j$ from the scaling stage, thus defining the likelihood for $\boldsymbol{\theta}$, $L(\boldsymbol{\theta};\ \boldsymbol{X}) = \prod_i L(\theta_i;\ \boldsymbol{\psi}_j = \hat{\boldsymbol{\psi}}_j, \boldsymbol{x}_i)$.

Unlike the proficiency distribution used in the scaling stage, the prior in the conditioning stage is a structured distribution. In fact, the distribution for θ used in the conditioning stage assumes proficiencies follow a normal linear model. That is,

$$\theta_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\gamma}'\boldsymbol{y}_i, \sigma^2) \quad i = 1, \ldots, n, \tag{7}$$

where $\boldsymbol{y}_i$ is the vector of background variables for examinee i, $\boldsymbol{\gamma}$ is the vector of regression effects, and σ^2 is the variance.

The marginal likelihood of $(\boldsymbol{\gamma}, \sigma^2)$ given the item response matrix $\boldsymbol{X}$ and the item parameters $\boldsymbol{\psi}$ is

$$L(\boldsymbol{\gamma}, \sigma^2; \boldsymbol{X}, \hat{\boldsymbol{\psi}}) = \sigma^{-2n} \prod_{i=1}^{n} \int_t L(t;\ \boldsymbol{x}_i)\phi\left(\frac{t - \boldsymbol{\gamma}'\boldsymbol{y}_i}{\sigma}\right) dt, \tag{8}$$

where $\phi(\cdot)$ is the standard normal density function. The operational analysis of NAEP data obtains maximum likelihood estimates of $\boldsymbol{\gamma}$ and σ^2, denoted $\hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}^2$, through the use of an estimation-maximazation (EM) algorithm (Dempster, Laird, & Rubin, 1977; Thomas, 1993), where student proficiencies are deemed nuisance parameters. The EM algorithm requires evaluation of the mean $\overline{\theta}_i$ and variance $\overline{\sigma}_i^2$ of the conditional posterior distribution

$$p(\theta_i \mid \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\gamma}, \sigma^2) \propto Pr\{\boldsymbol{x}_i \mid \theta_i, \hat{\boldsymbol{\psi}}\}\phi\left(\frac{\theta_i - \boldsymbol{\gamma}'\boldsymbol{y}_i}{\sigma}\right). \tag{9}$$

The integrals required to calculate the means $(\overline{\theta}_i)$ and variances $(\overline{\sigma}_i^2)$ of the conditional posterior distribution in (9) are intractable. The operational analysis approximates the means and variances by approximating the distribution with a discrete distribution over Q quadrature points

$(t_1,\ t_2, \ldots,\ t_Q)$ each with mass:

$$\hat{p}(t_k \mid \boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\gamma}, \sigma^2) = \frac{Pr\{\boldsymbol{x}_i \mid t_k, \hat{\boldsymbol{\psi}}\}\phi\left(\frac{t_k - \boldsymbol{\gamma}'\boldsymbol{y}_i}{\sigma}\right)}{\sum_q Pr\{\boldsymbol{x}_i \mid t_q, \hat{\boldsymbol{\psi}}\}\phi\left(\frac{t_q - \boldsymbol{\gamma}'\boldsymbol{y}_i}{\sigma}\right)}$$

Estimating group results. After completion of the EM algorithm, plausible values for each examinee's proficiency $[\theta_i]$ are drawn from an approximation of the conditional posterior distribution in (9) using the following algorithm:

1. Draw $\boldsymbol{\gamma}^{(m)} \sim \mathcal{N}(\hat{\boldsymbol{\gamma}}, V(\hat{\boldsymbol{\gamma}}))$, where $V(\hat{\boldsymbol{\gamma}})$ is the estimated variance of the maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$.

2. Conditional on the generated value $\boldsymbol{\gamma}^{(m)}$ and the fixed variance $\sigma^2 = \hat{\sigma}^2$, calculate the mean $\overline{\theta}_i$ and variance $\overline{\sigma}_i^2$ of the conditional posterior in (9) for each examinee i in the sample.

3. Draw a single $\theta_i \sim \mathcal{N}(\overline{\theta}_i, \overline{\sigma}_i^2)$ independently for each $i = 1, \ldots, n$.

These three steps are repeated M times producing M sets of imputed values $(\theta_i^{(1)}, \ldots, \theta_i^{(M)})$ for each examinee in the sample. Imputed values are used to approximate posterior expected values.

As noted earlier, one of the goals of NAEP is to estimate the performance of subgroups by examining the average value of some function $g(\cdot)$ over all individuals in that group. To estimate the average value of g for all individuals in group $\mathcal{G}$, the NAEP operational analysis calculates the posterior expected value of

$$\hat{g}_{\mathcal{G}}(\boldsymbol{\theta}) = \frac{1}{n_{\mathcal{G}}} \sum_{i \in \mathcal{G}_{\mathcal{S}}} w_i g(\theta_i), \tag{10}$$

where $n_{\mathcal{G}} = \sum_{i \in \mathcal{G}_{\mathcal{S}}} w_i$ and $\mathcal{G}_{\mathcal{S}}$ denotes the members of group $\mathcal{G}$ that are in the sample $\mathcal{S}$. To find the posterior expected value of $\hat{g}_{\mathcal{G}}(\boldsymbol{\theta})$, the NAEP methodology first approximates the conditional posterior expectation $E\left[\hat{g}_{\mathcal{G}}(\theta) \mid \boldsymbol{\gamma}^{(m)}\right]$ with $\hat{g}_{\mathcal{G}}(\boldsymbol{\theta}^{(m)})$, and then averages the $\hat{g}_{\mathcal{G}}(\boldsymbol{\theta}^{(m)})$s over the M plausible values. Let $\tilde{g}_{\mathcal{G}}$ denote this estimate of the expected value of g for individuals in group $\mathcal{G}$. That is, let

$$\tilde{g}_{\mathcal{G}} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{n_{\mathcal{G}}} \sum_{i \in \mathcal{G}_{\mathcal{S}}} w_i g(\theta_i^{(m)}).$$

