Chapter 9

OVERVIEW OF PART II: THE ANALYSIS OF 1998 NAEP DATA[1]

Nancy L. Allen, James E. Carlson, and John R. Donoghue
Educational Testing Service

9.1 INTRODUCTION

The purpose of this chapter is to summarize some information from previous chapters that is integral to the analysis of NAEP data, to summarize the analysis steps used for all subjects, and to indicate what information is in each of the remaining chapters. The overview of the analyses conducted on the 1998 NAEP data focuses on the common elements of the analyses used across the subject areas of the assessment. Some of this information is available only within this chapter. Details by subject area are provided in Chapters 14 through 24.

The organization of this chapter is as follows:

- Section 9.2 provides a short overview of the NAEP design for 1998. To provide additional background information, the section also provides a short description of the samples selected for 1998. Chapters 1 through 7 provide this same information in much more detail.

- Section 9.3 summarizes the steps in analysis common to all subject areas. Some of this information is described in more detail in other chapters. The rest is included only within this chapter. The topics covered are as follows:

  - Section 9.3.1 briefly describes the preparation of the final sampling weights. Detailed information about the weighting procedures is given in Chapters 10 and 11. Detailed information about the sampling design is in Chapters 3 and 4.
  - Section 9.3.2 provides information about the scoring reliability of constructed-response items. It provides information about the reliability measures used with the NAEP data during analysis. Chapter 7 contains information about the reliability procedures used during the scoring process.
  - Section 9.3.3 summarizes the information provided by the teacher questionnaires, and indicates its use during the analysis process.
  - Section 9.3.4 provides a description of the item properties examined for background questions and for cognitive items. It includes a description of the classical item statistics examined for both dichotomously (right versus wrong) and polytomously (more than two response categories) scored items. It also includes a description of the item-level results available from summary data tables. Chapter 13 contains more information about the conventions used in creating these summary tables. Finally, a thorough description of differential item functioning analyses is provided.

[1] Nancy L. Allen, James E. Carlson, and John R. Donoghue were responsible for the psychometric and statistical analysis of the 1998 national and state NAEP data.

- Section 9.3.5 summarizes the steps used to scale NAEP data. The steps include item response theory (IRT) scaling of the items, generating plausible values to account for measurement error, transforming the results to the final reporting scale, creating composite scores if necessary, and providing tables of reported statistics. Details of the theory behind these steps are available in Chapter 12.
- Section 9.3.6 provides some information about previous results of dimensionality analyses.
- Finally, Section 9.3.7 gives an introduction to hypothesis testing and drawing correct conclusions about NAEP data. Specific information about which hypothesis test procedures were used for different purposes is provided in Chapter 13.

- Section 9.4 contains a description of the information provided in Chapters 10 through 24 of this report.

9.2 SUMMARY OF THE NAEP DESIGN

As described in Chapter 1, the 1998 NAEP comprised three components. One component encompassed major assessments in reading, writing, and civics, providing detailed information about student scale scores at the fourth-, eighth-, and twelfth-grade levels of nonpublic and public schools. The second major component was the state assessment at the fourth- and eighth-grade levels in reading and at the eighth-grade level in writing. In addition to the two major components, special studies—a civics special trend study, a 50-minute writing study, and a classroom-based study of writing—were conducted. The results from and procedures used in these special studies are reported in separate documents.

Results from the analyses described in the following chapters were published in the following reports:

- *The NAEP 1998 Reading Report Card for the Nation and the States* (Donahue et al., 1999), which provides both public- and nonpublic-school data for major NAEP reporting subgroups for all of the jurisdictions that participated in the state assessment program, as well as selected results from the 1998 national reading assessment.

- *The NAEP 1998 Writing Report Card for the Nation and the States* (Greenwald et al., 1999), which provides both public- and nonpublic-school data for major NAEP reporting subgroups for all of the jurisdictions that participated in the state assessment program, as well as selected results from the 1998 national writing assessment.

- *The NAEP 1998 Civics Report Card for the Nation* (Lutkus et al., 1999), which provides both public- and nonpublic-school results for major NAEP reporting subgroups from the 1998 national civics assessment.

Because the samples of students included in the 1998 NAEP assessment are listed and described in detail in Chapter 1, only a brief description of these samples is given here. The 1998 national samples consisted of the main NAEP samples for reading, writing, and civics, which were based on a common set of assessment procedures including grade-level samples, and samples for these special studies; a study of

trends in civics performance (1988–1998); a study in which students were administered a 50-minute writing assessment; and a study of classroom writing.

