Research Report

Assessing Fit of Latent Regression Models

Sandip Sinharay

Zhumei Guo

Matthias von Davier

Bernard P. Veldkamp

December 2009

ETS RR-09-50

Assessing Fit of Latent Regression Models

Sandip Sinharay, Zhumei Guo, and Matthias von Davier
ETS, Princeton, New Jersey

Bernard P. Veldkamp
University of Twente, The Netherlands

December 2009

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

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

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

Abstract

The reporting methods used in large-scale educational assessments such as the National Assessment of Educational Progress (NAEP) rely on a *latent regression model*. There is a lack of research on the assessment of fit of latent regression models. This paper suggests a simulation-based model-fit technique to assess the fit of such models. The technique consists of investigating whether basic statistical summaries are predicted adequately by the latent regression model. Application of the suggested technique to an operational NAEP data set reveals important information regarding the fit of the latent regression model to the data.

Key words: Bootstrap, NAEP, p-value

Acknowledgments

The work was completed while Bernard P. Veldkamp was a visiting scholar at ETS. The authors thank Andreas Oranje, Shelby Haberman, Amy Dresher, John Donoghue, Catherine McClellan, Steve Isham, Frank Rijmen, and Dan Eignor for useful advice, and Amanda MacBride for help with copy-editing the manuscript.

1 Introduction

The National Assessment of Educational Progress (NAEP), the only regularly administered and congressionally mandated national assessment program (see, e.g., Beaton & Zwick, 1992), is an ongoing survey of the academic achievement of U.S. school students in a number of subject areas such as reading, writing, and mathematics. Since 1984, NAEP reporting methods have used a multilevel statistical model consisting of two components: (a) an item response theory (IRT) component and (b) a linear regression component (see, e.g., Beaton, 1987; Mislevy, Johnson, & Muraki, 1992). 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) also adopted essentially the same model. This model is often referred to as a *latent regression model* (LRM). An algorithm for estimating the parameters of this model is implemented in the DGROUP set of programs (Rogers, Tang, Lin, & Kandathil, 2006), which is an ETS product.

Standard 3.9 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, [AERA, APA, & NCME], 1999) requires evidence of model fit when an IRT model is used to make inferences from a data set. Hence, it is important to assess the fit of the LRM used by NAEP to ensure quality control and an overall improvement in the long run. Although some model-checking techniques have been applied to the NAEP model (e.g., Beaton, 2003; Dresher & Thind, 2007; Li, 2005), there is a substantial scope of further work on the topic.

This paper suggests a simulation-based procedure to assess the goodness of fit of the LRM used by NAEP and other large-scale assessments mentioned above. The procedure consists in investigating whether several summary statistics of the data are predicted adequately by the model. Similar to the parametric bootstrap (e.g., Efron & Tibshirani, 1993) and the posterior predictive model-checking method (e.g., Gelman, Carlin, Stern, & Rubin, 2003), our procedure generates predicted data sets under the model using the NAEP operational software and compares several aspects of the predicted data sets to those of the

observed data sets.

Section 2 gives some background, describing the current NAEP statistical model and estimation procedure, and the existing NAEP model-checking procedures. Section 3 describes our suggested model checks. Section 4 provides a real data example. Section 5 studies the Type I error rates of the method suggested. Section 6 gives conclusions and suggestions for future work.

2 The NAEP Latent Regression Model and Estimation

2.1 The Model

Assume that the unique p-dimensional latent proficiency variable for examinee i is $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots \theta_{ip})'$. In NAEP, p could be between 1 and 5. Let us denote the response vector to the test items for examinee i as $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{ip})$, where, $\boldsymbol{y}_{ik}$, a vector of responses, contributes information about θ_{ik}. For example, $\boldsymbol{y}_{ik}$ could be responses of examinee i to algebra questions in a mathematics test and θ_{ik} the algebra skill variable of the examinee. The likelihood for an examinee is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}_i) = \prod_{k=1}^{p} f_1(\boldsymbol{y}_{ik}|\theta_{ik}) \equiv L(\boldsymbol{\theta}_i; \boldsymbol{y}_i)\cdot \tag{1}$$

The expressions $f_1(\boldsymbol{y}_{iq}|\theta_{iq})$ in (1) is a product of terms 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 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots x_{im})$ are m fully measured demographic and educational characteristics for the examinee. Conditional on $\boldsymbol{x}_i$, the examinee proficiency vector $\boldsymbol{\theta}_i$ is assumed to follow a multivariate normal distribution, that is,

$$\boldsymbol{\theta}_i|\boldsymbol{x}_i \sim N(\boldsymbol{\Gamma}'\boldsymbol{x}_i, \boldsymbol{\Sigma}). \tag{2}$$

Together, (1) and (2) form the LRM or *conditioning model* employed in NAEP. For further details, see, for example, von Davier, Sinharay, Oranje, and Beaton (2006).

2.2 Estimation

NAEP uses a three-stage estimation process for fitting the previously discussed LRM to the data:

1. The first stage, *scaling*, fits the model given by (1) to the examinee response data and estimates the item parameters using the PARSCALE software (Allen, Donoghue, & Schoeps, 2000). The prior distributions on the components of the examinee proficiency are assumed to be independent discrete univariate distributions in this stage.

2. The second stage, *conditioning*, assumes that the item parameters are fixed at the estimates found in the scaling stage and fits the model given by (1) and (2) to the data, and estimates $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$. The following versions of the DGROUP program perform the conditioning step differently:

   - BGROUP (Beaton, 1987), employed when $p \leq 2$, uses numerical quadrature.
   - CGROUP (Thomas, 1993), employed when $p > 2$, uses Laplace approximations.
   - NGROUP (Mislevy, 1985), used to find the starting values for BGROUP or CGROUP, uses a normal approximation of $L(\boldsymbol{\theta}_i; \boldsymbol{y}_i)$.

3. The third stage of the NAEP estimation process generates *plausible values*, which are imputed values of the proficiency variables, for all the examinees using the parameter estimates obtained from the scaling and conditioning stages (the plausible values are used to estimate examinee subgroup averages). The third stage also estimates the variances corresponding to the examinee subgroup averages as the sum of two terms: the variance due to the latency of $\boldsymbol{\theta}_i$s 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 the terms involve the plausible values generated in the conditioning step.