Variance Calculation

The third step in the three-step NAEP operational analysis is to find the variance of the group estimators $\tilde{g}_{\mathcal{G}}$. There are two sources of uncertainty associated with the estimates. First,

NAEP only samples a small portion of the entire population of students. Second, the true values of the sampled students' proficiencies [θ] are not directly observed; examinees simply provide multiple discrete measures of their proficiency through the assessment items. Following (Rubin, 1987), the operational analysis applies a conditioning argument to split the total variation into the two additive pieces: the variation due to sampling, and the variance due to latency of proficiencies. NAEP uses the multiple imputation or plausible value methodology (Rubin, 1987) to estimate the variance due to latency and a jackknife estimate of the sampling variance.

Variance due to latency of θ. The variance due to latency of θ, or measurement error of θ, is the variance of the latent proficiency given the item responses. If the abilities θ were observed directly, then the measurement error would be zero. However, because we observe only a finite number of item responses for each examinee, the measurement error will be nonzero.

The operational analysis estimates the variance of $\tilde{g}_{\mathcal{G}}$ due to latency of θ for subgroup $\mathcal{G}$, denoted $U_{\mathcal{G}}$, by calculating the variance of $\hat{g}_{\mathcal{G}}^{(m)} = \hat{g}(\boldsymbol{\theta}^{(m)})$ over plausible values $m = 1, \ldots, M$. That is

$$U_{\mathcal{G}} = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{g}_{\mathcal{G}}^{(m)} - \tilde{g}_{\mathcal{G}})^2.$$

Variance due to sampling of students. The variance due to sampling of students is the sampling variance that would be observed if all proficiencies θ for students in the sample were actually observed. The NAEP estimation procedure uses a jackknife estimator to estimate the variance due to sampling.

Thirty-six PSU pairs are formed from the 72 noncertainty PSUs where the pairs are composed of PSUs from adjacent strata (thus PSU pairs were relatively similar in socioeconomic makeup). The schools from the 22 certainty PSUs are split into 26 groups based on information about the schools. These 26 groups are then split into two pseudo-PSUs. Combining the 36 noncertainty PSU pairs with the 26 pairs of pseudo-PSUs formed from the certainty PSUs leads to 62 PSU groups.

In turn, a single PSU is dropped from one of the 62 groups and the estimated mean $\tilde{g}_{\mathcal{G}}$ is recalculated with correctly adjusted weights. Let $\tilde{g}_{\mathcal{G}}^{(p)}$ denote the estimate for replicate group p. The NAEP jackknife estimate of sampling variance, denoted $V_{\mathcal{G}}$ is defined

$$V_{\mathcal{G}} = \sum_{p=1}^{62} (\hat{g}_{\mathcal{G}} - \hat{g}_{\mathcal{G}}^{(p)})^2.$$

In practice, the NAEP operational analyses use only the first plausible value to calculate the group estimate $\hat{g}_{\mathcal{G}}^{(p)}$ for replicate set p and as we note in the analysis sections this may bias the estimates of the standard errors, especially those for achievement level results.

Alternative approaches to estimating the sampling variance of the group statistics have been explored by (Longford, 1995) and (Raudenbush et al., 1999). In both cases the authors use superpopulation methods (Cassel et al., 1993; Potthoff et al., 1992) to derive estimates of sampling variance; (Longford, 1995) employs a classical approach and (Raudenbush et al., 1999) applies MCMC techniques (Gilks et al., 1996) to perform a Bayesian analysis. However, like the current NAEP estimation procedures, these methods are based on the plausible values, which assume the items' parameters are known and fixed at their estimates from scaling. The next section proposes a Bayesian hierarchical model that combines the scaling model with a superpopulation model for student proficiencies, which incorporates the background information and hence does not assume the item parameters are fixed and known.

5 Bayesian Hierarchical Model for Large-Scale Assessment

Model Formulation

For NAEP, we have a data set $\{X_{ij} : \ i = 1, \ldots, n; \ j = 1, \ldots, J\}$ consisting of the scores of n individuals to J test items (some responses will be missing since each student only receives a subset of the items).

We model constructed-response (CR) items with the GPCM (Muraki, 1992) defined in (4), and model multiple-choice (MC) items with the 3PL model (Lord, 1980) defined in (2).

$$[X_{ij} \mid \theta_i, \beta_j, \alpha_j, \pi_j, \boldsymbol{\delta}_j] \sim \begin{cases} \text{GPCM}(\theta_i, \beta_j, \alpha_j, \boldsymbol{\delta}_j) & \text{if item } j \text{ is CR} \\ \text{3PL}(\theta_i, \beta_j, \alpha_j, \pi_j) & \text{if item } j \text{ is MC} \end{cases} \tag{11}$$

Because this is a fully Bayesian model, we define prior distributions for all model parameters. The prior for the guessing parameter is $\pi_j \sim \text{Beta}(a_\pi, b_\pi)$. We set values for a_π and b_π based on the number of choices available (e.g., $a_\pi = 10$ and $b_\pi = 40$ might be used for a five-choice item). For the difficulty β_j and discrimination α_j parameters, we use uninformative prior distributions; the prior distribution for the discrimination parameter places all of its mass above zero.