As described in Chapters 1 and 2, for each subject area in the main and state assessments, blocks of items were used to create a large number of different assessment booklets according to a focused design. The 1998 civics assessment used a focused balanced incomplete block (BIB) design. The 1998 reading and writing assessments used focused partially balanced incomplete block (focused PBIB) designs. In a focused BIB design, each block of cognitive items appears in the same number of booklets. To balance possible block-position main effects, each block appears an equal number of times in each position. In addition, the focused BIB design requires that each block of items be paired in a booklet with every other block of items. If one of the features that define a focused BIB design is not evident, then the design is called a focused partially balanced incomplete block (PBIB) design.

9.3 ANALYSIS STEPS

Because the analysis methods are not identical across subject areas, a separate analysis chapter has been included for each major assessment. The procedures used depended on whether assessment items were scored dichotomously (right versus wrong) or polytomously (more than two categories of response) and whether links across grade levels were required. Basic procedures common to most or all of the subject area analyses are summarized here. The order is essentially that in which the procedures were carried out.

9.3.1 Preparation of Final Sampling Weights

Because NAEP uses a complex sampling design (Chapters 3 and 4) in which students in certain subpopulations have different probabilities of inclusion in the sample, the data collected from each student must be assigned a weight to be used in analyses. The 1998 NAEP weights were provided by Westat, the NAEP contractor in charge of sampling. Detailed information about the weighting procedures is available in Chapters 10 and 11 and in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

9.3.2 Reliability of Scoring Constructed-Response Items

A minimum of 25 percent of the responses for reading, writing, and civics items involved only in the national assessment and 6 percent of the responses for reading and writing items involved in both the national and state assessments were scored by a second reader to obtain statistics on interreader (interrater) reliability. Ranges for percentage of exact agreement for the combined state and national assessments of reading, writing, and civics can be found in Table 7-2. This reliability information was also used by the team leaders to monitor the capabilities of all readers and maintain uniformity of scoring across readers. More information about this use of the reliability information is provided in Chapter 7.

In addition to reliability information calculated and used during the scoring process, several additional reliability measures are calculated for constructed-response items after the item response data has been placed in the NAEP database. They appear in Appendix C. These include a final percentage exact agreement, the intraclass correlation, Cohen's Kappa (Cohen, 1968), and the product-moment correlation between the scores for the first and second readers. These measures are summarized in Zwick (1988), Kaplan and Johnson (1992), and Abedi (1996). Each measure has advantages and disadvantages for use in different situations. In this report, the percentage exact agreement is reported for all

constructed-response items, Cohen's Kappa is reported for dichotomously scored constructed-response items, and the intraclass correlation is reported for polytomously scored constructed-response items.

9.3.3 Teacher Questionnaires

Teachers of assessed students were asked to complete a two-part questionnaire. The first part of the questionnaire pertained to the teacher's background and training. The second part pertained to the procedures used by the teacher for specific classes containing assessed students. See Chapter 2 for a description of the teacher questionnaires.

To analyze the data from the teacher questionnaires at grades 4 and 8 with respect to the students' data, each teacher's questionnaire had to be matched to all of the sampled students who were taught by that teacher. In the subsequent chapters, two separate match rates for each grade are given. The first is the percentage of students that could be matched to both the first and second parts of the teacher questionnaire. For these students, information is available about the background and training of their teachers and about the methods used in the particular class they attended. The second match rate is the percentage of students that could be matched to the first part of the teacher questionnaire. This match rate is larger because more students could be matched with information about a teacher than with information about the particular class they attended. Note that these match rates only reflect the student-level missing data. They do not reflect the additional missing data due to item-level nonresponse on the part of teachers. Variables derived from the teacher questionnaires were used as reporting variables at the student level and as variables that contributed to conditioning for the appropriate samples.

Teachers of students who were in the grade 4 assessment sample were asked to complete a two-part questionnaire. As with the grade 8 teacher questionnaire, the first part pertained to the teacher's background and training. Unlike the grade 8 teacher questionnaire, the second part pertained to only a single class that the teacher taught. In development of the questionnaires, it was thought that fourth-grade teachers would teach one class in each subject. In practice, that was found to be untrue for a number of teachers. A single student-teacher match rate matching students to the first part of the questionnaire is reported for grade 4 in the following chapters.

9.3.4 Analysis of Item Properties: Background and Cognitive Items

The first step in the analysis of the 1998 data was item-level analysis of all instruments. Item analyses were performed separately for each grade on each item in each subject area. Each block of items was analyzed separately by grade, with the total score on the block (including the analyzed item) used as the criterion score for statistics requiring such a score. In the cases where final weights were not available, preliminary weights were used in these preliminary analyses. The item analysis of cognitive items was repeated after scaling of the items was completed.

9.3.4.1 Background Items

For each NAEP background item, the unweighted and weighted percent of students who gave each response were examined, as well as the percent of students who omitted the item and the percent who did not reach the item. The number of respondents was also tabulated. These preliminary analyses were conducted within grade cohorts and within major reporting categories. If unexpected results were found, the item data and the encoding of responses were rechecked.

