Assessing fit of latent regression models

Sandip Sinharay, Zhumei Guo, and Matthias von Davier
Educational Testing Service, Princeton, NJ, USA[1]

Bernard P. Veldkamp
University of Twente, Enschede, The Netherlands

The reporting methods used in large-scale educational survey assessments such as the National Assessment of Educational Progress (NAEP) rely on a latent regression model. Research assessing the fit of latent regression models is lacking. This article suggests a simulation-based model fit procedure to assess the fit of such models. The procedure involves investigating whether the latent regression model adequately predicts basic statistical summaries. Application of the suggested procedure to an operational NAEP data set revealed important information regarding the fit of the latent regression model to the data.

[1] The opinions expressed herein are those of the author and do not necessarily represent those of Educational Testing Service.

INTRODUCTION

The National Assessment of Educational Progress (NAEP), the only regularly administered and mandated national assessment program in the United States (see, for example, Beaton & Zwick, 1992), is an ongoing survey of the academic achievement of school students in the United States in a number of subject areas, such as reading, writing, and mathematics. In 1984, researchers reporting NAEP results began using a statistical model consisting of two components: (i) an item response theory (IRT) component, and (ii) a linear regression component (see, for example, Beaton, 1987; Mislevy, Johnson, & Muraki, 1992). Researchers conducting other large-scale educational assessments, such as the International Adult Literacy Study (IALS; Kirsch, 2001), the Trends in Mathematics and Science Study (TIMSS; Martin & Kelly, 1996), and the Progress in International Reading Literacy Study (PIRLS; Mullis, Martin, Gonzalez, & Kennedy, 2003) adopted a very similar model. This model is often referred to as a *latent regression model* (LRM). The DGROUP set of programs (Rogers, Tang, Lin, & Kandathil, 2006), which is a product of Educational Testing Service (ETS), can be used to estimate the parameters of this model.

Standard 3.9 of the *Standards for Educational and Psychological Testing* (American Psychological Association, National Council on Measurement in Education, & American Educational Research Association, 1999) demands evidence of model fit when an IRT model is used to make inferences from a data set. It is therefore important to assess the fit of the LRM used in NAEP to ensure quality control and an overall improvement in the long term. Although some model-checking procedures have been applied to the NAEP model (e.g., Beaton, 2003; Dresher & Thind, 2007; Li, 2005), there is scope for further work in this area.

In this article, we recommend use of a simulation-based procedure to assess fit of the LRM used in NAEP and other large-scale assessments. The procedure involves investigating whether the model adequately predicts several summary statistics of the observed data. The suggested procedure, which generates predicted data sets under the assumption that the model is true, involves use of the NAEP operational software, and it compares several summary statistics computed from the observed data set to those computed from the predicted data sets. Our procedure is therefore similar to the parametric bootstrap (e.g., Efron & Tibshirani, 1993) and the posterior predictive model checking method (e.g., Gelman, Carlin, Stern, & Rubin, 2003).

We provide, in Section 1 of this article, background information on the current NAEP statistical model and estimation procedure, and on the existing NAEP model-checking procedures. We then, in Sections 2 and 3, describe our suggested model checks and the NAEP data set that we used for our study. In the fourth section, we explore the Type I error rates of the suggested procedure. We provide a real data example in the fifth section. In the final section of the article (Section 6), we present conclusions and suggestions for future work.

1. THE NAEP LATENT REGRESSION MODEL AND ESTIMATION

1.1 The Model

In NAEP, the latent proficiency variable for student *i* is assumed to be *p*-dimensional, where *p* could be between 1 and 5. Let us denote it as $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots \theta_{ip})'$.

Let us denote the response vector to the test items for student *i* as $\mathrm{y}_i = (\mathrm{y}_{i1}, \mathrm{y}_{i2}, \ldots \mathrm{y}_{ip})$, where y_{ik}, a vector of responses, contributes information about θ_{ik}. For example, y_{ik} could be responses of student *i* to algebra questions in a mathematics test and θ_{ik} the algebra skill variable of the student. Let us denote the item parameters of the items that are designed to elicit information on θ_{ik}'s (i.e., items that measure the *k*-th subscale) as β_k. Suppose $\beta = (\beta_1, \beta_2, \ldots \beta_p)$. The likelihood for a student is given by

$$f(\mathrm{y}_i | \theta_i, \beta) = \prod_{k=1}^{p} f_1(\mathrm{y}_{ik} | \theta_{ik}, \beta_k) \equiv L(\theta_i, \beta; \mathrm{y}_i). \qquad (1)$$

The expressions $f_1(\mathrm{y}_{ik} | \theta_{ik}, \beta_k)$, above, consist of factors contributed by a univariate IRT model, usually the two- or three-parameter logistic (2PL, 3PL) model for dichotomous items and the generalized partial-credit model (GPCM) for polytomous items.

Suppose $\mathrm{x}_i = (\mathrm{x}_{i\,1}, \mathrm{x}_{i\,2}, \ldots \mathrm{x}_{im})$ are *m* covariates for the *i*-th student. Typically, NAEP collects information on demographic and educational characteristics, converts them to numerical variables, and then uses a principal component analysis for extraction of principal components that explains 90% of the variance of these variables (see, for example, Allen, Donoghue, & Schoeps, 2000); the values of the principal components play the role of the x_{ij}'s in further analyses. Conditional on x_i, the student proficiency vector θ_i is assumed to follow a multivariate normal distribution, that is,

$$\theta_i | \mathrm{x}_i, \mathrm{y}_i, \mathrm{y}, \beta, \Gamma, \Sigma \sim N(\Gamma' \mathrm{x}_i, \Sigma). \qquad (2)$$

Together, Equations 1 and 2 form the LRM or *conditioning model* employed in NAEP. Equations 1 and 2 imply that

$$p(\theta_i | \mathrm{y}_i, \mathrm{x}_i, \beta, \Gamma, \Sigma) \propto L(\theta_i, \beta; \mathrm{y}_i)\, N(\Gamma' \mathrm{x}_i, \Sigma). \qquad (3)$$

where $p(\theta_i | \mathrm{y}_i, \mathrm{x}_i, \beta, \Gamma, \Sigma)$ is the conditional posterior distribution of θ_i. (For further details, see, for example, von Davier, Sinharay, Oranje, & Beaton, 2006).

1.2 Estimation

NAEP uses a three-stage estimation process for fitting the above-mentioned LRM to the data.

1. The first stage, *scaling*, uses the PARSCALE software (Allen et al., 2000) to fit the model given by Equation 1 to the student response data and to estimate the item parameters. During this stage, the prior distributions of the components of the student proficiency are assumed to be independent, discrete univariate distributions.