We model the student ability distribution with a model similar to that used in the conditioning stage of the current NAEP estimation strategy. Like the distribution assumed in the conditioning

stage, the mean is a linear function of background variables $\boldsymbol{y}_i$. However, to properly account for the clustered nature of NAEP data, the model uses a linear mixed effects model (Laird & Ware, 1982) to cluster examinees by the school they attend (examinee i attends school $s(i)$) and to cluster schools by the PSU they belong to (school s belongs to PSU $p(s)$). This is similar to the treatment proposed by (Longford, 1995) and (Raudenbush et al., 1999) for the analysis of known student proficiencies,

$$\left.\begin{array}{rcl} \left[\theta_i \mid \nu_{s(i)}, \boldsymbol{\gamma}, \boldsymbol{y}_i, \sigma\right] & \sim & \mathcal{N}\left(\nu_{s(i)} + \boldsymbol{\gamma}'\boldsymbol{y}_i, \sigma_{s(i)}\right) \\ \left[\nu_s \mid \tau\right] & \sim & \mathcal{N}(\eta_{p(s)}, \tau^2) \\ \left[\eta_p \mid \omega\right] & \sim & \mathcal{N}(0, \omega^2). \end{array}\right\} \tag{12}$$

Assume the within-school variances σ_s^2 are constant across schools (i.e., $\sigma_s = \sigma$ for $s = 1, \ldots, S$). The between-school variance τ^2 measures the variance among school-intercepts ν_s, and the between-PSU variance ω^2 measures the variance among PSU-intercepts η_p. By integrating the school-intercepts out of the conditional distribution of θ, we find

$$\Sigma_{jk} = \mathrm{Cov}(\theta_j, \theta_k) = \begin{cases} Var(\theta_j) = \omega^2 + \tau^2 + \sigma^2 & j = k \\ \omega^2 + \tau^2 & s(j) = s(k),\ j \neq k \\ \omega^2 & p(s(j)) = p(s(k)),\ s(j) \neq s(k) \\ 0 & p(s(j)) \neq p(s(k)). \end{cases} \tag{13}$$

So students in the same school and in the same PSU are correlated with one another. We use inverse-gamma distributions for the prior distributions of σ^2, τ^2, and ω^2. Let (a_σ, λ), (a_τ, b_τ), and (a_ω, b_ω) denote the hyperparameters of these prior distributions respectively. All hyperparameters except λ are fixed constants and chosen to reflect any prior information that may be available about the particular variance component.

Markov Chain Monte Carlo Estimation

Given the item response matrix $\boldsymbol{X}$, estimation of the posterior distributions for the item response model parameters in (11) using MCMC methods is a straightforward exercise (Patz & Junker, 1999). We extend the MCMC estimation procedure to include a superpopulation, mixed effects model to describe the distribution of latent proficiencies θ. The layer of the hierarchical model that relates the proficiencies to background variables is similar to the models proposed by (Longford, 1995) and (Raudenbush et al., 1999) when student proficiencies are fixed and known.

The complete conditional distribution of the background effects $\boldsymbol{\gamma}$, the school intercepts ν, and the PSU-intercepts η are conditionally independent of the item responses and the item parameters given the student proficiencies θ and the variance components σ^2, τ^2, and ω^2. To improve the mixing of the MCMC algorithm, we choose to draw school- and PSU-intercepts and background effects $\boldsymbol{\gamma}$ from the joint conditional posterior distribution

$$f_{\boldsymbol{\gamma},\boldsymbol{\nu},\boldsymbol{\eta}}(\boldsymbol{\gamma},\boldsymbol{\nu},\boldsymbol{\eta} \mid \boldsymbol{\theta},\sigma,\tau,\omega,\boldsymbol{Y}) = f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma} \mid \boldsymbol{\theta},\boldsymbol{Y},\sigma,\tau,\omega) f_{\eta}(\boldsymbol{\eta} \mid \boldsymbol{\gamma},\boldsymbol{\theta},\sigma,\tau,\omega) f_{\nu}(\boldsymbol{\nu} \mid \boldsymbol{\gamma},\boldsymbol{\eta},\boldsymbol{\theta},\sigma,\tau)$$

by first drawing $\boldsymbol{\gamma}$ conditional on the variance-covariance matrix defined in (13) and the student proficiency vector $\boldsymbol{\theta}$

$$[\boldsymbol{\gamma} \mid \boldsymbol{\theta},\boldsymbol{Y},\sigma,\tau,\omega] \sim \mathcal{N}\left((\boldsymbol{Y}^t\Sigma^{-1}\boldsymbol{Y})^{-1}\boldsymbol{Y}^t\Sigma^{-1}\boldsymbol{\theta},(\boldsymbol{Y}^t\Sigma^{-1}\boldsymbol{Y})^{-1}\right).$$