9.3.4.2 Cognitive Items

All NAEP cognitive items were subjected to analyses of item properties. These analyses included conventional item analyses and incorporated examinee sampling weights. Item analysis was conducted at the block level so that the "number correct" scores for students responding to an item, selecting each option of an item, omitting an item, or not reaching an item, is the average number of correct responses for the block containing that item. Because of the inclusion of polytomously scored items in the cognitive instruments, it was necessary to use special procedures for these items. The resulting statistics are analogous to those for the dichotomously scored items, as listed below.

Dichotomously Scored Items. These items were analyzed using standard procedures that result in a report for each item that includes:

- for each option of the item, for examinees omitting and not reaching the item, and for the total sample of examinees:
  - the number of examinees,
  - the percentage of examinees,
  - the mean of number-correct scores for the block in which the item appears, and
  - the standard deviation of number-correct scores for the block in which the item appears;
- the percentage of examinees providing a response that was "off-task";
- p+, the proportion of examinees who received a correct score on the item (ratio of number correct to number correct plus wrong plus omitted);
- Δ, the inverse-normally transformed p+ scaled to mean 13 and standard deviation 4;
- the biserial correlation coefficient between the item and the number-correct scores for the block in which the item appears; and
- the point-biserial correlation coefficient between the item and the number-correct scores for the block in which the item appears.

***Polytomously Scored Items*.** Enhanced procedures were employed for polytomously scored items. Methods parallel to those used for dichotomously scored items resulted in values reported for each distinct response category for the item. Response categories for each item were defined in two ways—one based on the original codes for responses as specified in the scoring rubrics used by the scorers, and one used in defining the item response theory (IRT) model scales. The latter was based on a scoring guide developed by subject-area and measurement experts and it defined the treatment of each response category in scaling. For example, a constructed-response item with four response categories would initially have seven categories (not-reached, omitted, off-task, and the four valid response categories). Another set of statistics resulted from mapping the response categories (excluding not-reached) into a new set of categories reflecting the scoring guide for the items as scaled. A constructed-response item with ordered categories, for example, would be mapped into a set of integers in a corresponding order. The scoring guide could result in the collapsing of (combining of) some response categories. The response categories, based on the final scoring guide developed by subject-area and measurement experts, were used to calculate the polytomously scored item statistics.

The following statistics, analogous to those for dichotomously scored items, were computed:

- The percentage of examinees providing a response that was "off-task."

- In place of p+, the ratio of the mean item score to the maximum-possible item score was used.
- In place of Δ, the inverse-normally transformed ratio of the mean item score to the maximum-possible item score scaled to mean 13 and standard deviation 4.
- The polyserial correlation coefficient was used in place of the biserial.
- The Pearson correlation coefficient, or R-polyserial was used in place of the point-biserial.

9.3.4.3 Tables of Item-Level Results

Tables were created of the percentages of students choosing each of the possible responses to each item within each of the samples administered in 1998. The results for each item were cross-tabulated against the basic reporting variables such as region, gender, race/ethnicity, public/nonpublic school, and parental education. All percentages were computed using the sampling weights. These tables are referred to as the test question section of the electronically available summary data tables for each sample. In the summary data tables, the sampling variability of all population estimates was obtained by the jackknife procedure used by ETS in previous assessments.

9.3.4.4 Tables of Block-Level Results

Tables summarizing the item statistics for all of the items within each block are provided in Chapters 16, 17, 20, 21, and 24. These tables contain statistics calculated using student weights to account for NAEP's complex sampling of students, as well as the unweighted sample size. Weighted summary statistics estimate the results for the whole population of students in the NAEP sampling frame.

- The **unweighted sample size** is the number of students in the reporting sample who receive each block in the assessment. It is the number of students contributing to the other statistics presented in the tables.

- The **weighted average item score** for the block is the average, over items, of the score means for each individual weighted items in the block. Missing responses to polytomous items before the last observed response in a block are also considered intentional omissions and scored so that the response is in the lowest category. Occasionally, extended constructed-response items are the last item in a block of items. Because considerably more effort is required of the student to answer these items, nonresponse to an extended constructed-response item at the end of a block is considered an intentional omission (and scored as the lowest category) unless the student also did not respond to the item immediately preceding that item. In that case, the extended constructed-response item is considered not reached and treated as if it had not been presented to the student. In the case of the main and state writing assessment, there is a single constructed-response item in each separately-timed block. In the writing assessment when a student does not respond to the item or when the student provides an off-task response, the response is also treated as if the item had not been administered. Scaling areas in NAEP are determined a priori by grouping items into content areas for which overall performance is deemed to be of interest, as defined by the frameworks developed by the National Assessment Governing Board (NAGB). A scale score θ_k is defined a priori by the collection of items representing that scale. What is important, therefore, is that the models capture salient information in the response data to effectively summarize the overall performance on the content area of the populations and subpopulations being assessed in the content areas.