2.3 Existing Works on Assessing Fit of the NAEP Model

NAEP rigorously monitors data quality and employs a number of qualitative checks of the results of their statistical analysis. As first-level checks, NAEP employs several plausibility analysis (which involves examining the computer outputs to make sure that they make sense) and computer-based checks at different stages of statistical analysis; these ensure that the data-analysis process is working as intended. The first-level checks include going through several carefully designed checklists, such as an item analysis checklist

and a CGROUP/BGROUP conditioning checklist. However, it can be argued that these first-level checks provide quality control measures that are necessary but not sufficient. That is, even if they reveal no problems and the programs are running as expected on the appropriate data sets, the inferences may be problematic if the model does not explain the data adequately.

Therefore, as second-level checks, additional steps are taken to check the appropriateness and quality of the IRT model (Allen et al., 2002, p. 233). The item parameter estimates are examined—extreme estimates often indicate problems. Differential item functioning (DIF) analyses are used to examine issues of multidimensionality (see, for example, Roussos & Stout, 1996, for the connection between DIF and multidimensionality). NAEP operational analyses also employ graphical item fit analyses using residual plots and a related χ^2-type item fit statistic (Allen et al., p. 233) that provide guidelines for treating 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 acknowledged by Allen et al. (p. 233). Another second-level check used in NAEP operational analyses is the comparison of observed and model-predicted proportions of examinees obtaining a particular score on an item (Rogers, Gregory, Davis, & Kulick, 2006); these analyses, however, do not use the variability of the quantities involved. It will 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 addresses this issue.

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

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

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). Then, a bootstrap method is used to determine the null distribution of these statistics. Li (2005) employed Beaton's statistics

to operational test data to determine the effect of accommodations on students with disabilities. Dresher and Thind (2007) employed Beaton's statistics to 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 obtained the null distribution of the statistic from its values for one simulated data set. However, these methods have their limitations. For example, Sinharay (2005, p. 379) argued that fit statistics based on examinee-level residuals are unreliable because of their excessive variability, a limitation that applies to Beaton's statistics.

In a practical application of model fit analysis, it is important to examine the appropriate aspects of the model using appropriate test statistics. Standard recommendation (see, e.g., Gelman et al., 2003, p. 172) on this issue is that model checking in an application should focus on aspects of the model that are relevant to the purposes to which the inference will be applied. For example, if one is interested in estimating the mean income of a population using a statistical model, the model fit analysis should focus on the mean. However, there has been little focus on this issue, both in the context of IRT model fit in general (e.g., Sinharay, 2005) and in the context of NAEP.

Thus, there is substantial scope of further work on assessing the fit of LRMs to NAEP data. Note that such work has to properly take into account the idiosyncrasies of the NAEP model and data such as the matrix sampling (that refers to the fact that each examinee sees only some of the questions), sampling weights, and missing values.

3 Methods

This study applies a simulation-based model fit technique to NAEP statistical analysis to investigate whether several data summaries (or *test statistics*) are predicted adequately by the LRM employed in NAEP.

3.1 Description of the Suggested Technique

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. Hence, we use a simulation-based method, which makes use of existing NAEP

programs, to determine the null distribution of the test statistics. Plausible values are generated as in NAEP operational procedures using the DGROUP program. The operational NAEP analysis generates five plausible values for each candidate, but we generate 200 plausible values each. These are like draws of $\boldsymbol{\theta}_i$ from its posterior distribution. Using the generated plausible values and the item parameters estimated in the calibration step, 200 predicted data sets are generated from the model given by Equations 1 and 2. Values of the test statistics are computed for each of these predicted data sets. The flowchart in Figure 1 shows the procedure for the average raw score statistic that will be described shortly. These values of the statistics (which can be considered to have been predicted under the model) are compared to the corresponding observed values 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 misfit of the model. The comparison of the observed and predicted values of the statistics was performed graphically by plotting the observed and predicted values of the statistics: an observed value that lies at the tail of the distribution of the predicted values would indicate that the corresponding statistic is not predicted adequately by the model. In addition, we computed p-values for the statistics; a p-value is the proportion of the predicted values of a statistic that is larger than the corresponding observed value. A very low or very high p-value would indicate that the corresponding statistic is not predicted adequately by the model.

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

1. Generating a sample (mostly using a Markov chain Monte Carlo method; Gelman et al., 2003) from the joint posterior distribution of the model parameters
2. Simulating data sets using the generated parameter values
3. Comparing the observed value of a test statistic with its values computed from the generated data sets

Our suggested method performs the last two of these three steps and hence is similar to the PPMC method. However, it is an approximation of the latter because it performs only