2. The second stage, *conditioning*, assumes that the item parameters are fixed at the estimates found in the scaling stage and that they fit the model given by (1) and (2) to the data, and estimates Γ and Σ. The following versions of the DGROUP program perform this conditioning step differently.
   - BGROUP (Beaton, 1987) is employed when $p \leq 2$ and uses numerical quadrature.
   - CGROUP (Thomas, 1993) is employed when $p > 2$ and uses Laplace approximations.
   - NGROUP (Mislevy, 1985) is employed to find the starting values for BGROUP or CGROUP and uses a normal approximation of $L(\theta_i;\mathrm{y}_i)$.[1]
3. The third stage of the NAEP estimation process generates *plausible values* (imputed values of the proficiency variables) for all the students using the parameter estimates obtained from the scaling and conditioning stages. The plausible values are generated according to the following three-step process:
   - Draw $\Gamma \sim N(\hat{\Gamma},\hat{S}(\hat{\Gamma}))$, where $\hat{\Gamma}$ and $\hat{S}(\hat{\Gamma})$ are estimates of Γ and the corresponding standard deviation, respectively, and are obtained using DGROUP.
   - Compute from Equation 3 the posterior mean and the covariance of θ_i, conditional on the generated value of Γ and the fixed variance matrix $\Sigma = \hat{\Sigma}$.
   - Draw θ_i from a multivariate normal distribution, with the mean and variance computed in the above step.

The plausible values are used to estimate student subgroup averages. The third stage also estimates the variances corresponding to the student subgroup averages as the sum of two terms—the variance due to the latency of θ_is and the variance due to sampling of students. The computation of the second term involves the use of a jackknife approach, while the computations of both terms involve the plausible values generated in the conditioning step.

1.3 Existing Work on Assessing Fit of the NAEP Model

NAEP researchers rigorously monitor data quality and employ a number of qualitative checks of the results of their statistical analyses. When conducting first-level checks, NAEP researchers employ several plausibility analyses. (These involve examining the computer outputs to make sure that they make sense.) The researchers also conduct computer-based checks at different stages of the statistical analysis; these ensure that the data analysis process is working as intended. The first-level checks involve working through several carefully designed checklists, such as an item analysis checklist and a DGROUP conditioning checklist. However, these first-level checks provide quality control measures that are necessary but not sufficient. Thus, even if the checks reveal no problems and show that the programs are running as expected on the appropriate data sets, the inferences may be problematic if the model does not adequately explain the data. Therefore, as second-level checks, additional steps are taken to ensure the appropriateness and quality of the IRT model (Allen et al., 2000, p. 233).

[1] Because the item parameters are assumed to be known in this step, the symbol of the item parameters does not appear in this expression for the likelihood.

This check involves examination of item parameter estimates—extreme estimates often indicate problems—and use of differential-item-functioning (DIF) analyses to examine issues of multidimensionality (see, for example, Roussos & Stout, 1996, for the connection between DIF and multidimensionality). Those conducting NAEP operational analyses also employ graphical item fit analyses. These require use of residual plots and a related χ^2-type item fit statistic (Allen et al., 2000, p. 233) that provides guidelines on how to treat the items, such as collapsing categories of polytomous items, treating adjacent-year data separately in concurrent calibration, or dropping items from the analysis. However, the null distribution of these residuals and of the χ^2-type statistic are unknown, as Allen et al. (2000) acknowledge (p. 233).

Another second-level check used in NAEP operational analyses is comparison of observed and model-predicted proportions of students obtaining a particular score on an item (Rogers, Gregory, Davis, & Kulick, 2006). These analyses, however, do not use the variability of the quantities involved. We considered it would be useful to make the comparison of the observed and predicted proportions more meaningful by providing a methodology that incorporates the variability. As will be clear later, our work partially addressed this issue.

Beaton (2003) suggests item fit measures involving sums and sum of squares of residuals obtained from the responses of each student to each question. Assuming that Y_{ij} denotes the response of the *i*-th student to the *j*-th item, Beaton's fit indices are of the form

$$\sum W_i \; \frac{(Y_{ij} - E(Y_{ij}|\Theta))^k}{(\sqrt{Var(Y_{ij}|\Theta)})^k}$$

where k could be 1 or 2, Θ is the collection of all model parameters, and W_i is the NAEP sampling weight (Allen et al., 2000, pp. 161–225). A bootstrap method is then used to determine the null distribution of these statistics. Li (2005) used Beaton's statistics when analyzing operational test data sets in order to determine the effect of accommodations on students with disabilities. Dresher and Thind (2007) used Beaton's statistics when analyzing 2003 NAEP and 1999 TIMSS data. Dresher and Thind also employed the χ^2-type item fit statistic provided by the NAEP-PARSCALE program, but computed the null distribution of the statistic from its values for one simulated data set. These methods have their limitations, however. For example, Sinharay (2005, p. 379) argues that fit statistics based on examinee-level residuals are unreliable because of their excessive variability, a limitation that applies to Beaton's fit statistics. (See also Li, Bolt, & Fu, 2006, who found such statistics questionable.)

With any practical application of model fit analysis, it is important that analysts use suitable test statistics to examine the appropriate aspects of the model. The standard recommendation (see, for example, Gelman et al., 2003, p. 172) is that those checking a model in an application should focus on aspects of the model that are relevant to the purposes for which the inference will be applied. For example, if we were interested in using a statistical model to estimate the mean income of a population, we would need to focus the model fit analysis on the mean.

This issue has received little attention with respect to the IRT model fit in general (e.g., Sinharay, 2005) and with respect to NAEP in particular. Thus, there is substantial scope for further work directed at assessing the fit of LRMs that use NAEP data. Note that such work has to take full account of the idiosyncrasies of the NAEP model and of data such as the matrix sampling (so that each student sees only some of the questions), sampling weights, and missing values.

2. THE SUGGESTED PROCEDURE

During our study, we applied a simulation-based model fit procedure to NAEP statistical analysis to investigate whether the LRM used for NAEP adequately predicts several data summaries (or *test statistics*).