- The **weighted average R-polyserial correlation** is the average, over items, of the item-level R-polyserial correlations (R-biserial for dichotomous items) between the item and the number-correct block score. For each item-level R-polyserial, total block number-correct score (including the item in question, and with students receiving zero points for all not-reached items) was used as the criterion variable for the correlation. The number-correct score was the sum of the item scores for a student where correct dichotomous items are assigned 1 and correct polytomous (or multiple-category) items are assigned the score category for the response. Data from students classified as not reaching the item were omitted from the calculation of the statistic.

- The **weighted alpha reliability** is the average of the polyserial correlations for polytomous items and the biserial correlation for the dichotomous items within a block. As for the weighted average R-polyserial correlations, the total block number-count score was used as the criterion.

- The **weighted proportion of students attempting the last item** of a block (or, equivalently, one minus the proportion of students not reaching the last item) is often used as an index of the degree of speededness associated with the administration of that block of items. Mislevy and Wu (1988) discussed these conversions.

9.3.4.5 Differential Item Functioning Analysis of Cognitive Items

Differential item functioning (DIF) analysis refers to procedures that assess whether items are differentially difficult for different groups of examinees. DIF procedures typically control for overall between-group differences on a criterion, usually test scores. Between-group performance on each item is then compared within sets of examinees having the same total test scores.

DIF analyses were conducted for items in the national main assessments in reading, writing, and civics that had not previously been studied for differential item functioning. Each set of analyses involved three reference group/focal group comparisons: male/female, White/Black, and White/Hispanic.

The Mantel-Haenszel Procedure. The DIF analyses of the dichotomous items were based on the Mantel-Haenszel chi-square procedure (Mantel & Haenszel, 1959), as adapted by Holland and Thayer (1988). The procedure tests the statistical hypothesis that the odds of correctly answering an item are the same for two groups of examinees that have been matched on some measure of proficiency (usually referred to as the matching criterion). The DIF analyses of the polytomous items were completed using the Mantel-Haenszel ordinal procedure which is based on the Mantel procedure (Mantel, 1963), (Mantel & Haenszel, 1959). These procedures compare proportions of matched examinees from each group in each polytomous item-response category.

For both types of analyses, the measure of proficiency used is typically the total item score on some collection of items. Since, by the nature of the BIB or PBIB design, booklets comprise different combinations of blocks, there is no single set of items common to all examinees. Therefore, for each student, the measure of proficiency used was the total item score on the entire booklet. These scores were then pooled across booklets for each analysis. This procedure is described by Allen and Donoghue (1994, 1996). In addition, because research results (Zwick & Grima, 1991) strongly suggest that sampling weights should be used in conducting DIF analyses, the weights were used.

For each dichotomous item in the assessment, an estimate of the Mantel-Haenszel common odds ratio, α_{MH} , expressed on the ETS delta scale for item difficulty, was produced. The estimates indicate the difference between reference group and focal-group item difficulties (measured in ETS delta scale units), and typically run between about +3 and -3. Positive values indicate items that are differentially easier for the focal group than the reference group after making an adjustment for the overall level of proficiency in

the two groups. Similarly, negative values indicate items that are differentially harder for the focal group than the reference group. It is common practice at ETS to categorize each item into one of three categories (Petersen, 1988): "A" (items exhibiting no DIF), "B" (items exhibiting a weak indication of DIF), or "C" (items exhibiting a strong indication of DIF). Items in category "A" have Mantel-Haenszel common odds ratios on the delta scale that do not differ significantly from 0 at the alpha = .05 level or are less than 1.0 in absolute value. Category "C" items are those with Mantel-Haenszel values that are significantly greater than 1 and larger than 1.5 in absolute magnitude. Other items are categorized as "B" items. A plus sign (+) indicates that items are differentially easier for the focal group; a minus sign (-) indicates that items are differentially more difficult for the focal group.

The ETS/NAEP DIF procedure for polytomous items uses the Mantel-Haenszel ordinal procedure (Mantel & Haenszel, 1959). The summary tables of identified polytomous items contain generalizations of the dichotomous A, B, and C categories: "AA," "BB," or "CC."

SIBTEST Procedure. For the first time in the 1998 assessment, ETS introduced the SIBTEST (Shealy & Stout, 1993) DIF procedure into the analyses of NAEP items. All items new in 1998 were examined using both Mantel-Haenszel and SIBTEST procedures for DIF. Like the Mantel-Haenszel procedure, SIBTEST seeks to compare the performance of the focal and reference group members of similar ability. The Mantel-Haenszel procedure uses matching on total score to establish comparability; SIBTEST uses a linear "regression correction" (see [Shealy & Stout, 1993] for details) to obtain more accurate matching of the groups. Simulation results (Chang, et al., 1995; Roussos & Stout, 1996) indicate that the Mantel-Haenszel procedure and SIBTEST function similarly for most items, although SIBTEST maintains better Type I error control for items with extreme discrimination IRT(a-parameters).