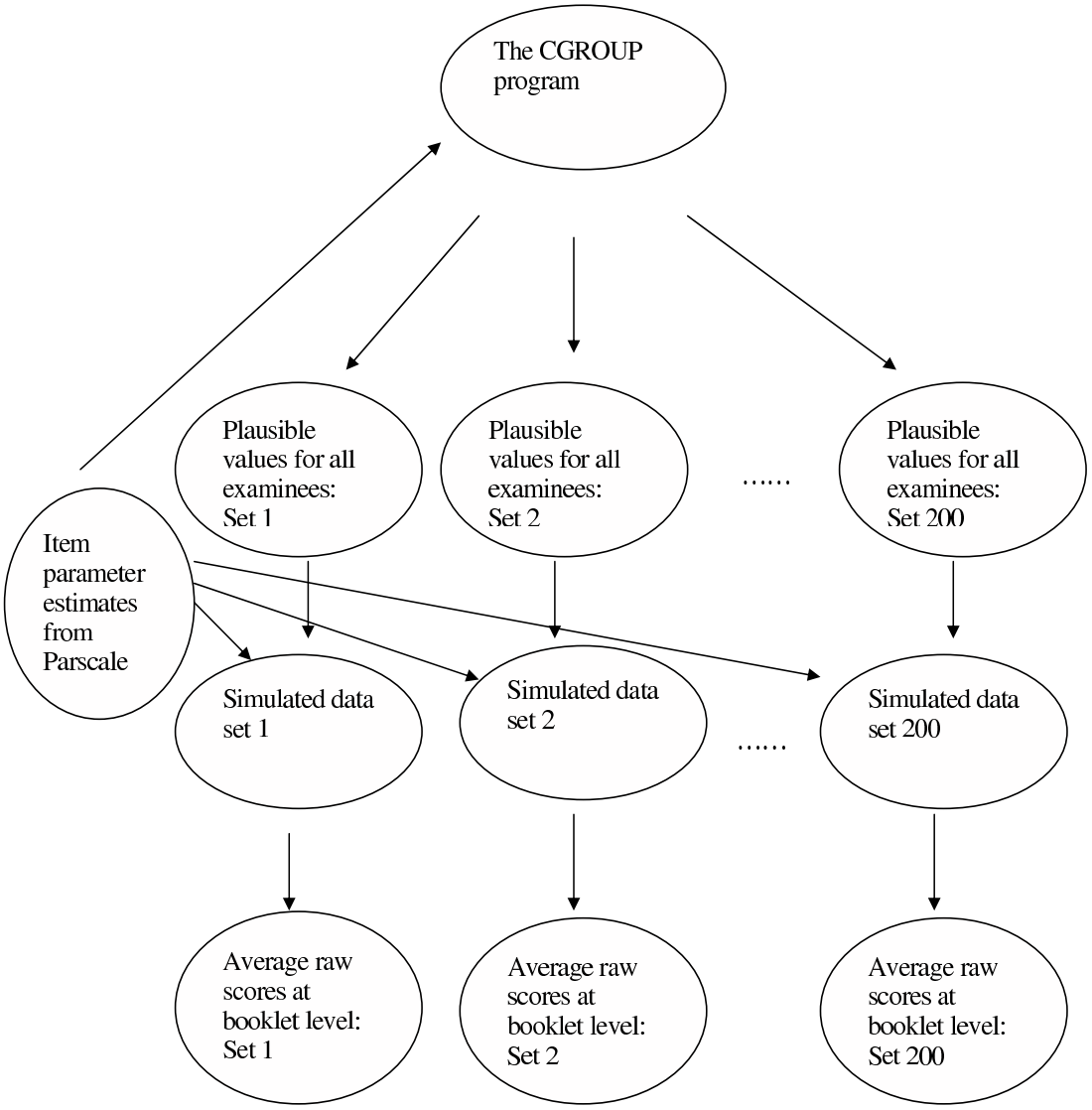

***Figure* 1. A flowchart showing the steps of the simulation procedure to determine the null distribution of the average raw scores. The item-parameter estimates from PARSCALE are used in the CGROUP version of the DGROUP program, which generates 200 sets of plausible values. The plausible values and the item-parameter estimates will be used to generate 200 simulated data sets, which will result in 200 sets of average raw scores for each booklet for each examinee subgroup. The observed average raw scores will then be compared to the corresponding 200 simulated values.**

part of the first step involved in a PPMC: it draws plausible values (which are approximate draws from the examinee posterior distribution), but it does not draw item parameter values. The PPMC method has been successfully used to detect misfit of simple IRT models by Sinharay, Johnson, and Stern (2006) and Sinharay (2005), who applied some fit statistics similar to those in this paper.

The suggested technique 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).

Our suggested method is simple to understand, as it is similar to two popular model-checking methods. In addition, it uses existing NAEP software so that operational implementation of the procedure will be straightforward.

3.2 Description of the Test Statistics

This paper examined if the LRM used in NAEP adequately predicts the test statistics listed following this paragraph. All of these statistics are computed for each booklet separately. Researchers van der Linden and Hambleton (1997, p. 16) recommended the collection of 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-parameters IRT models. Sinharay (2005) and Sinharay et al. (2006) followed the recommendation to assess the fit of simple IRT models using a variety of simple summaries of the data, which are similar to the ones listed in this section. The recommendation of van der Linden and Hambleton is equally appropriate for any IRT model, including the LRM employed in NAEP. Our suggested method, together with the statistics described here, provides a tool kit that can be used to collect a variety of evidence about the fit of the LRM to NAEP data or other similar data.

- Average raw score: Let Y_{ij} denote the response of the i-th examinee to the j-th item in a booklet. For a dichotomous item, Y_{ij} is 0 or 1, while for a $k-$category polytomous item, Y_{ij} takes one among the values 0, 1, $\ldots k-1$. NAEP encounters a substantial percentage of omitted and not-reached responses. In NAEP, not-reached items are treated as not-presented items, and an omitted response is assigned a fractional score

equal to the reciprocal of the number of options in a multiple-choice item and assigned the score for the lowest scoring category for a constructed-response item (Allen et al., 2000, pp. 231-232). Hence, the following variation of the average raw score statistic is used.

$$A_s = \frac{\sum_{i \in s} W_i \sum_j Y_{ij}/R_i}{\sum_{i \in s} W_i},$$

where s denotes a subgroup and R_i is the sum of the maximum raw score points on the items that the i-th examinee reached. The statistic A_s denotes the average proportion score in a booklet for the s-th subgroup. Note that if examinee 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. As argued earlier, this statistic is related to the examinee subgroup means.

- 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 is closely related to the statistic of interest in Rogers, Gregory, Davis, & Kulick (2006), the main difference being that p_j is defined for a booklet.

- Biserial correlation coefficients: Because of the way NAEP operational analysis treats the omitted and the not-reached items, the standard definition of the biserial correlation is not appropriate here. Hence, for each item in a booklet, we compute the correlation between the item response vector and the vector of proportion correct scores $\sum_j Y_{ij}/R_i$ (using notations introduced earlier) using the sampling weights W_i in the computations.

- Item pair correlation: This is the correlation between the response vectors for two items, where the sampling weights W_i were used in the computations.

We chose the above mentioned statistics because they are simple data summaries and most of them were found to be useful in research by Sinharay (2005) and Sinharay et al. (2006). The average raw score statistic deserves special mention. Ideally, model-checking in an application should focus on aspects of the model that are relevant to the purposes to which the inference will be applied (Gelman et al., 2003, p. 172). The quantities of primary

interest in NAEP are the mean scale scores for the different subgroups, so it is necessary to determine if these quantities are estimated adequately. Ideally, one would like to compare the observed value of a test statistic based on the mean scale scores to the model-predicted values of the test statistic. However, the mean scale scores are functions of the unobserved examinee proficiency variables, so it is impossible to obtain a test statistic based on these that will have an observed value. The average raw scores in examinee subgroups are observed proxies of the unobserved mean scale scores; these raw scores, while simple to compute, are expected to be highly correlated with the mean scale scores. If the NAEP model predicts the average raw scores for examinee subgroups accurately for the data sets under consideration, one should be confident that the subgroup estimates provided by the NAEP model are accurate.