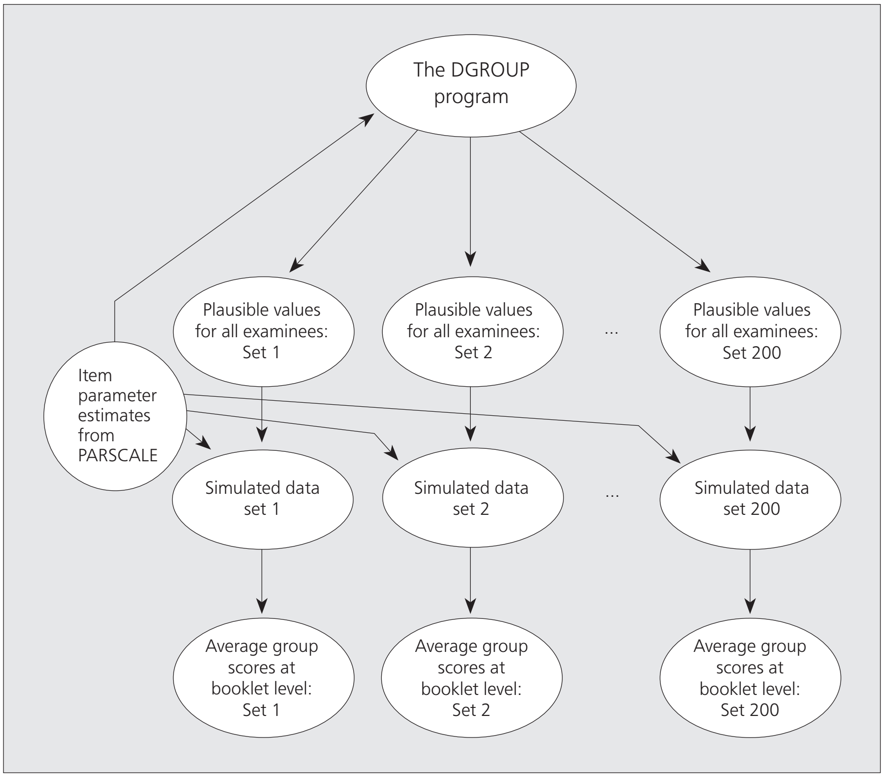
2.1 Description of the Procedure

The determination of the null distribution (or the computation of the *p*-values) of a test statistic is not straightforward, given the complicated nature of the LRM applied in NAEP. Because of this, we employed a simulation-based procedure that used existing NAEP software programs to determine the null distribution of the test statistics and to perform the model fit assessment. The steps in the procedure follow:

1. Computation, from the original NAEP data set, of several test statistics (such as biserial correlation). We describe these statistics in Section 2.2.
2. Estimation, using PARSCALE, of item parameters (as in the scaling stage of the NAEP three-stage estimation process).
3. Generation, using the DGROUP program of plausible values (as in the third stage of the NAEP estimation process). The operational NAEP estimation process generated five plausible values for each candidate, but we generated 200 plausible values for each candidate. These are like draws of θ_i from its posterior distribution.
4. Simulation, from the model given by Equation 1, of 200 data sets, using the generated plausible values and the item parameters estimated in Step 2, above. The simulated data sets can be considered to be those predicted by the model, that is, those we would observe if the model were valid.
5. Computation of the values of several test statistics for each of the 200 simulated data sets (resulting in 200 *simulated/predicted values* for each statistic). We then compared the predicted values of each statistic to the corresponding observed values in order to judge the goodness of fit of the model. An observed value that is extreme with respect to the distribution of the predicted values indicates model misfit. We performed the comparison of the observed and predicted values of the statistics graphically, by plotting the observed and predicted values of the statistics. An observed value located at the tail of the distribution of the predicted values indicates that the model does not adequately predict the corresponding statistic. We also computed *p*-values for the statistics. A *p*-value is the proportion of the predicted values of a statistic that is greater than the corresponding observed value. A very low or a very high *p*-value indicates that the model does not adequately predict the corresponding statistic.

The flowchart in Figure 1 provides a graphical description of the procedure for the average group score statistic (described below).

Figure 1: Steps of the simulation procedure to determine the null distribution of the average group score statistic


The DGROUP program
Plausible values for all examinees: Set 1
Plausible values for all examinees: Set 2
...
Plausible values for all examinees: Set 200
Item parameter estimates from PARSCALE
Simulated data set 1
Simulated data set 2
...
Simulated data set 200
Average group scores at booklet level: Set 1
Average group scores at booklet level: Set 2
Average group scores at booklet level: Set 200

Note: The item parameter estimates from PARSCALE were used in the DGROUP program, which generates 200 sets of plausible values. The plausible values and the item parameter estimates were used to generate 200 simulated data sets, which resulted in 200 simulated/predicted average group scores for each booklet for each student group. Each booklet-level-observed average group score was then compared to the corresponding 200 simulated values for model fit assessment.

The procedure just described is an approximation of the posterior predictive model-checking (PPMC) method (e.g., Gelman et al., 2003; Sinharay, 2005), a popular Bayesian model-checking procedure. The PPMC method involves the following four steps:

1. Generating a sample of size *n*, mostly using a Markov chain Monte Carlo method (Gelman et al., 2003) from the joint posterior distribution of the model parameters;
2. Simulation of *n* data sets using the generated parameter values;
3. Computation of the values of a test statistic of interest for each of these *n* simulated data sets; and

4. Comparison of the observed value of the corresponding test statistic with the *n* values computed in the above step.

Because our suggested procedure required us to perform the last three of these steps, the procedure is similar to the PPMC method. However, it is only an approximation of the PPMC method. This is because we performed only part of the first step involved in a PPMC: we drew plausible values (which are approximate draws from the student posterior distribution), but we did not draw item parameter values and assumed that these were fixed at their estimates (obtained from the scaling state of the NAEP estimation process). Sinharay (2005) and Sinharay, Johnson, and Stern (2006) successfully used the PPMC method to detect misfit of simple IRT models. Our suggested procedure involved application of several fit statistics similar to those recorded in these two articles.

Our procedure is also similar to the parametric bootstrap method (e.g., Efron & Tibshirani, 1993) that has been successfully applied to assess the fit of IRT models and other similar models (see, for example, Stone, 2000; von Davier, 1997). We consider our procedure fairly easy to understand because it is similar to two popular model-checking methods. In addition, because it uses existing NAEP software, operational implementation of the procedure is straightforward.

2.2 Description of the Test Statistics

With NAEP, unlike several other large-scale assessments, not all students are asked all items; instead, each student has to answer the items in one of several booklets. (A booklet is a collection of test items.) We computed all the test statistics separately for each booklet.

Researchers van der Linden and Hambleton (1997, p. 16) recommend collecting a wide variety of evidence about model fit and then making an informed judgment about model fit and usefulness of a model with a particular set of data for assessing the fit of two- and three-parameter IRT models. Sinharay (2005) and Sinharay et al. (2006) took heed of this recommendation when assessing the fit of simple IRT models. They used a variety of simple summaries of the data—similar to the ones listed below—to do this. The recommendation put forward by van der Linden and Hambleton (1997) is equally appropriate for any IRT model, including the one employed in NAEP.