The algorithm then draws the PSU-intercepts $\boldsymbol{\eta}$ from conditional posterior distribution given the vector of background effects $\boldsymbol{\gamma}$

$$[\eta_p \mid \boldsymbol{\gamma},\boldsymbol{\theta},\sigma,\tau,\omega] \sim \mathcal{N}\left(V\sum_s \frac{\overline{r}_s}{\sigma^2+n_s\tau^2}, V\right),$$

where $\overline{r}_s = \frac{1}{n_s}\sum_{i\in s}\{\theta_i - \boldsymbol{\gamma}'\boldsymbol{y}_i\}$ is the average residual for school s, and $V = \left(\frac{1}{\omega^2} + \sum_s \frac{n_s}{\sigma^2+n_s\tau^2}\right)^{-1}$. Conditional on the draws of the background effects $\boldsymbol{\gamma}$, the PSU-intercepts $\boldsymbol{\eta}$ and the variance components, we draw the school-intercepts from the complete conditional posterior distribution

$$[\nu_s \mid \boldsymbol{\gamma},\boldsymbol{\eta},\tau,\boldsymbol{\sigma}] \sim \mathcal{N}\left(\frac{n_s\tau^2\overline{r}_s + \sigma_s^2\eta_{p(s)}}{n_s\tau^2+\sigma_s^2}, \frac{\tau^2\sigma_s^2}{n_s\tau^2+\sigma_s^2}\right).$$

The complete conditional posterior distributions for the variance components σ^2, τ^2, and ω^2 are inverse-gamma distributions. For the within-school variance σ_s^2 for school s, the complete conditional distribution is

$$\left[\sigma_s^2 \mid \boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{Y},\boldsymbol{\nu}\right] \sim \text{Inverse-Gamma}\left(a_\sigma + n_s/2, \lambda_\sigma + \frac{1}{2}\sum_{i\in s}(r_i - \nu_s - \eta_{p(s)})\right),$$

where $r_i = \theta_i - \boldsymbol{\gamma}'\boldsymbol{Y}$. The between-school variance τ^2 depends on the school- and PSU-intercepts, $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$,

$$\left[\tau^2 \mid \boldsymbol{\nu},\boldsymbol{\eta}\right] \sim \text{Inverse-Gamma}\left(a_\tau + S/2, b_\tau + \sum_{s=1}^{S}(\nu_s - \eta_{p(s)})^2\right),$$

where S is the number of schools in the sample. The final variance component is the between-PSU variance; the complete conditional distribution for the between-PSU variance depends only on the PSU-intercepts,

$$\left[\omega^2 \mid \boldsymbol{\eta}\right] \sim \text{Inverse-Gamma}\left(a_\omega + Q/2, b_\omega + \sum_{k=1}^{Q} \eta_k^2\right),$$

where Q is the number of PSUs in the sample.

Estimating Group Quantities

In a superpopulation formulation of the NAEP model, we want to estimate the expected value of the function $g(\theta)$ for a randomly sampled individual from group $\mathcal{G}$. Specifically, we examine the posterior distribution of $g(\theta)$ for a student from group $\mathcal{G}$ given the data observed in the sample, denoted $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$. The posterior mean of $g(\theta)$ serves as a point estimate

$$E[g(\theta_i) \mid \mathcal{D}, i \in \mathcal{G}] = \int_{\boldsymbol{y}} E[g(\theta) \mid \mathcal{D}, \boldsymbol{Y} = \boldsymbol{y}] dF(\boldsymbol{y}|\mathcal{D}, \mathcal{G}), \tag{14}$$

where $F(\boldsymbol{y} \mid \mathcal{G})$ is the distribution of background characteristics for individuals in group $\mathcal{G}$ and

$$E[g(\theta) \mid \mathcal{D}, \boldsymbol{Y} = \boldsymbol{y}] = \int_{\boldsymbol{\gamma}} E[g(\theta) \mid \boldsymbol{\gamma}, \boldsymbol{Y} = \boldsymbol{y}] dF(\boldsymbol{\gamma} \mid \mathcal{D}). \tag{15}$$

Calculating the quantity $E[g(\theta) \mid \boldsymbol{\gamma}, \boldsymbol{Y}]$ is straightforward because, by assumption the distribution of θ conditional on the background vector $\boldsymbol{Y}$ is normal,

$$[\theta \mid \boldsymbol{Y} = \boldsymbol{y}] \sim \mathcal{N}(\boldsymbol{\gamma}'\boldsymbol{y}, \sigma^2 + \tau^2 + \omega^2).$$

The MCMC algorithm estimates the posterior distribution $F(\boldsymbol{\gamma} \mid \mathcal{D})$. However, estimating the population distribution $F(\boldsymbol{y} \mid \mathcal{D}, \mathcal{G})$ of background characteristics for a given group can be as complicated as the estimation of the distribution of the latent proficiencies for a group. Although we could develop a complicated model to derive an estimate of the distribution $F(\boldsymbol{y} \mid \mathcal{D}, \mathcal{G})$, that is beyond the scope of this paper. Instead, we use the sampling weights to estimate the distribution. Specifically, we estimate the population density of the background characteristics $\boldsymbol{y}$ with the weighted sample proportion

$$f(\boldsymbol{y} \mid \mathcal{D}, \mathcal{G}) = \frac{\sum_{\{i:\boldsymbol{Y}_i = \boldsymbol{y},\ i \in \mathcal{G}\}} w_i}{\sum_{i \in \mathcal{G}} w_i}.$$

Given the output from the MCMC algorithm, we estimate the posterior expectation $g(\theta)$ for group $\mathcal{G}$ according to the following algorithm.

For MCMC iterations $t = 1, \ldots, T$

1. Draw $\boldsymbol{\gamma}^{(t)}$ from the MCMC estimated posterior distribution.

2. For each individual i in the sample calculate

$$g_i^{(t)} = E[g(\theta) \mid \boldsymbol{\gamma}^{(t)}, \boldsymbol{Y}_i] = \int_\theta g(\theta) d\Phi(\theta;\ \boldsymbol{\gamma}^{(t)}\boldsymbol{Y}_i, \sigma^2 + \tau^2 + \omega^2),$$

   where $\Phi(\cdot; \mu, \sigma^2)$ is the cumulative distribution function (cdf) for a normal random variable with mean μ and variance σ^2. The integral in the equation above is simple for many quantities of interest such as the mean, variance, or the probability of an individual having an proficiency greater some cutpoint. For more complex functions of $g(\cdot)$, we use numerical integration techniques. The numerical integration could be achieved using either Monte Carlo or quadrature techniques. For computational efficiency, we suggest numerical quadrature.