4 Analysis of Data From 2002 NAEP Reading Assessment

4.1 The Data

The NAEP 2002 Reading assessment grade-12 data, with about 15,000 examinees, was used for this study. Primary reasons for choosing the reading assessment are that reading is a No Child Left Behind subject, and, historically, reading items have been more likely to display problematic item fit than mathematics (another No Child Left Behind subject) items. The reading assessment measures three skills: (a) reading for literary experience, (b) reading for information, and (c) reading to perform a task. The reading assessment had 38 booklets. The first 36 booklets were given to a few hundred students each, while the remaining two booklets were given to a few thousand students each. Another difference is that each of the remaining two booklets consists of one long block (out of a total of two long blocks) of items, whereas each of the first 36 booklets consists of two shorter blocks (out of a total of nine short blocks) of items. The number of items in a booklet is around 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 remaining booklet. About 50% of all students taking the reading assessment were male, about 65% were White, and about 15% were Black. The proportion of omitted and not-reached responses ranged from 4% to 10% for the different booklets.

4.2 The procedure

The data analysis consisted of the following steps:

1. Calculating the test statistics described in Section 3.2 from the original data set
2. Estimating the item parameters using the NAEP PARSCALE program
3. Running the CGROUP version of the DGROUP program to fit the LRM to the data and to generate 200 plausible values
4. Simulating 200 data sets using the plausible values generated in the third step and the estimated item parameters obtained in the second step
5. Calculating the test statistics for each of the 200 simulated data sets (resulting in 200 *predicted values* for each statistic) and comparing the results with the corresponding *observed value*s of the test statistics computed from the original data set using graphical plots and p-values.

4.3 Results

4.3.1 Average raw scores

Figure 2 shows the p-values for the average raw-score statistic. The four panels show results for all the examinees, for male students, for White students, and for Black students. Each panel shows 38 points, each point denoting the p-value for a booklet. A horizontal dashed line in each panel denotes the value of 0.025; a p-value below that will indicate that the predicted values of the statistic are significantly lower than the corresponding observed value. The range of the Y-axis is the same in all the four panels.

The figure shows the following:

- There is some evidence of misfit for all students (top left panel), with about half of the p-values lying below 0.025.
- In all the panels, the p-values lie mostly below 0.5, which indicates that the predicted values are mostly lower than the observed value of the statistic.

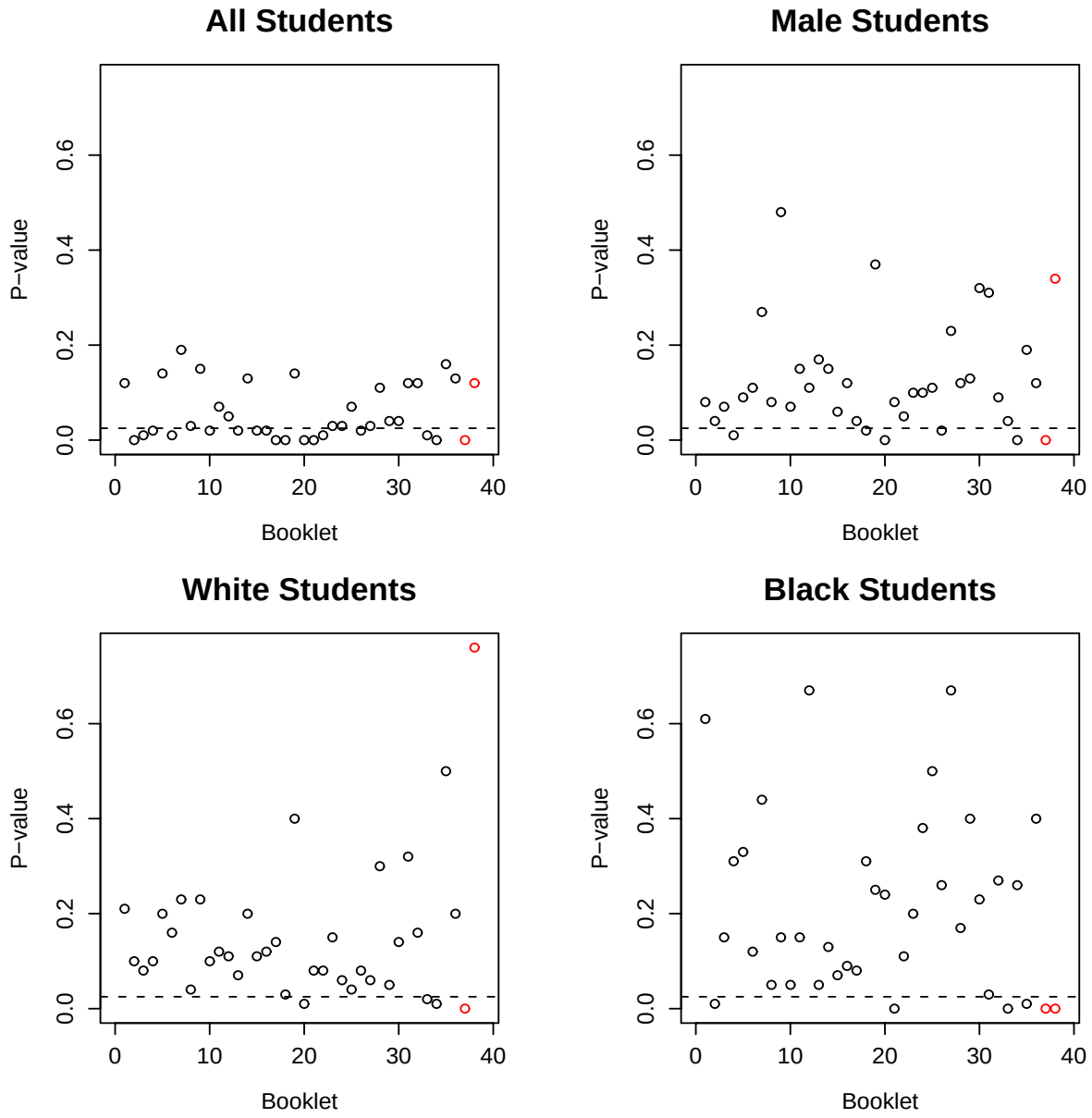

***Figure* 2. The p-values for the average score statistics for all students, male students, white students, and black students for all booklets.**

- The plots for the male, White, and Black students do not show overwhelming evidence of misfit.

Figure 3 shows the observed value and predicted value of the average score statistic for four booklets (1, 2, 37, and 38) for all the examinees, for male students, for White students, and for Black students. Each row corresponds to a booklet and has four panels. In any panel, a histogram denotes the predicted values and a vertical dashed line denotes the observed value.