Our suggested procedure, together with the statistics described below, provides a tool kit that researchers can use to collect a variety of evidence to determine the fit of the LRM to NAEP data.

- *Average group score:* Let Y_{ij} denote the response of the *i*-th student to the *j*-th item in a booklet. For a dichotomous item, Y_{ij} is 0 or 1. For a *k*-category polytomous item, Y_{ij} takes one from among the values 0, 1, ... *k*–1. NAEP encounters a substantial percentage of omitted and not-reached responses. In NAEP, not-reached items are treated as not-presented items. An omitted response is assigned a fractional score equal to the reciprocal of the number of options if the item is multiple-choice and is assigned the score for the lowest scoring category if the item is a constructed-response item (Allen et al., 2000, pp. 231–232). To obtain a statistic that appropriately takes into account the omitted and not-reached responses, we defined

$$s_i = \sum_j Y_{ij} / R_i$$

as the proportion-correct score of student *i*, where R_i is the sum of the maximum raw score points for the items that the *i*-th student reached. We then computed the weighted average of the s_is for a student group as

$$A_g = \frac{\sum_{i \in g} W_i s_i}{\sum_{i \in g} W_i}$$

where *g* denotes a student group (such as all students or male students, or White students). The statistic A_g denotes the average proportion score in a booklet for the *g*-th group. Note that if student *i* omitted item *j*, Y_{ij} is $1/m$ for a *m*-option multiple-choice item and is equal to the lowest scoring category for a constructed-response item. Because student subgroup means are reported in NAEP, the decision to examine how the NAEP model predicts the statistic A_g is a natural one.

- *Average item score:* We used the weighted average item score for item *j*,

  $$p_j = \frac{\sum_i W_i Y_{ij}}{\sum_i W_i}$$

  as a test statistic. This statistic is closely related to the proportion score statistic used by Rogers et al. (2006). The main difference between the two is that p_j is defined for a booklet.
- *Biserial correlation coefficients:* Because of the way the NAEP operational analysis treats the omitted and the not-reached items, the standard definition of the biserial correlation is not appropriate here. Accordingly, for each item in a booklet, we computed the correlation between the vector of responses to an item and the vector of proportion-correct scores s_i (using the notation introduced earlier). We used the sampling weights W_i in the computations.
- *Item pair correlation:* This is the correlation between the response vectors for two items. We again used the sampling weights W_i in the computations.

We chose the above statistics not only because they are simple data summaries but also because Sinharay (2005) and Sinharay et al. (2006) found results similar to these in their research. The average group score statistic deserves special mention. Ideally, model checking in an application should focus on those aspects of the model that are relevant to the purposes for which the inference will be applied (Gelman et al., 2003, p. 172). Because the quantities of primary interest in NAEP are the mean scale scores for the different subgroups, it is necessary to determine if the model adequately predicts these quantities. The ideal would be to compare the observed value[2] of a test statistic based on the mean scale scores to the model-predicted values of the test statistic. However, because mean scale scores are functions of model-estimated student proficiency variables, it is impossible to obtain a test statistic based on mean

[2] When we refer to an "observed value," we mean a value that can be computed from the data set before an appropriate model is fitted to it.

scale scores that will have an observed value. The average group scores of student subgroups of interest are thus best-possible observed proxies of the mean scale scores of these subgroups. We can expect that these average group scores, although simple to compute, will have strong correlations with their corresponding mean scale values. Accurate prediction of the average scores of important student subgroups by the NAEP model should thus provide strong evidence that the subgroup estimates provided by the NAEP model are accurate.

3. DATA

We obtained a data set from the NAEP 2002 reading assessment for Grade 12. This set contained data for about 15,000 students. Our primary reasons for choosing the reading assessment were that reading is a No Child Left Behind[3] subject and that reading items have typically been more likely than mathematics items to display problematic item fit (mathematics is another No Child Left Behind subject). The reading assessment considered here measured three skills—reading for literary experience, reading for information, and reading to perform a task.

The reading assessment had 38 booklets. Each of the first 36 booklets was given to a few hundred students, while each of the last two booklets was given to a few thousand students. In addition, each of the last two booklets consisted of one long block (out of a total of two long blocks) of items,[4] whereas each of the first 36 booklets consisted of two shorter blocks (out of a total of nine short blocks) of items. The number of items in a booklet was approximately 20 (about one-third multiple-choice items and about two-thirds constructed-response items) for the first 37 booklets and about 10 (all constructed-response items) for the last booklet. About 50% of all students taking the reading assessment were male, approximately 65% were White, and about 15% were Black. The proportion of omitted and not-reached responses ranged from 4% to 10% for the various booklets.

4. STUDYING THE TYPE I ERROR RATE OF THE SUGGESTED PROCEDURE

Studying the Type I error rate of any statistic used for model checking is important. As we noted earlier, our suggested procedure is similar to the bootstrap method (Efron & Tibshirani, 1993) and the posterior predictive model checking method (Gelman et al., 2003). Researchers have found that both of these methods have Type I error rates close to the nominal level for a wide variety of models, including IRT models. However, we performed a limited study to make sure that the Type I error rate of our suggested procedure was not too high.

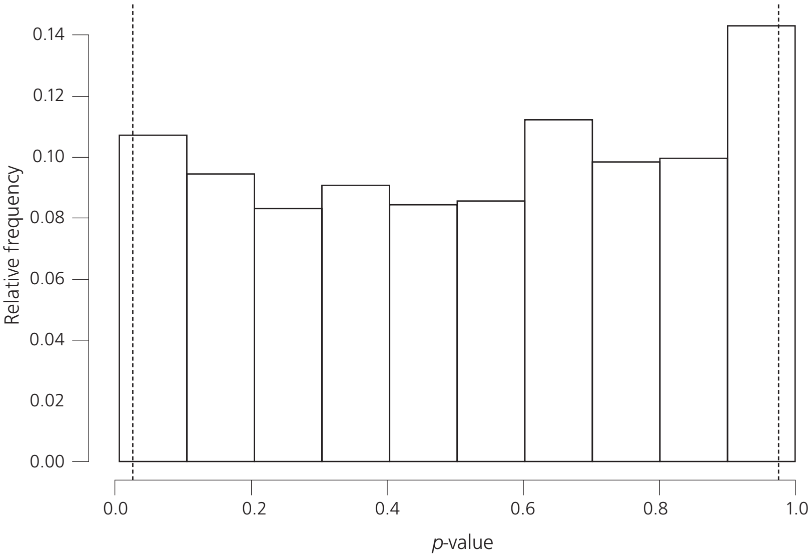
[3] The No Child Left Behind Act of 2001 requires states to develop assessments in basic skills to be given to all students in certain grades, if those states are to receive federal funding for schools.