Like the Mantel-Haenszel procedure, SIBTEST analyses used the entire booklet score in forming the matching variable. These results were then pooled across the booklets using a procedure described by Chang, et al. (1995) and implemented by Donoghue (1998b). Sampling weights were used for SIBTEST analyses.

The SIBTEST measure of DIF, β, is in the metric of Dorans and Kulick's (1986) standardized mean difference (SMD). As an effect size measure, the SMD divided by the item standard deviation was used (as was done for polytomous items with the Mantel procedure). For an item to receive the designation C (dichotomous items) or CC (polytomous items), two criteria had to be met: (a) the estimate of β had to be significantly different from zero, and (b) the absolute value of the effect size (SMD/std. dev.) had to be at least .25.

In 1998, results for the SIBTEST procedure were quite similar to those for the Mantel-Haenszel procedure. All but 1 C or CC item identified by the Mantel-Haenszel procedure was also identified by SIBTEST. No C or CC items were uniquely identified by SIBTEST. All C or CC items identified by either procedure were referred to DIF committees (described below).

Standardization Method. In standard DIF analyses such as Mantel-Haenszel and SIBTEST, it is well established that a moderately long matching test is required for the procedures to be valid (i.e., identify DIF in items unconfounded by other irrelevant factors [e.g., Donoghue, Holland, & Thayer, 1993]). In the main and state NAEP writing assessments, the booklets contain two 25-minute blocks, with one writing prompt per block. Thus, each examinee has (at most) two responses on six-category prompts. This is too little information for the test statistics associated with Mantel (1963) or SIBTEST (Shealy & Stout, 1993) procedures to function effectively. Thus, standard DIF approaches based on statistical tests of items are likely to function poorly, and so were not used in the writing assessment analysis.

In the writing assessment, the standardization method of Dorans and Kulick (1986) was used to produce descriptive statistics. The matching variable was the total score on the booklet. As in other NAEP DIF analyses, the statistics were computed based on pooled booklet matching; the results are accumulated over the booklets in which a given item appears (e.g., Allen & Donoghue, 1996). This analysis was accomplished using the standard NAEP DIF program NDIF that also calculates the Mantel-Haenszel statistic. The statistic of interest appears under the label SMD for "standardized mean difference." First, differences in the item score between the two comparison groups are calculated for each level of the booklet score. Then, the SMD for the item is the average of these differences divided by their standard deviation.

Significance testing was not performed, due to the low reliability of the matching variable. Instead, the standardized mean difference values were used descriptively, to identify those items that demonstrate the most evidence of DIF. A rough criterion used in the past to describe DIF for polytomous items has been to create the ratio of the SMD to the item's standard deviation and flag any item with a ratio of at least .25. A criteria of at least .10 could also be arbitrarily used to identify items with the most evidence of DIF.

All NAEP DIF Procedures. All NAEP DIF analyses used rescaled sampling weights. A separate rescaled weight was defined for each comparison as

$$Rescaled\ Weight = Original\ Weight \bullet \frac{Total\ Sample\ Size}{Sum\ of\ the\ Weights}$$

where the total sample size is the total number of students for the two groups being analyzed (e.g., for the White/Hispanic comparison, the total number of White and Hispanic examinees in the sample at that grade), and the sum of the weights is the sum of the sampling weights of all the students in the sample for the two groups being analyzed. Three rescaled weights were computed for White examinees—one for the gender comparison and two for the race/ethnicity comparisons. Two rescaled overall weights were computed for the Black and Hispanic examinees—one for the gender comparison and another for the appropriate race/ethnicity comparison. The rescaled weights were used to ensure that the sum of the weights for each analysis equaled the number of students in that comparison, thus providing an accurate basis for significance testing.

In the calculation of total item scores for the matching criterion, not-reached, off-task, and omitted items were considered to be wrong responses. Polytomous items were weighted more heavily in the formation of the matching criterion, proportional to the number of score categories. For each item, calculation of the Mantel-Haenszel statistic did not include data from examinees who did not reach the item in question.

Each DIF analysis was a two-step process. In the initial phase, total item scores were formed and the calculation of DIF indices was completed. Before the second phase, the matching criterion was refined by removing all identified C or CC items, if any, from the total item score. The revised score was used in the final calculation of all DIF indices. Note that when analyzing an item classified as C or CC in the initial phase, that item score is added back into the total score for the analysis of that item only.