3. Set

$$\overline{g}_{\mathcal{G}}^{(t)} = \frac{\sum_{i \in \mathcal{G}} w_i g_i^{(t)}}{\sum_{i \in \mathcal{G}} w_i}.$$

The values $\overline{g}_{\mathcal{G}}^{(t)}$ are draws from the posterior distribution of $g(\theta)$ for individuals in group $\mathcal{G}$. We use these draws to perform Monte Carlo integration in order to estimate features of the distribution $f(g(\theta)|\mathcal{D}, \mathcal{G})$ (e.g., median, mean, quantiles) using the discrete set $\{\overline{g}_{\mathcal{G}}^{(t)} :\ t = 1, \ldots, T\}$.

6 Analysis

We examine the performance of our proposed hierarchical model using two data sets. The first data set is a simulated data set meant to mimic the real NAEP data. The second data set is a subset from an actual data set from the NAEP 1998 reading assessment at grade 8.

Simulated Data Set

A data set was simulated to mimic a true observed NAEP data set. The latent abilities (θ) for 4,700 examinees were simulated according to the following specifications. First, the PSU means ($\boldsymbol{\eta}$) for 94 PSUs were sampled such that the between PSU variance explains 6.2% of the total variance. Second, school means ($\boldsymbol{\nu}$) were drawn for five schools within each PSU, such that the the between school variance within each PSU accounted for 16.3% of the total variance. Finally 10 students were simulated for each school. Each student was classified into one of 32 demographic groups defined by race, gender, and parents' education. The between group variation accounted for 24.2% of the total variance. The remaining 53.3% is within-school, within-group variance.

The item responses of these 4,700 simulated examinees were generated according the 2PL and 3PL models and the GPCM. We used the item types and item parameters from the *reading to perform a task* scale in the NAEP 1998 reading assessment at grade 8. The items in this scale are divided into three blocks. We assume that each student only receives two blocks of the items. Each examinee in the simulated data set receives two blocks at random according to a spiraled design and the third block is treated as missing by design.

In order to assess the performance of the Bayesian estimation procedure, we examine how well the program is able to recover the simulating parameters. We examine the performance by comparing the variance components, group mean estimates, and achievement level results.

Recoverability of variance components. The posterior median for the percentage of the total variance attributable to between-PSU variation is 5.5% and the 95% posterior credible interval is (2.8, 9.1). Although the point estimate is lower than the simulating percentage (6.2%), the credible interval clearly contains the true value. The model estimates the proportion of variation explained by between-school variation to be 15.5% with a 95% credible interval of (12.5, 19.0). Again the point estimate is a little lower than the simulating 16.3%, but the credible interval does contain that value. Conversely the within-school, or unexplained variance component is overestimated. The estimate suggests that the within-school variance makes up 56.2% of the total variation with a 95% credible interval equal to (52.8, 60.2), which contains the simulating value.

Recoverability of group means. Table 1 contains the true simulating means and the estimated group means for the demographic groups used to simulate the data set. The first column is an identifier of the group, the second column is the simulating value, and the third column is the posterior median, followed by the 95% posterior credible interval estimated using our MCMC estimation procedure; the final three rows are estimates derived from an operational analysis of the simulated data.

In each of the 10 groups, the true simulating parameter is contained inside both the 95% credible interval from the Bayesian analysis and within the 95% confidence interval from the current estimation procedure. However, the confidence intervals derived from the current operational strategy are 50% narrower than those found with the Bayesian analysis. Part of this difference is attributable to the fact that the NAEP analysis treats the item parameters as fixed and known.

No interaction effects were included in the model. therefore if we refine the table and look at

Table 1.

The True and Estimated Mean Proficiencies for Groups in the Simulated Data Set

Group	True mean	MCMC–posterior Median	MCMC–posterior 95% C.I.	NAEP MLE	NAEP 95% C.I.
Male	−0.21	−0.21	(−0.28, −0.14)	−0.20	(−0.24, −0.16)
Female	0.17	0.16	(0.09, 0.23)	0.17	(0.13, 0.21)
White & Other	0.27	0.26	(0.19, 0.33)	0.27	(0.24, 0.30)
Black	−0.61	−0.61	(−0.70, −0.53)	−0.61	(−0.68, −0.55)
Hispanic	−0.48	−0.49	(−0.57, −0.41)	−0.47	(−0.53, −0.40)
Asian	0.28	0.32	(0.20, 0.45)	0.29	(0.17, 0.42)
Less than H.S.	−0.63	−0.62	(−0.70, −0.53)	−0.61	(−0.68, −0.54)
High School	−0.30	−0.31	(−0.39, −0.23)	−0.29	(−0.35, −0.24)
More than H.S.	0.11	0.13	(0.05, 0.22)	0.13	(0.05, 0.21)
College	0.30	0.28	(0.21, 0.35)	0.29	(0.25, 0.33)

Note. The 95% C.I. for the MCMC approximated posterior calculations is the 95% equal-tailed posterior credible interval for the group.

the mean of the simulated proficiencies for the crossed groups (e.g., White, female, < high school; see Table 2) and compare to the estimated means for these subgroups, we find that our estimates are unsatisfactory. Just over 59% (19 out of 32) of the 95% posterior credible intervals contain the simulating parameter. As is true in all regression-type analyses, it is best to be sure that there are no interaction effects before interpreting a simple main-effects-only model.