There are some differences in observed and predicted values of the test statistic, especially for Booklets 2 and 37. For Booklet 37, the observed values are significantly different from the corresponding predicted values. However, the magnitude of the difference is not too large, even for this booklet. For example, for all students (the first panel in the third row in Figure 3), the observed value of the average score statistic is about 0.58 while the mean of the predicted values is approximately 0.57. So the differences between the observed and predicted values, while statistically significant, are most likely not practically significant. However, further research is required to study the practical significance of these differences.

4.3.2 Average Item Score

Figure 4 shows all the p-values for the statistic. Two vertical lines are drawn at values 0.025 and 0.975. If an item appears in two different booklets, it is treated as two different items and has two p-values associated with it.

The figure shows that more than half of the p-values are less than 0.5. The percentage of p-values that are larger than 0.975 or smaller than 0.025 is 9–not much larger than the nominal level.

4.3.3 Biserial Correlation

Figure 5 shows all the p-values for the biserial correlation. Two vertical lines are drawn at values 0.025 and 0.975. If an item appears in two different booklets, it is treated as two different items and has two p-values associated with it.

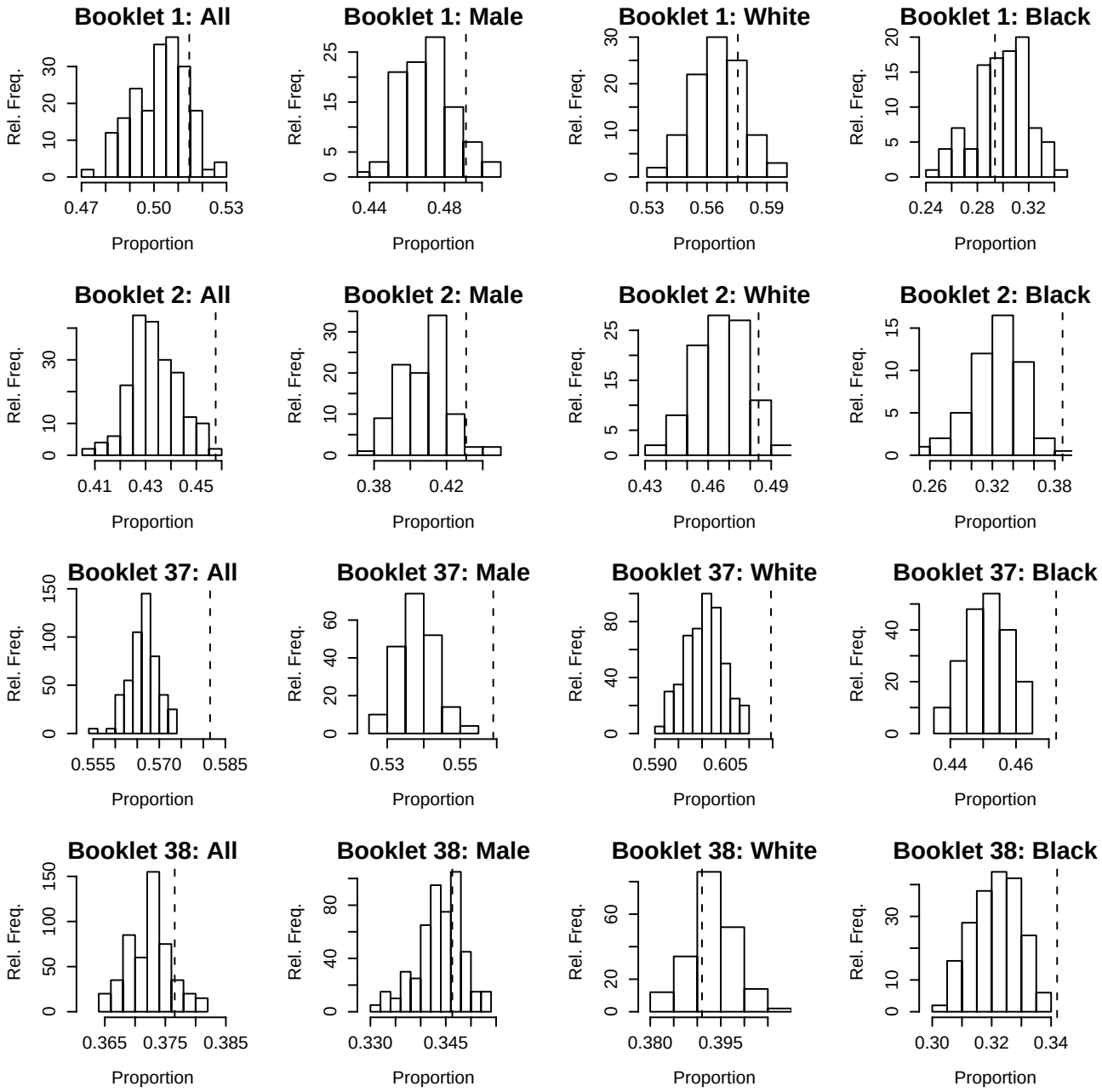

***Figure* 3. The observed and predicted values of the average score statistics for all students, male students, White students and Black students for Booklets 1, 2, 37, and 38.**