Following standard practice at ETS for DIF analyses conducted on final forms, all C or CC items were reviewed by a committee of trained test developers and subject-matter specialists. Such committees are charged with making judgments about whether or not the differential difficulty of an item is unfairly related to group membership. The committees assembled to review NAEP items include both ETS staff and outside members with expertise in the field. The committees carefully examine each identified item

to determine if either the language or contents would tend to make the item more difficult for an identified group of examinees. As pointed out by Zieky (1993):

> It is important to realize that DIF is not a synonym for bias. The item response theory based methods, as well as the Mantel-Haenszel and standardization methods of DIF detection, will identify questions that are not measuring the same dimension(s) as the bulk of the items in the matching criterion . . . .Therefore, judgment is required to determine whether or not the difference in difficulty shown by a DIF index is unfairly related to group membership. The judgment of fairness is based on whether or not the difference in difficulty is believed to be related to the construct being measured . . . .The fairness of an item depends directly on the purpose for which a test is being used. For example, a science item that is differentially difficult for women may be judged to be fair in a test designed for certification of science teachers because the item measures a topic that every entry-level science teacher should know. However, that same item, with the same DIF value, may be judged to be unfair in a test of general knowledge designed for all entry-level teachers. (p. 340)

9.3.5 Scaling

Scales based on item response theory (IRT) were derived for each subject area. Three scales were created for national main reading grade 8 and grade 12 assessment data, one for each purpose for reading. Only two of these scales—Reading for Literary Experience and Reading to Gain Information—were assessed at grade 4. A single scale was created for national main writing assessment data, and one scale was created for national main civics assessment data. NAEP uses the methodology of multiple imputations (plausible values) to estimate characteristics of the scale score distributions. Chapter 12 describes in detail the theoretical underpinnings of NAEP's scaling methods and the required estimation procedures. The basic analysis steps are outlined here.

1. Use the NAEP BILOG/PARSCALE computer program (described in Chapter 12) to estimate the parameters of the item response functions on an arbitrary provisional scale. This program uses an IRT model incorporating the two- and three-parameter logistic forms for dichotomously scored items and the generalized partial-credit form for polytomously scored items. In order to select starting values for the iterative parameter-estimation procedure for each dataset, the program is first run to convergence, imposing the condition of a fixed normal prior distribution of the scale score variable. Once these starting values are computed, the main estimation runs model examinee scale score ability as a multinomial distribution. That is, no prior assumption about the shape of the scale score distribution is made. In analyses involving more than one population, estimates of parameters are made with the overall mean and standard deviation of all subjects' proficiencies specified to be 0 and 1, respectively.

2. Use a version of the MGROUP program (described in Chapter 12), which implements the method of Mislevy (see Chapter 10 or Mislevy, 1991) to estimate predictive scale score distributions for each respondent on an arbitrary scale, based on the item parameter estimates and the responses to cognitive items and background questions.

3. Use random draws from these predictive scale score distributions (plausible values, in NAEP terminology) for computing the statistics of interest, such as mean proficiencies for demographic groups.

4. Determine the appropriate metric for reporting the results and transform the results as needed. This includes the linking of current scales to scales from the past or the selection of the mean and variance of new scales. After scale score distributions for the scaling are transformed, composite scale score distributions are created for the reading, writing, and civics assessments.

5. Use the jackknife procedure to estimate the standard errors of the mean proficiencies for the various demographic groups.

As explained in Chapter 10, the plausible values obtained through the IRT approach are not optimal estimates of individual scale score; instead, they serve as intermediate values to be used in estimating subpopulation characteristics. Under the assumptions of the scaling models, these subpopulation estimates are statistically consistent, which would not be true of subpopulation estimates obtained by aggregating optimal estimates of individual scale score.

9.3.5.1 Scaling the Cognitive Items

The data from the national main assessment samples were scaled using IRT models. For dichotomously scored items two- and three-parameter logistic forms of the model were used, while for polytomously scored items the generalized partial-credit model form was used. These two types of items and models were combined in the NAEP scales. Item parameter estimates on a provisional scale were obtained using the NAEP BILOG/PARSCALE program. The fit of the IRT model to the observed data was examined within each scale by comparing the empirical item response functions with the theoretical curves, as described in Chapter 12. Plots of the empirical item response functions and theoretical curves were compared across assessments for items in the reading trend assessment. The DIF analyses previously described also provide information related to the model fit across subpopulations.

The national main assessments of reading, writing, and civics each have special characteristics that determine the procedures that were followed for the scaling of each subject. For reading, a key consideration was the degree of similarity between the 1998 assessment and earlier assessments in terms of the populations assessed and the characteristics of the assessment instrument used. The civics and writing scales were not linked to any previously defined scales.

The frameworks for the different subject areas dictate differences in the numbers of scales. For reading, item parameter estimation was performed separately for each of three scales defined in its framework, using data from each grade sample separately.