[4] In NAEP, the item pool is divided into several blocks of items; each booklet typically consists of three blocks of items.

When carrying out our study, we began by simulating a data set from the NAEP model, using the techniques described in Section 2 of this article. Let us denote this simulated data set as *D*. The structure of *D* is the same as that described in Section 3 of this article. For example, *D* has responses from about 15,000 students to questions in 38 booklets (remember that each student worked on just one booklet).[5] We performed the model fit analysis (this involved running PARSCALE and DGROUP, simulating 200 data sets, computing the test statistics, and computing the *p*-values) described in Section 2 on *D*. Because *D* was simulated from the NAEP model, the model fitted *D* perfectly. So, ideally, we would have expected not to see any sign of misfit of the NAEP model to this data set.

The proportion of significant *p*-values was close to the nominal level of 5% for all the statistics—average group score, average item score, biserial correlation, and item-pair correlation. Figure 2 provides an example—a histogram of all 744 *p*-values for the biserial correlation statistic. The y-axis in this figure shows the relative frequency (frequency of an interval divided by the total frequency). Note the two vertical lines drawn at values .025 and .975. The figure shows that the *p*-values are more or less uniformly distributed between 0 and 1. Six percent of the *p*-values were greater than .975 or less than .025, very close to the nominal level. This outcome points to the acceptable Type I error rate of our suggested procedure.

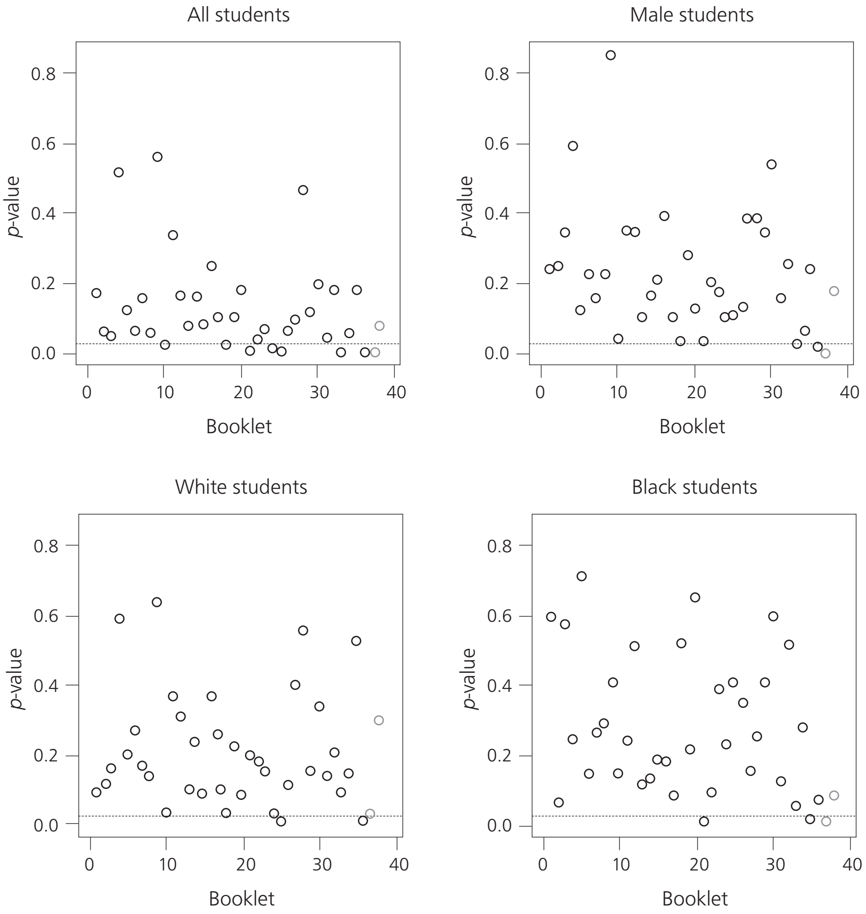
Figure 2: The *p*-values for the biserial correlations in the Type I error study



[5] This simulated data set is one of those simulated for the next section, where we study the fit of the NAEP model to the NAEP data set described in Section 3.

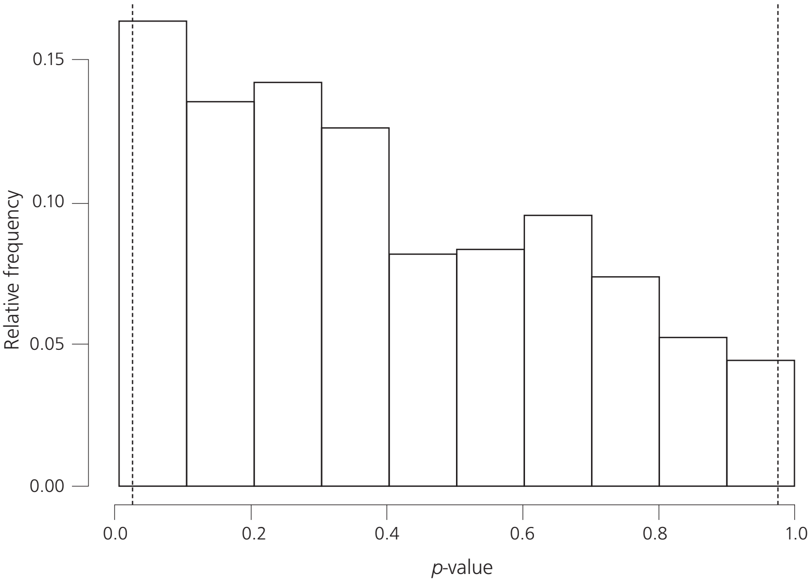
However, for the average group score statistic and the average item score statistic, the *p*-values did not seem to follow a uniform distribution and had a mean less than .5, as is evident in Figure 3. This figure shows all the *p*-values for the average group score statistic for four groups—male students, White students, Black students, and all students.[5] Each panel displays 38 points, and each point denotes the *p*-value for a booklet. The horizontal dashed line in each panel denotes the value of .025; a *p*-value below this indicates that the predicted values of the statistic were significantly lower than the corresponding observed value. The range of the y-axis is the same in all four panels of Figure 3. Figure 4 presents a histogram of all the 744 *p*-values for the average item score statistic.

Figure 3: The p-values for the average group score statistic in the Type I error study



[5] The first three of these groups are actually important subgroups in NAEP reporting.