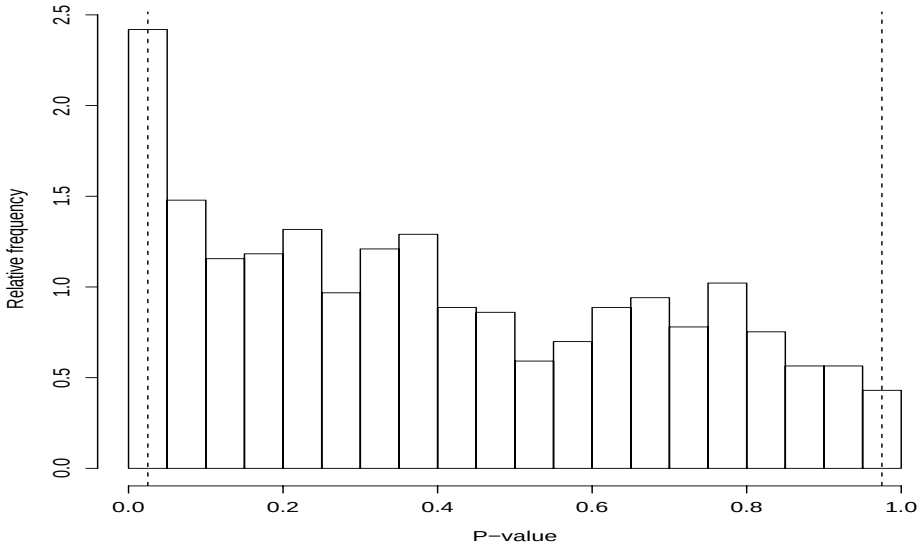

***Figure* 4. The p-values for the average item score statistic.**

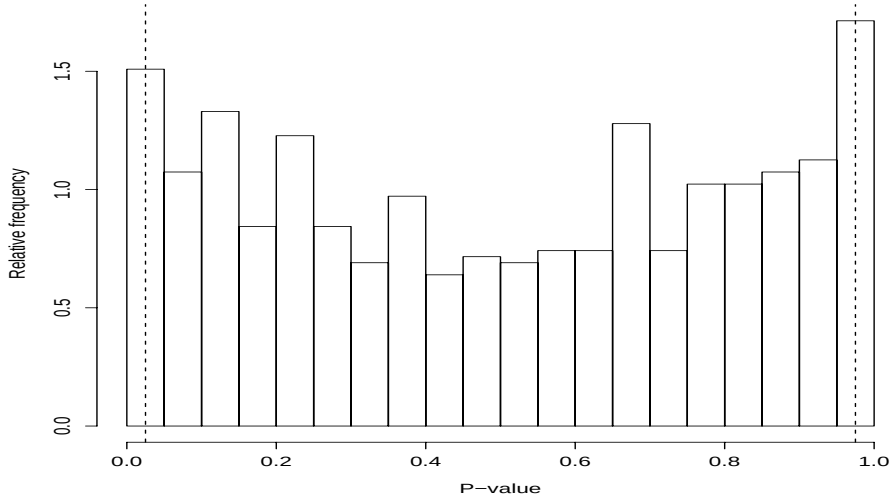

***Figure* 5. The p-values for the biserial correlations.**

The figure shows that there are not too many extreme p-values for the biserial correlation—that is exactly what is expected when a model predicts a statistic adequately. The percentage of p-values that are larger than 0.975 or smaller than 0.025 is 10–not much larger than the nominal level.

4.3.4 Item-Pair Correlation

Figure 6 shows all the p-values for the item-pair correlation statistic. Two vertical lines are drawn at values 0.025 and 0.975. If an item appears in two different booklets, it is treated as two different items.

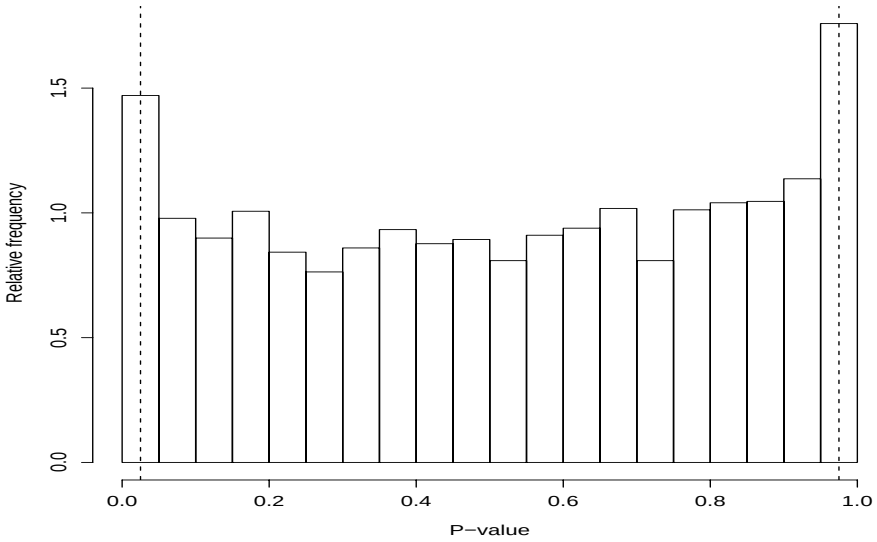

***Figure* 6. The p-values for the item-pair correlations.**

The figure shows that there are not too many extreme p-values for the item-pair correlations. The percentage of p-values that are larger than 0.975 or smaller than 0.025 is 9–not much larger than the nominal level.

4.3.5 Discussion

We find the LRM employed in NAEP to adequately predict the average item score, the biserial correlation, and item-pair correlation. The model does not seem to predict the average raw scores of the students adequately. It often underpredicts the scores; however, the differences between the observed and predicted scores seem to be rather small so that the differences are most likely not practically significant. Overall, as the model is found to adequately predict several summaries of the NAEP data, we are quite confident that the model is adequate for the data.

5 Studying the Type I Error Rate of the Suggested Method

It is important to study the Type I error rate of any statistic used for model checking. Our suggested technique is similar to the bootstrap method (Efron & Tibshirani, 1993) and the posterior predictive model-checking method (Gelman et al., 2003). Both of these methods are expected to have Type I error rates close to the nominal level. However, we still perform a limited study to make sure that the Type I error rate of our suggested method is not too high. To perform this study, we need a data set generated from the model; The earlier section described the procedure to generate several data sets that can be considered to have come from the NAEP model, that is, they are data sets that we would have seen if the NAEP model were true. We replaced the item responses in the observed data set by those in the first generated data set and repeated the whole analysis (which included running PARSCALE, DGROUP, simulating 200 data sets, and computing the test statistics) described earlier. The results from applying our method are described in the following section for different test statistics.

5.1 Average Raw Scores

Figure 7 shows the p-values for the average raw-score statistic from all the 38 booklets from all students, from male students, from White students, and from Black students. The proportion of significant p-values in Figure 7 is not much higher than the nominal level of 5%, which points to the respectable Type I error rate of our suggested method.

However, Figure 7 shows some patterns similar to those in Figure 2. For example, all p-values are below 0.6 for all students, and there are no p-values larger than 0.85 in any of the four panels. Ideally, under the null model, the p-values for a good-fit statistic should be distributed uniformly between 0 and 1, which is not the case here. Further research might be done to explore this phenomenon.