9.3.5.2 Generation of Plausible Values for Each Scale

After the scales were developed, plausible values were drawn from the predictive distribution of scale score values for each student (this process is called conditioning). For the writing and civics scales, plausible values were drawn separately for each grade. For the reading scale, vectors of multivariate plausible values were drawn from the joint distribution of scale score values for the assessed student. The scales within an assessment are correlated. Multivariate generation utilizes this shared variation among the scales in generating the plausible values. This procedure properly reflects the dependency between the scale proficiencies. Multivariate plausible values were computed separately for each grade. All plausible values were later rescaled to the final scale metric using appropriate linear transformations.

The variables used to calculate plausible values for a given national main assessment scale or group of scales included a broad spectrum of background, attitude, and experiential variables and composites of such variables. All standard reporting variables were included. To enhance numerical

stability for the national main assessment scales, the original background variables were standardized and transformed into a set of linearly independent variables by extracting principal components from the correlation matrix of the original contrast variables. The principal components, rather than the original variables, were used as independent variables to calculate plausible values for those scales. Details of the conditioning process and of the NAEP BGROUP and NAEP CGROUP (Thomas, 1994) computer programs that implement the process are presented in Chapter 12. The variables used in conditioning are listed in Appendix F.

9.3.5.3 Transformation to the Reporting Metric

Reading short-term trend scales were linked to previous assessment scales via common population linking procedures described in the subject-specific data analysis chapters. Essentially, the 1994 and 1998 data were calibrated together. Data from the two assessments were scaled together in the same BILOG/PARSCALE run, specifying the samples for each assessment as coming from different populations. For each scale, the mean and standard deviation of the 1994 data from this joint calibration were matched to the mean and standard deviation of the 1994 data as previously reported. This then linked the 1998 data to the previously established scale. New scales were established for the writing and civics national main assessment. Then the metrics for the newly established scales were set to have a mean of 150 and a standard deviation of 35.

The transformations were of the form

$$\theta_{target} = A \bullet \theta_{calibrated} + B$$

where

θ_{target} = scale level in terms of the system of units of the final scale used for reporting;

$\theta_{calibrated}$ = scale level in terms of the system of units of the provisional NAEP-BILOG/PARSCALE scale;

$A = SD_{target} / SD_{calibrated}$;

$B = M_{target} - A \bullet M_{calibrated}$;

SD_{target} = the estimated or selected standard deviation of the scale score distribution to be matched;

$SD_{calibrated}$ = the estimated standard deviation of the sample scale score distribution on the provisional NAEP-BILOG/PARSCALE scale;

M_{target} = the estimated or selected mean of the scale score distribution to be matched; and

$M_{calibrated}$ = the estimated mean of the sample scale score distribution on the provisional NAEP-BILOG/PARSCALE scale.

After the plausible values were linearly transformed to the new scale, any plausible value less than 0 was censored to 0. For the reading assessment, any value greater than 500 was censored to 500; for the

writing and civics assessments, any value greater than 300 was censored to 300. Fewer than 1 percent of the students in any sample were censored in this way. The final transformation coefficients for transforming each provisional scale to the final reporting scale are given in subsequent chapters.

9.3.5.4 Definition of Composites for the Multivariate Scales in Reading

In addition to the plausible values for each scale, a composite of the individual reading assessment scales was created as a measure of overall proficiency. The composite scale score was a weighted average of the plausible values of the individual scales. The weights reflected the relative importance of the scales and were provided in the framework developed by the subject-area committee. The weights are approximately proportional to the number of items in each scale at a given grade level.

9.3.5.5 Tables of Scale Score Means and Other Reported Statistics

Scale scores and trends in scale scores were reported by grade for a variety of reporting categories. Additionally, the percentages of the students within each of the reporting groups who were at or above achievement levels were reported to provide information about the distribution of achievement within each subject area. All estimates based on scale score values have reported variances or standard errors based on scale score values, including the error component due to the latency of scale score values of individual students as well as the error component due to sampling variability. These tables are part of the electronically delivered summary data tables.

9.3.6 Dimensionality Analysis

Over the years a number of studies have been conducted in order to seek answers to the question of how many dimensions underlie the various NAEP assessment instruments, and whether there is a sufficiently strong first dimension to support inferences about a composite scale in subjects such as reading. For the 1992 mathematics and reading assessments, a study was conducted (Carlson, 1993) to determine whether the increasing emphasis on extended constructed-response items that are scored polytomously has any effect on the dimensionality. It was determined that for the 1992 NAEP data, item type was not related to any of the dimensions identified.

9.3.6.1 Previous Dimensionality Analyses of NAEP Data

In an early study, the dimensionality of NAEP reading assessment data collected during the 1983–84 academic year was examined by Zwick (1986, 1987). Zwick also studied simulated data designed to mirror the NAEP reading item response data but having known dimensionality. Analysis of the simulated datasets allowed her to determine whether the BIB spiraling design artificially increases dimensionality. Zwick found substantial agreement among various statistical procedures, and that the results using BIB spiraling were similar to results for complete datasets. Overall she concluded that “it is not unreasonable to treat the data as unidimensional” (1987, p. 306).