Figure 4: The *p*-values for the average item score statistic in the Type I error study



Figures 3 and 4 show that the distribution of these *p*-values is not uniform; for example, more than half of the *p*-values are less than .5 for each of these statistics. However, because we were studying the misfit of the model to a data set that was simulated from the model, we would have expected the distribution of these *p*-values to be uniform, and the proportion of the *p*-values that are less than .5 to be very close to .5. Further research is needed on this issue. Fortunately, the percentage of *p*-values for these two statistics that was greater than .975 or less than .025 was close to 5, the nominal level.

We repeated all the analyses reported in this section using another simulated data set. The results, however, were similar to the preceding analysis; that is, the distribution of these *p*-values was not uniform.

5. RESULTS FROM THE ANALYSIS OF DATA FROM THE 2002 NAEP READING ASSESSMENT

In this section, we describe the results of our application of the procedure suggested in Section 2 to the 2002 NAEP reading data set described in Section 3. We first provide results for the four statistics and then provide a discussion of the results.

5.1 Average Group Score

Figure 5, which is similar to Figure 3, presents the *p*-values for the average group score statistic for all the booklets as well as for male students, White students, Black students, and all students. The figure shows the following:

1. *Some evidence of misfit for all students (top left panel):* about half of the *p*-values lie below .025.
2. *Most of the p-values in all panels lie below .5, which indicates that the predicted values were generally lower than the observed value of the statistic:* the *p*-values in Figure 5 are substantially smaller overall than those observed in Figure 3. Hence, even if we consider the distribution observed in Figure 3 as the true null distribution of *p*-values, the model seems to under-predict the average group score statistic in Figure 5.
3. *Little evidence of misfit in the plots for the male, White, and Black students:* however, note that most of the *p*-values corresponding to the last two booklets are 0.0 in these plots.

Figure 5: The *p*-values for the average group score statistic for the 2002 NAEP reading data set

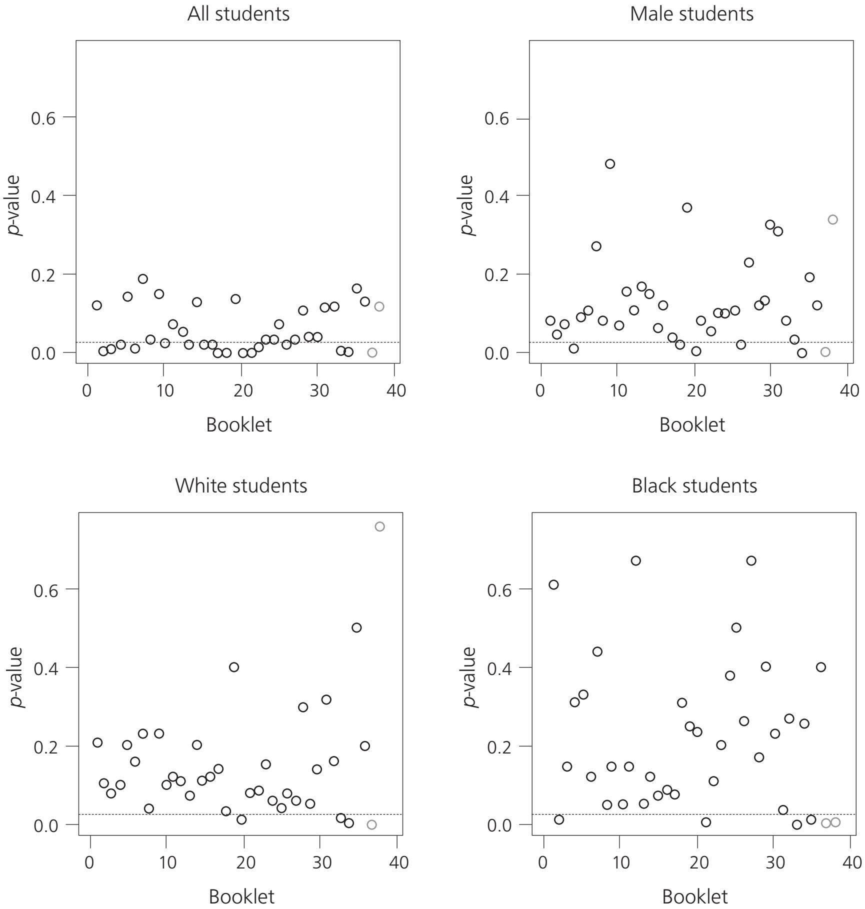
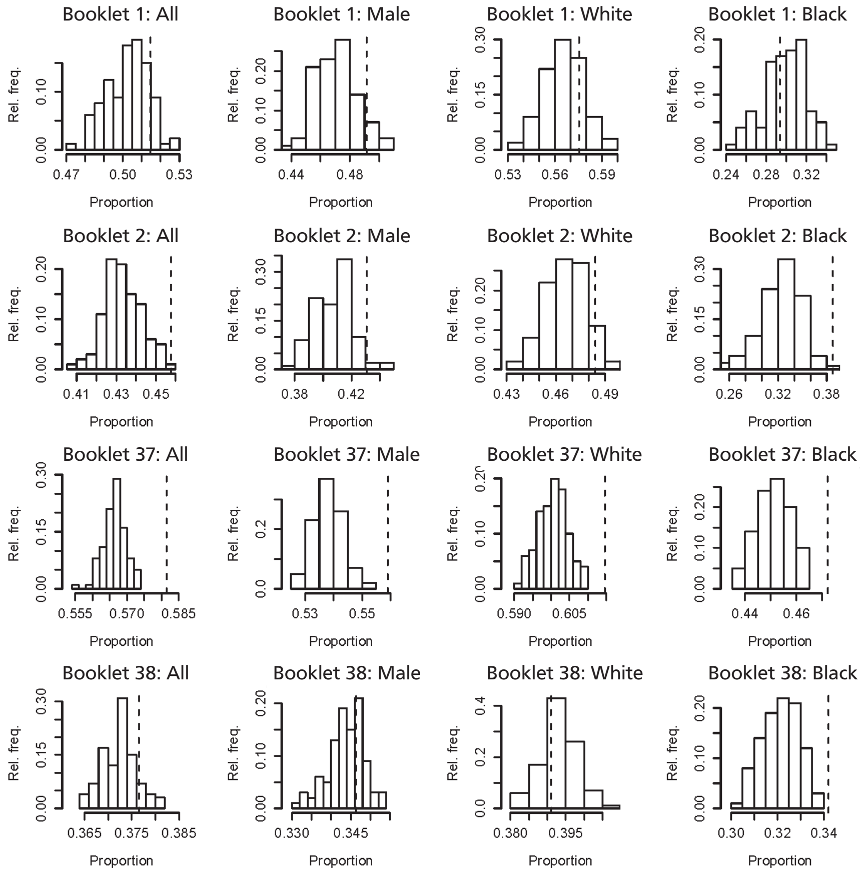


Figure 6 shows the observed value and the predicted value of the average score statistic for four booklets (1, 2, 37, and 38) for all the students, male students, White students, and Black students. Each row corresponds to a booklet and has four panels. In each panel, the histogram denotes the predicted value, and the vertical dashed line denotes the observed value. There are some differences in the figure in the observed and predicted values of the test statistic, especially for Booklets 2 and 37. However, the magnitude of these differences is not too large, even for Booklet 37, where we found the largest differences. For example, for Booklet 37, for all students (the first panel in the third row in Figure 6), the observed value of the average score statistic is about .58, while the mean of the predicted values is approximately .57. As such, the differences between the observed and predicted values, while statistically significant, are likely to have little practical significance. Further research to study the practical significance of these differences would be beneficial.