5.2 Average Item Score

Figure 8 shows all the p-values for the average item-score statistic. The percentage of p-values that are larger than 0.975 or smaller than 0.025 is 5, the same as the nominal

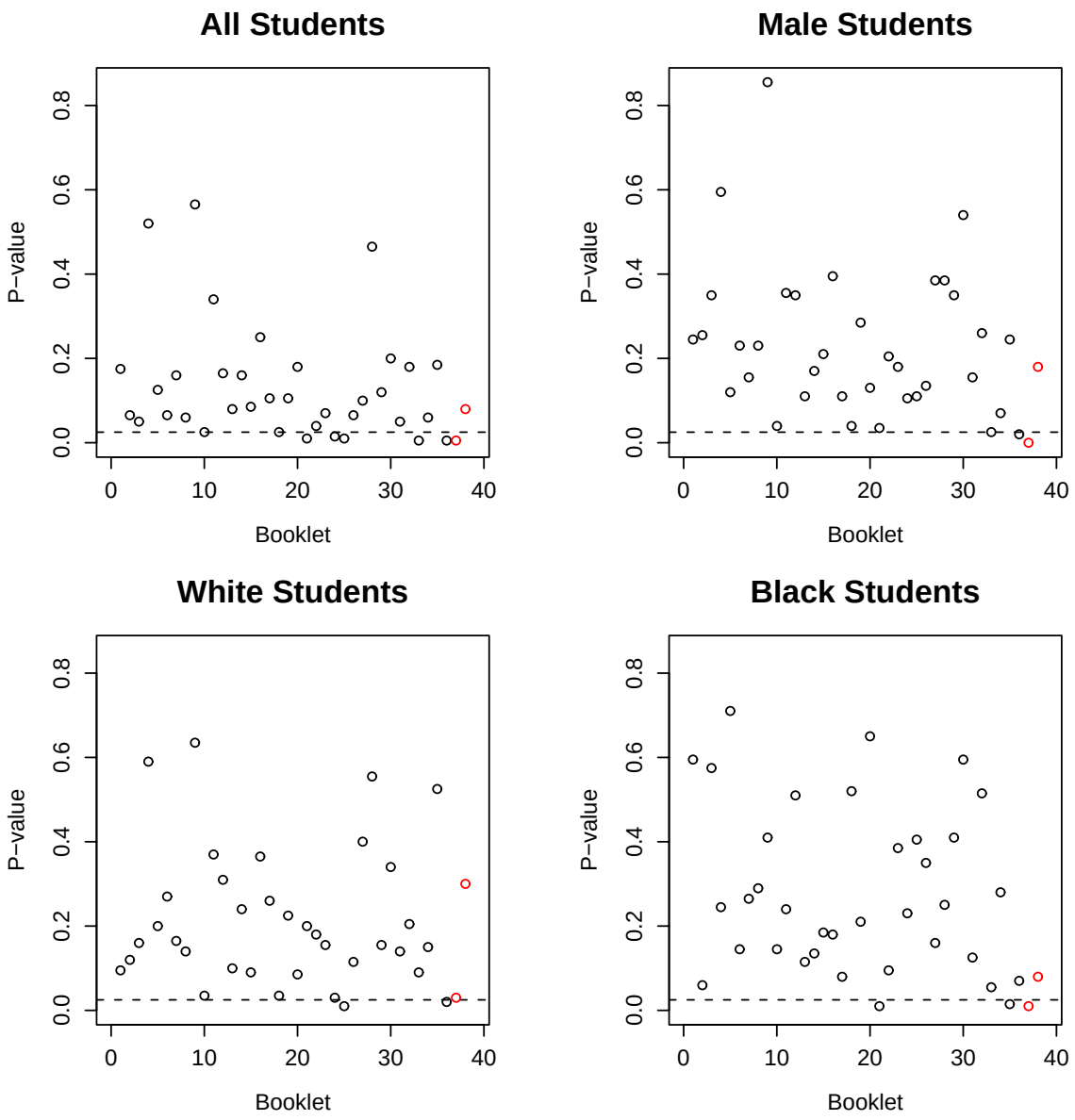

***Figure* 7. The *p*-values for the average score statistics for all students, male students, White students and Black students for all booklets in the Type I error study.**

level. However, the figure shows, as does Figure 4, that more than half of the p-values lie below 0.5.

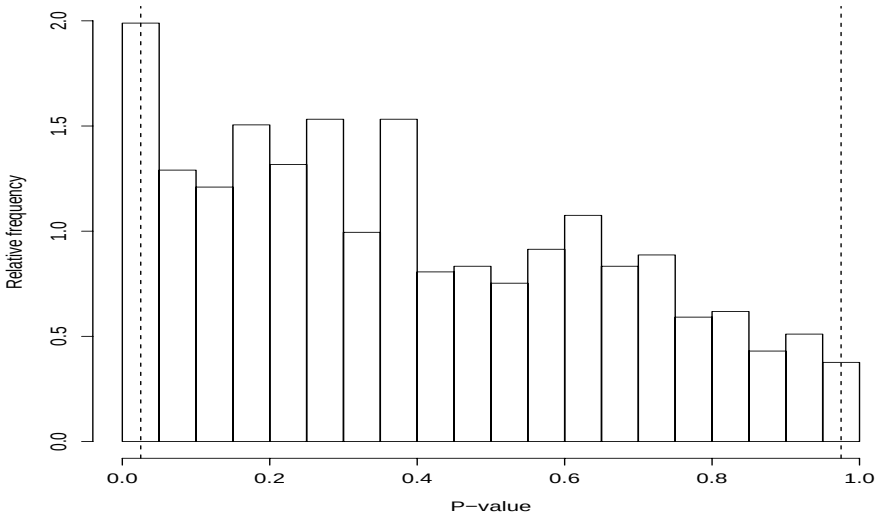

***Figure* 8. The p-values for the average item-score statistic in the Type I error study.**

5.3 Biserial Correlation

Figure 9 shows all the p-values for the statistic. The figure shows that the p-values are more or less uniformly distributed between 0 and 1. The percentage of p-values that are larger than 0.975 or smaller than 0.025 is 6, very close to the nominal level.

5.4 Item-Pair Correlation

Figure 10 shows all the p-values for the statistic. The figure shows that p-values are more or less uniformly distributed between 0 and 1. The percentage of p-values that are larger than 0.975 or smaller than 0.025 is 5, exactly the same as the nominal level.