Recoverability of proportion in achievement levels. Finally, we examine the performance of the MCMC algorithm for large-scale assessment data in terms of achievement levels. Let all students with a proficiency greater than the cutoff point 1.59 be classified as advanced students, and all students with a proficiency greater than -0.51 be classified as students with an above basic understanding.

Table 3 contains the true proportion of the simulated proficiencies that are in the two

Table 2.

Group Means for the Simulated Data for Subgroups

Group	Male		Female	
	True mean	MCMC	True mean	MCMC
White & Other, < High School	−0.61	(−0.57, −0.37)	−0.22	(−0.18, 0.02)
White & Other, High School	−0.20	(−0.33, −0.15)	0.13	(0.06, 0.24)
White & Other, > High School	0.16	(0.1, 0.28)	0.51	(0.49, 0.68)
White & Other, College	0.35	(0.2, 0.36)	0.70	(0.59, 0.75)
Black,< High School	−1.22	(−1.43, −1.2)	−0.73	(−1.03, −0.81)
Black, High School	−1.21	(−1.18, −0.97)	−0.63	(−0.78, −0.59)
Black, > High School	−0.59	(−0.75, −0.54)	−0.10	(−0.35, −0.14)
Black, College	−0.59	(−0.65, −0.45)	−0.18	(−0.26, −0.06)
Hispanic, < High School	−0.92	(−1.11, −0.91)	−0.58	(−0.72, −0.52)
Hispanic, High School	−0.71	(−0.87, −0.67)	−0.28	(−0.48, −0.29)
Hispanic, > High School	−0.41	(−0.45, −0.24)	0.01	(−0.06, 0.15)
Hispanic, College	−0.36	(−0.35, −0.16)	0.14	(0.04, 0.23)
Asian, < High School	0.05	(−0.55, −0.27)	−0.19	(−0.17, 0.12)
Asian, High School	0.16	(−0.32, −0.05)	0.03	(0.06, 0.35)
Asian, > High School	0.22	(0.11, 0.38)	0.50	(0.5, 0.77)
Asian, College	0.22	(0.21, 0.46)	0.58	(0.6, 0.86)

Note. Subgroups defined by crossing the three demographic variables gender, race, and parents' highest level of education.

achievement levels, the 95% posterior credible intervals estimated from the MCMC algorithm, and the 95% confidence intervals approximated with the operational NAEP analysis for the 10 groups defined by gender, race, and parents' education. All of the credible intervals and confidence intervals contain the proportion calculated directly from the proficiency scores used to simulate the data set.

Table 3.

The True and Estimated Proportion of Students Above Basic Proficiency Levels and in Advanced Proficiency Level for Groups in Simulated Data Set

Group	Above basic			Advanced		
	True prop.	MCMC	NAEP	True prop.	MCMC	NAEP
Male	0.62	(0.59, 0.64)	(0.60, 0.64)	0.029	(0.028, 0.042)	(0.024, 0.041)
Female	0.75	(0.72, 0.77)	(0.74, 0.77)	0.071	(0.068, 0.094)	(0.056, 0.085)
White & Other	0.80	(0.77, 0.82)	(0.79, 0.82)	0.072	(0.069, 0.096(	(0.063, 0.089)
Black	0.44	(0.41, 0.49)	(0.39, 0.49)	0.009	(0.007, 0.013)	(0.001, 0.014)
Hispanic	0.52	(0.47, 0.54)	(0.48, 0.56)	0.014	(0.011, 0.019)	(0.004, 0.018)
Asian	0.83	(0.78, 0.85)	(0.78, 0.89)	0.072	(0.071, 0.119)	(0.016, 0.104)
Less than H.S.	0.44	(0.41, 0.48)	(0.41, 0.49)	0.011	(0.007, 0.013)	(0.001, 0.017)
High School	0.60	(0.54, 0.61)	(0.56, 0.63)	0.022	(0.018, 0.030)	(0.009, 0.029)
More than H.S.	0.76	(0.72, 0.78)	(0.72, 0.79)	0.044	(0.054, 0.082)	(0.039, 0.075)
College	0.80	(0.78, 0.82)	(0.79, 0.82)	0.081	(0.074, 0.104)	(0.067, 0.095)

Note. The columns labeled MCMC contain the approximated 95% equal-tailed credible intervals from the Bayesian analysis and the columns labeled NAEP contain the 95% confidence intervals derived from the operational analysis.

The widths of the credible intervals and the confidence intervals are very similar, which is somewhat counterintuitive. Recall in the analysis of the group mean estimates we found that the intervals from the NAEP analysis tended to be narrower, because the NAEP analysis treats item parameters as known whereas the Bayesian analysis treats them as unknown.

One possible explanation for why the intervals are wider than expected from the NAEP analysis is that the approximation of the sampling variance that uses only a single plausible value provides an overestimate of the sampling variance. When all five plausible values are used to approximate the sampling variance in the NAEP procedure, the resulting intervals tend to be slightly narrower than those obtained from the Bayesian analysis. For example, the confidence

interval for the proportion of students whose parents finished college that are in the advanced level is (0.068, 0.094), which is narrower than the Bayesian interval for the same group (0.074, 0.104).

NAEP 1998 Reading Assessment at Grade 8

We apply the hierarchical model introduced in Section 5 to the *national reading to perform a task* scale in the NAEP 1998 reading assessment at grade 8. There were 4,494 eighth grade students that were assigned a booklet that contained items from this scale. Three background variables are included in the analysis: gender, race, and parent's education. The model is fit under two conditions: (a) fixing the item parameters at their estimates from the operational analysis of the data, and (b) treating the items as unknown and estimating them as part of the hierarchical model.