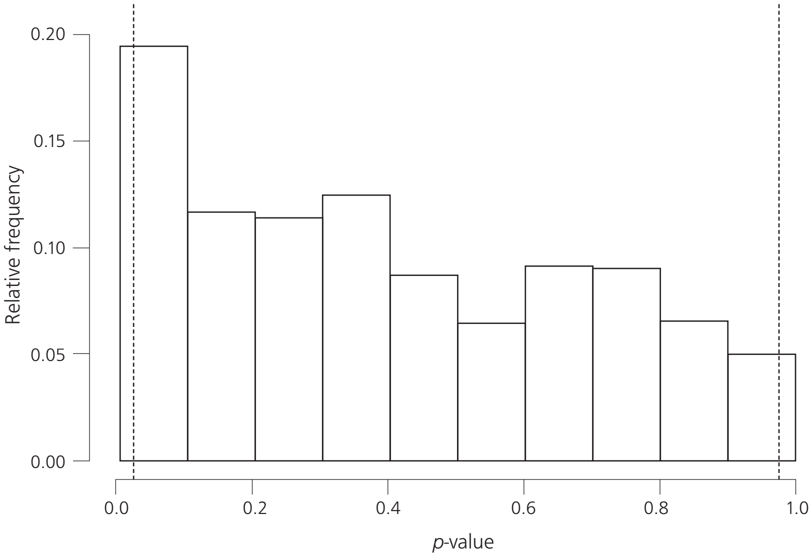
Figure 6: The observed and the predicted values of the average group score statistic for Booklets 1, 2, 37, and 38 for the 2002 NAEP reading data set



5.2 Average Item Score

Figure 7 presents all 744 *p*-values for the average item score statistic. Note the two vertical lines drawn at values .025 and .975. Note also that an item which appeared in two different booklets is treated as two different items and so has two *p*-values associated with it. The figure shows that just over half of the *p*-values were less than .5. The percentage of *p*-values greater than .975 or is less than .025 is 9, not much more than the nominal level.

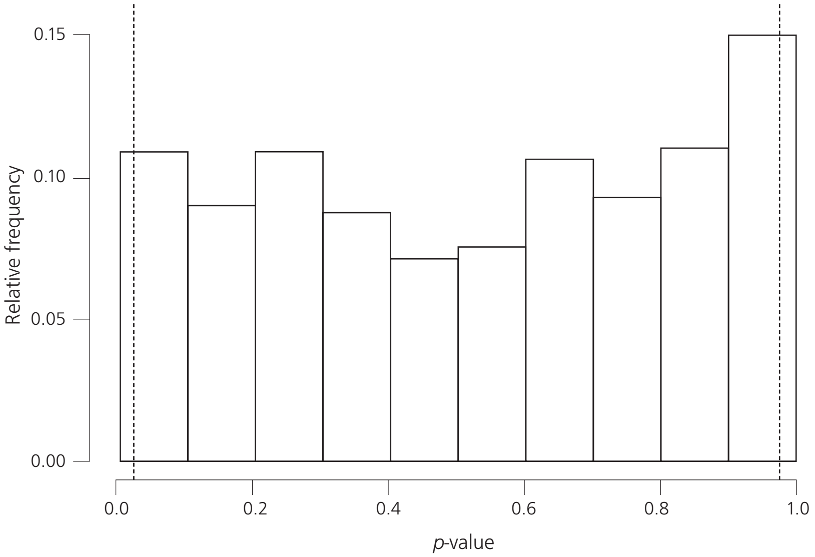
Figure 7: The *p*-values for the average item score statistic for the 2002 NAEP reading data set



5.3 Biserial Correlation

Figure 8 presents all 744 *p*-values for the biserial correlation. Note the two vertical lines drawn at values .025 and .975. Note also that an item which appeared in two different booklets was treated as two different items and so has two *p*-values associated with it. The figure shows only a few extreme *p*-values for the biserial correlation, which means that the NAEP model adequately predicted the statistic. The percentage of *p*-values greater than .975 or less than .025 is 10, not much more than the nominal level.

Figure 8: The *p*-values for the biserial correlations for the 2002 NAEP reading data set



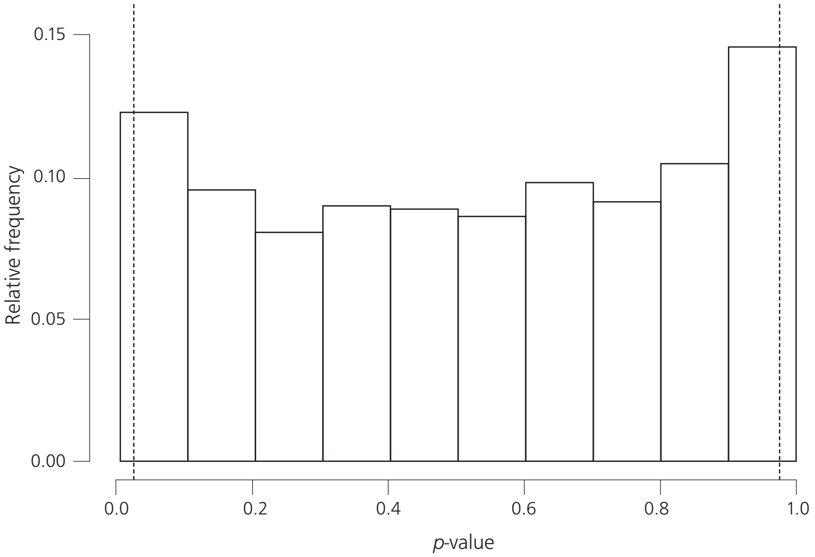
5.4 Item-pair Correlation

Figure 9 shows all 13,784 *p*-values for the item-pair correlation statistic. Note the two vertical lines drawn at values .025 and .975. Note also that an item which appeared in two different booklets was treated as two different items. The figure shows only a few extreme *p*-values for the item-pair correlations. The percentage of *p*-values greater than .975 or less than .025 is 9, not much more than the nominal level.