5.5 Discussion on the Type I Error Rate

The results discussed in the previous section indicate that our suggested method has a satisfactory Type I error rate. The proportion of significant p-values is very close to or the same as the nominal level for all the test statistics we examined. However, for the

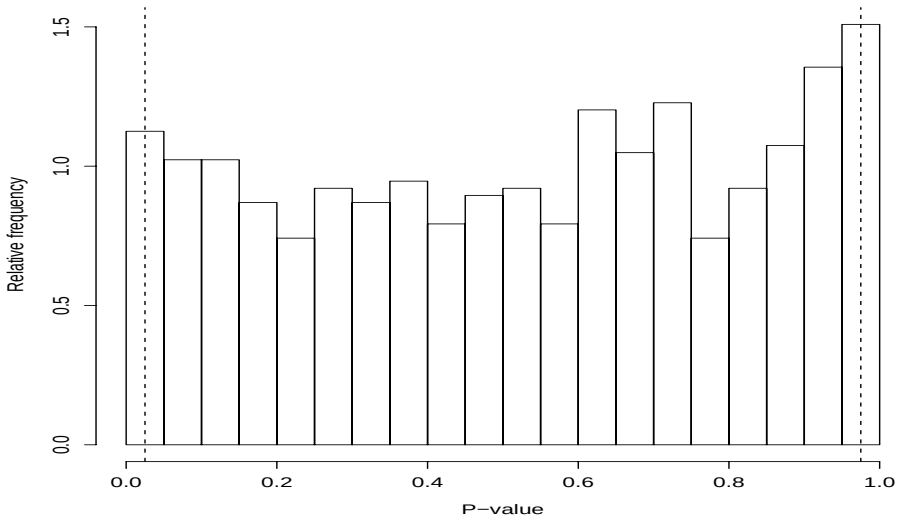

***Figure* 9. The *p*-values for the biserial correlations in the Type I error study.**

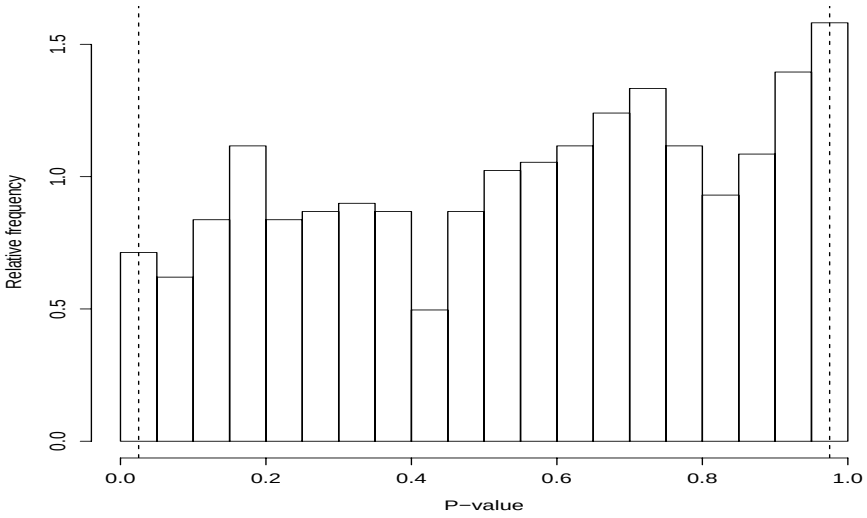

***Figure* 10. The *p*-values for the Item-Pair correlations in the Type I error study.**

average raw-score statistic and the average item score statistic, the p-values do not seem to follow a uniform distribution; further research might be done on this issue.

6 Conclusions

To ensure quality control and overall improvement of the NAEP statistical analysis, there is a constant need to 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. This paper applied a simulation-based model-fit technique to NAEP data to investigate whether basic statistical summaries like the average raw scores for different subgroups are predicted adequately by the LRM employed in NAEP. Our suggested procedure is simple to understand and uses existing NAEP software, so that operational implementation of the procedure will be easy. The analysis of a real data set provided us with limited evidence of misfit of the NAEP model. However, the magnitude of the misfit does not seem to be drastic (i.e., the misfit most likely does not have any practical significance).

The distribution of the p-values for the average raw-score and the average item-score statistics under the null model were found to be non-uniform and not centered around 0.5: the model seems to underpredict these quantities. We do not have an explanation for this phenomenon as of now. It is possible that the phenomenon has to do with the way NAEP generates plausible values or with the discrepancy between the scaling stage (where the ability parameters are assumed to follow independent discrete univariate distributions) and the conditioning stage (where the ability parameters are assumed to follow a multivariate regression model) of NAEP estimation. Further research might be conducted on this issue.

There are a number of related issues that could be examined in the future. Our suggested technique is similar to the PPMC method and the parametric bootstrap method, both of which were found to have satisfactory Type I error and power rate for a variety of IRT models. Hence, we expect the suggested method to have satisfactory Type I error rate and power. However, a more detailed study of Type I error rate and power of the method are a future research issue to pursue. It is possible to examine raw-score-based graphical item fit, like that employed in Sinharay (2006). It is also possible to compute correlations between the scaled composite scores and the raw scores of the examinees—a high correlation and adequate prediction of booklet-level raw scores by the model will provide additional assurance that the subgroup averages of scaled scores are accurate. In

this work, we examined booklet-level statistics only. It may be informative to examine test statistics that combine information from several booklets (e.g., an overall average item score) by combining information over booklets and a weighted average of booklet averages for a subgroup. As NAEP reports the percentage of examinees at or above different performance levels (e.g., basic, proficient), it will be helpful to focus on a statistic related to percentages. Running an MCMC algorithm and then employing the PPMC method to assess the fit of the NAEP model can be a topic of future research, especially after the recent works on an MCMC algorithm for fitting the NAEP model (e.g., Johnson & Jenkins, 2004).

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (Eds.). (2000). *The 1998 NAEP technical report.* (NCES 2001509). Washington, DC: National Center for Education Statistics.

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

Beaton, A. (1987). *The NAEP 1983–84 technical report.* Princeton, NJ: ETS.

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

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.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman & Hall.

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 Rep. No. RR-04-38). Princeton, NJ: ETS.

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

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, QU.

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

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika, 44,* 358–381.

Mislevy, R. (1985). Estimation of latent group effects. *Journal 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*. Princeton, NJ: ETS.

Rogers, A., Tang, C., Lin, M.-J., & Kandathil, M (2006). DGROUP [Computer software]. Princeton, NJ: ETS.

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, 42*, 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.

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. (Eds.) (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 Online, 2*, 29–48.

von Davier, M., & Sinharay, S. (2004). *Application of the stochastic EM method to latent regression models* (ETS Research Rep. No. RR-04-34). Princeton, NJ: ETS.

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. 26. *Psychometrics.* Amsterdam: Elsevier.