The posterior median of the total variance ($\sigma^2 + \tau^2 + \omega^2 +$ variance between groups) is fixed to equal one so that the results from the hierarchical models can be compared to the results from the NAEP analysis of the same data. The total between-school variance ($\tau^2 + \omega^2$) accounted for approximately one eighth of the total within-subgroup variation

$$\frac{\tau^2 + \omega^2}{\sigma^2 + \tau^2 + \omega^2} = \frac{1}{8}.$$

The between-PSU variance accounted for 19% of the total between-school variance,

$$\frac{\omega^2}{\tau^2 + \omega^2} = 0.19.$$

Group means. Table 4 contains the point estimate and standard errors (the MCMC SEs are actually posterior standard deviations) for the mean proficiency score for each of the subgroups defined by gender, race, and parent's education under the two hierarchical models and the NAEP estimation methodology. In addition, the table contains point estimates of the variability of the proficiency scores within each group.

In most cases, the estimates of the average proficiency score for each group calculated using the hierarchical models are very close to the estimates derived using the NAEP methodology. The groups with the largest differences are the Hispanic students (difference 0.08) and the students whose parents did not finish high school (difference 0.06).

The Bayesian model that fixes the item parameters at their maximum likelihood estimates reports standard deviations that are, on average, 10% smaller than those reported by the model

Table 4.

Estimates of the Mean Proficiency for Several Groups Under Two Hierarchical Models and NAEP Operational Model

	MCMC Estimated items			MCMC Fixed items			NAEP		
Group	Mean	SE	Var	Mean	SE	Var	Estimate	SE	Var
Overall	0.000	(0.026)	1.00	0.000	(0.022)	1.00	0.000	(0.023)	1.00
Male	−0.229	(0.031)	0.94	−0.237	(0.027)	0.94	−0.211	(0.034)	0.96
Female	0.234	(0.030)	0.95	0.242	(0.027)	0.95	0.221	(0.027)	0.95
White/AI/Other[a]	0.211	(0.029)	0.89	0.210	(0.026)	0.89	0.221	(0.028)	0.86
Black	−0.566	(0.049)	0.89	−0.564	(0.046)	0.89	−0.608	(0.038)	0.86
Hispanic	−0.494	(0.046)	0.90	−0.491	(0.042)	0.90	−0.576	(0.062)	0.93
Asian	0.100	(0.075)	0.89	0.100	(0.072)	0.90	0.117	(0.202)	0.96
Less than H.S.[b]	−0.559	(0.045)	0.90	−0.556	(0.041)	0.90	−0.620	(0.055)	0.87
High School	−0.258	(0.039)	0.90	−0.258	(0.037)	0.90	−0.269	(0.048)	0.91
More than H.S.	0.173	(0.043)	0.90	0.174	(0.040)	0.90	0.143	(0.041)	0.83
College	0.266	(0.032)	0.89	0.264	(0.028)	0.89	0.306	(0.030)	0.88

Note. SE = standard error. Var = variability. [a] The White/AI/Other racial group contains all students that were classified as a race other than Black, Hispanic, or Asian. [b] The less than H.S. category also contains students whose parents' education could not be determined.

that correctly treats them as unknown. The posterior standard deviations for the group means are similar to the standard errors derived from the NAEP methodology. The biggest difference between the posterior standard deviations from the hierarchical models and the NAEP standard errors in Table 4 is the difference for Asian students where the posterior standard deviations derived from the hierarchical models are less than half the size of the standard error derived from the operational NAEP analysis.

Further inspection of the data reveals why the standard error derived from the jackknife is so much larger than that found using the hierarchical method. The 62 replicate means for the

Asian students are plotted in a series in Figure 1. The figure shows that the mean calculated from replicate group #28 is very different from the means calculated from the other replicate groups. When the unusual set of replicate weights is replaced with the set of *overall* weights, the

***Figure* 1. The mean proficiency scores for Asian students calculated using each of the 62 replicate weights. The 28th replicate set of weights produces a mean that is very different from the other 61 sets.**

standard error of the mean estimate for Asian students is estimated to be 0.09, much closer to the posterior standard deviations estimated from the hierarchical models. The unusual replicate mean can be attributed to the fact that the PSUs or schools that make up the 28th PSU pair perform very different from one another. Whether the difference in the estimated standard deviation and the standard errors should be considered a weakness of the hierarchical model or the jackknife estimator has yet to be determined. The operational analysis is clearly less robust to unusual PSUs in contrast to the Bayesian analysis.

Achievement levels. Table 5 and Table 6 contain the estimates of the proportion of students who have an above basic and advanced understanding of the *reading to perform a task* scale, respectively. We report the mean of the posterior distribution of the proportion as the Bayesian point estimate and the posterior standard deviation as the measure of uncertainty about that Bayesian estimate. The Bayesian estimates are compared to the estimates calculated from the NAEP three-stage estimation procedure.

Table 5.