5.5 Discussion of the Results from the 2002 NAEP Reading Data Set

The results show that the LRM employed in NAEP adequately predicted the average item score, the biserial correlation, and the item-pair correlation. The model did not appear to adequately predict the average group scores of the students; it often underpredicted these scores. However, because the differences between the observed scores and the predicted scores seemed negligible, they were probably not practically significant. Overall, the model adequately predicted several summaries of the NAEP data. We can therefore conclude that the NAEP operational model was adequate for the NAEP data analyzed in this study.

Figure 9: The *p*-values for the item-pair correlations for the 2002 NAEP reading data set



6. CONCLUSIONS

To ensure quality control and overall improvement of the NAEP statistical analysis, it is important to frequently assess the fit of the NAEP statistical model. The task is far from straightforward, given the complex nature of the NAEP statistical model and estimation procedure.

As documented in this article, we applied a simulation-based model fit procedure to NAEP data to investigate whether the LRM employed in this assessment adequately predicts basic statistical summaries, such as the average group scores. Our suggested procedure is easily understood. Also, because it permits use of existing NAEP software, operational implementation of the procedure is simple. Analysis of a real data set provided us with some evidence of a misfit of the NAEP model. However, the magnitude of the misfit was small, which means that the misfit probably had no practical significance.

We found that the distribution of the *p*-values for the average group score and the average item score statistics under the null model were non-uniform and not centered around .5; the model seems to under-predict these quantities. We do not have an explanation for this phenomenon as yet and intend to conduct further research on this potential issue. It is possible that the phenomenon may be associated with how NAEP generates plausible values, or with the discrepancy between the scaling stage (where the ability parameters are assumed to follow independent univariate distributions) and the conditioning stage (where the ability parameters are assumed to follow a multivariate regression model) of NAEP estimation.

Another possible reason could be associated with the inclusion of several hundred principal components as covariates in the latent regression model. The smaller eigenvalues associated with the principal components were usually very small, which may have contributed to some level of instability of these components. This issue is likely to be a particular concern whenever large numbers of eigenvalues and associated principal components are involved. Anderson's (1963) asymptotic theory for principal component analysis and Krzanowski's (1987) jackknife-based standard errors for eigenvalues in principal component analysis may prove useful when assumptions necessary for deriving asymptotic results are not met.

It is not uncommon for simulation-based *p*-values to have non-uniform null distributions. Von Davier (1997) found that the bootstrap-based null distribution is non-uniform for each of two goodness-of-fit statistics. Researchers such as Sinharay and Stern (2003) and Sinharay et al. (2006) found PPMC-based *p*-values have non-uniform null distributions. Conceptually, it is possible to apply a double-simulation procedure to calibrate non-uniform *p*-values in order to obtain calibrated *p*-values that follow a uniform distribution.[6] However, this approach is too time-consuming if applied to a typical NAEP data set.

We would like to see several related issues studied in the future. In our limited examination, we found some evidence of misfit (which means that the procedure has some power), and we found the Type I error rate to be satisfactory. However, we consider it necessary to conduct a more detailed study of Type I error rate and the power of the suggested procedure. We also found greater misfit for the average group score statistic for the last two booklets, which were given to several thousand students, in contrast to the first 36 booklets, which were given to a few hundred. A possible reason for the misfit is the greater power of model fit measures for larger samples. Another reason could be that the last two booklets have one long block each, whereas the first 36 booklets have two short blocks each.

It is possible to examine raw-score-based graphical item-fit analyses, such as that conducted by Sinharay (2006). Because we examined booklet-level statistics only, it may be informative to study test statistics that combine information from several booklets. We could, for example, obtain an overall average item score by combining information across booklets and a weighted average of booklet averages for a group. Also, because NAEP reports the percentage of students at or above different performance levels (e.g., *basic*, *proficient*, etc.), it would be helpful to focus on a statistic related to percentages. Running an MCMC algorithm and then employing the PPMC method (Gelman et al., 2003) to assess the fit of the NAEP model could be another avenue of future research, especially given relatively recent work on an MCMC algorithm for fitting the NAEP model (see, for example, Johnson & Jenkins, 2004).

[9] Hjort, Dahl, and Steinbakk (2006), for example, obtained uniformly distributed calibrated PPMC-based *p*-values.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2000). *The 1998 NAEP technical report.* Washington, DC: U. S. Department of Education.

American Psychological Association, National Council on Measurement in Education & American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Anderson, T. W. (1963). Asymptotic theory for principal components analysis. *Annals of Mathematical Statistics*, *34*, 122−148.

Beaton, A. (1987). *The NAEP 1983−84 technical report*. Princeton, NJ: Educational Testing Service.

Beaton, A. (2003). *A procedure for testing the fit of IRT models for special populations: Draft.* Unpublished manuscript.

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

Dresher, A. R., & Thind, S. K. (2007, April). *Examination of item fit for individual jurisdictions in NAEP*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, USA.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York: Chapman & Hall.

Hjort, N. L., Dahl, F. A., & Steinbakk, G. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*, 1157−1174.

Johnson, M. S., & Jenkins, F. (2004). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress* (ETS Research Report No. RR-04-38). Princeton, NJ: Educational Testing Service.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (ETS Research Report No. RR-01-25). Princeton, NJ: Educational Testing Service.

Krzanowski, W. J. (1987). Cross-validation in principal component analysis. *Biometrics*, *43*, 575−584.

Li, J. (2005). *The effect of accommodations for students with disabilities: An item fit analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

Martin, M. O., & Kelly, D. L. (1996). *TIMSS technical report: Vol. I. Design and development.* Chestnut Hill, MA: Boston College.

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

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

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools.* Chestnut Hill, MA: Boston College.

Rogers, A., Gregory, K., Davis, S., & Kulick, E. (2006). *User's guide to NAEP model-based p-value programs*. Unpublished manuscript. Princeton, NJ: ETS.

Rogers, A., Tang, C., Lin, M.-J. & Kandathil, M. (2006). *DGROUP* (computer software). Princeton, NJ: Educational Testing Service.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355–371.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *4*2, 375–394.

Sinharay, S. (2006). Bayesian item fit analysis for dichotomous item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 429–449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321.

Sinharay, S., & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, *111*, 209–221.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*(1), 58–75.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*(3), 309–322.

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

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research, 2*(2), 29–48.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Marginal estimation of population characteristics: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Vol. 1. Psychometrics* (pp. 1039–1055). Amsterdam, The Netherlands: Elsevier.