Proportion of Students Who Perform Above Basic Proficiency Level on Reading to Perform a Task Scale

	Above basic ($\theta > -0.5$)			
	MCMC		NAEP	
	Estimated items			
Group	Mean	SE	Estimate	SE
Overall	0.693	(0.010)	0.700	(0.011)
Male	0.613	(0.012)	0.625	(0.017)
Female	0.775	(0.010)	0.779	(0.011)
White	0.775	(0.010)	0.784	(0.011)
Black	0.473	(0.021)	0.461	(0.030)
Hispanic	0.502	(0.019)	0.500	(0.028)
Asian	0.738	(0.026)	0.722	(0.097)
Less than H.S.	0.474	(0.019)	0.473	(0.032)
High school	0.602	(0.016)	0.590	(0.022)
More than H.S.	0.760	(0.014)	0.776	(0.020)
College	0.792	(0.010)	0.810	(0.012)

The achievement level results are similar to the group average results when comparing point estimates. The estimates of the proportion of students in the basic achievement levels are very similar under the two estimation strategies. All differences in the estimates for the proportion of students above the basic achievement level are within ± 0.02 and within two posterior standard

deviations.

In a few cases, the estimates are significantly different for the advanced level. The differences between the proportions of Black and Hispanic students, as well as the proportion of students whose parents did not finish high school and attended some school after high school, differ significantly. In each case, the Bayesian model provides estimates higher than those from the operational analysis. This is likely the result of *shrinkage* towards the mean, a common occurrence in Bayesian estimates.

Table 6.

Proportion of Students Who Perform at Advanced Proficiency Level on Reading to Perform a Task Scale

	Advanced ($\theta > 1.6$)			
	MCMC w/ PSU		NAEP	
Group	Mean	SE	Estimate	SE
Overall	0.053	(0.004)	0.050	(0.005)
Male	0.028	(0.003)	0.026	(0.005)
Female	0.079	(0.006)	0.075	(0.009)
White	0.070	(0.006)	0.066	(0.007)
Black	0.011	(0.002)	0.005	(0.003)
Hispanic	0.014	(0.002)	0.008	(0.003)
Asian	0.056	(0.010)	0.057	(0.020)
Less than H.S.	0.011	(0.002)	0.006	(0.004)
High School	0.025	(0.003)	0.025	(0.011)
More than H.S.	0.066	(0.007)	0.051	(0.012)
College	0.078	(0.006)	0.078	(0.008)

Unlike the group mean results, the posterior standard deviation and the standard errors are not that similar. On average, the standard errors derived from the NAEP three-stage procedure are nearly 1.5 times the posterior standard deviations of the proportion of students above basic.

The average standard error for the advanced achievement level estimates are almost twice as large as the posterior standard deviations. These results are in line with what was observed in the analysis of the simulated data sets: Using only a single plausible value provides an estimate of the sampling variance that is too large.

7 Conclusions

This paper has proposed a Bayesian hierarchical model for the analysis of large-scale assessments such as NAEP. The model builds on standard item response models by incorporating levels of the model that correctly cluster the assessed respondents within schools and schools within geographic regions. Our model, in contrast to the method currently employed for the analysis of NAEP data, allows the user to estimate all model parameters simultaneously, rather than in stages. The analysis of the simulated data shows that the model can recover the simulating parameters, and the analysis of the *reading to perform a task scale* for the NAEP 1998 reading assessment at grade 8 shows that our model provides estimates similar to those derived using the current NAEP operational analysis method.

Although our Bayesian model has the advantage that it incorporates all sources of uncertainty into a single model, it requires a computationally intensive MCMC algorithm for the approximation of the posterior distribution of the model parameters. If the assessment is relatively small, such as the examples contained in this paper where we had only a few background variables and less than 5,000 examinees, our model is probably preferable; the examples discussed in this paper required approximately 12 hours of computer time on a 2.0 GHz Pentium workstation. However, when assessments are much larger the computer time required would be too great. For example, the full NAEP 1998 reading assessment consisted of both a national assessment and approximately 45 state/district assessments; separate assessments were conducted at fourth, eighth, and twelfth grade; the assessments were designed to measure a three-dimensional proficiency; and the assessments each contained well over 100 background variables. For assessments the size of NAEP, the current operational methodology described in Section 4 is probably preferable because it can handle multidimensional assessments and it requires far less computer power, but as computers become faster and cheaper our model will be able to handle these huge assessments.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (No. NCES 2001-509). Washington, DC: National Center for Educational Statistics.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, *32*, 283–301.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistics theories of mental test scores.* Reading, MA: Addison-Wesley.

Cassel, C.-M., Särndal, C.-E., & Wretman, J. (1993). *Foundations of inference in survey sampling.* Malabar, FL: Krieger Publishing Company.

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

Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, *6*, 721–741.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice.* Boca Raton, FL: Chapman & Hall/CRC.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Linden, W. J. van der, & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Longford, N. T. (1995). *Model-based methods for analysis of data from the 1990 NAEP trial state assessment* (Research and Development Report 95-696). Washington, DC: National Center for Educational Statistics.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex

samples. *Psychometrika*, *56*, 177–196.

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

Muraki, E., & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating scale data [Computer software]. Chicago, IL: Scientific Software International.

Patz, R., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Potthoff, R., Woodbury, M., & Manton, K. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, *87*(418), 383–396.

Raudenbush, S., Fotiu, R., & Fai, C. (1999). Synthesizing results from the trial state assessment. *Journal of Educational and Behavioral Statistics*, *24*(4), 413–438.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley, John & Sons, Inc.

Scott, S. L., & Ip, E. H. (2002, June). Empirical Bayes and item-clustering in a latent variable hierarchical model: A case study from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, *97*(458), 409-419.

Thomas, N. (1993). *The E-step of the MGROUP EM algorithm* (ETS RR-93-37). Princeton, NJ: Educational Testing Service.

Zwick, R. (1992). National Assessment of Eductional Progress [Special issue]. *Journal of Educational Statistics*, *17*(2), 93–232.