Rock (1991) studied the dimensionality of the NAEP mathematics and science tests from the 1990 assessment using confirmatory factor analysis. His conclusion was that there was little evidence for discriminant validity except for the geometry scale at the eighth-grade level, and that “we are doing little damage in using a composite score in mathematics and science” (p. 2).

A second-order factor model was used by Muthén (1991) in a further analysis of Rock’s mathematics data, to examine subgroup differences in dimensionality. Evidence of content-specific variation within subgroups was found, but the average (across seven booklets) percentages of such

variation was very small, ranging from essentially 0 to 22, and two-thirds of these percentages were smaller than 10.

Carlson and Jirele (1992) examined 1990 NAEP mathematics data. Analyses of simulated one-dimensional data were also conducted, and the fit to these data was slightly better than that to the real NAEP data. Although there was some evidence suggesting more than one dimension in the NAEP data, the strength of the first dimension led the authors to conclude that the data "are sufficiently unidimensional to support the use of a composite scale for describing the NAEP mathematics data, but that there is evidence that two dimensions would better fit the data than one" (p. 31).

Carlson (1993) studied the dimensionality of the 1992 mathematics and reading assessments. The relative sizes of fit statistics for simulated as compared to actual data suggested that lack of fit may be more due to the BIB spiraling design of NAEP than the number of dimensions fitted. Kaplan (1995) similarly found that the chi-squared goodness of fit statistic in the maximum likelihood factor analysis model was inflated when data were generated using a BIB design. The sizes of the fit statistics for incomplete simulation conditions (a BIB design as in the actual NAEP assessment) were more like those of the real data than were those of the case of simulation of a complete data matrix. Consistent with findings of Zwick (1986, 1987), however, the incomplete design for data collection used in NAEP does not appear to be artificially inflating the number of dimensions identified using these procedures.

9.3.7 Drawing Inferences from the Results

Drawing correct inferences from the results of the assessments depends on several components. First, the hypothesis of no difference between groups must be tested statistically. For the 1998 assessment, the use of *t*-tests was introduced for most comparisons. These tests are more appropriate than *z*-tests based on normal distribution approximations when the statistics that are being compared are from distributions with thicker tails than those from the normal distribution. The statistical significance tests used in NAEP are described in detail in Chapter 13.

A second component contributing to drawing correct inferences is the way in which error rates are controlled when multiple comparisons are made. If we wish to make a number of comparisons in the same analysis, say White students versus Black, Hispanic, Asian/Pacific Island, and American Indian students, the probability of finding "significance" by chance for at least one comparison increases with the family size or number of comparisons. By the Bonferroni inequality, for a family size of 4, for example, the probability of a false positive (Type I error) using $\alpha = 0.05$ is less than or equal to $4 \times 0.05 = 0.20$, larger than most decision makers would accept.

One general method for controlling error rates in multiple comparisons is based on the Bonferroni inequality. In this method, the Bonferroni inequality is applied and α is divided by the family size, *n*. Now $\alpha = .05/4 = .0125$, and using α, the combined probability of one or more errors in the four comparisons remains controlled at less than or equal to .05. Note that dividing the probability by *n* is not the same as multiplying the critical value or the confidence band by *n*. Indeed, in moving from a family size of 1 to 4, we increase the critical value only from 1.960 to 2.498, a 27.4 percent increase. Doubling the family size again, to 8, increases the critical value to 2.735, an additional 9.5 percent increase. To double the initial critical value to 3.92, the family size would have to be increased to 564.

The power of the tests thus depends on the number of comparisons planned. There may be cases for which, before the data are seen, it is determined that only certain comparisons will be conducted. As an example, with the five groups above, interest might lie only in comparing the first group with each of the others (family size 4), rather than comparing all possible pairs of groups (family size 10). This means

that some possibly significant differences will not be found or discussed, but the planned comparisons will have greater power to identify real differences when they occur.

In 1998, a different criterion was used to increase the power of statistical tests in NAEP. Unlike other multiple-comparison procedures (e.g., the Bonferroni procedure) that control the familywise error rate (i.e., the probability of making even one false rejection in the set of comparisons), the false discovery rate (FDR) controls the expected proportion of falsely rejected hypotheses. So, if an α of .05 is selected, about 95 percent of the hypothesis tests made rejected or accepted the hypothesis correctly, while about 5 percent of the hypothesis tests made rejected or accepted the hypothesis incorrectly. Familywise procedures are considered conservative for large families of comparisons. Therefore, the FDR procedure is more suitable for multiple comparisons in NAEP than other procedures (Williams, Jones, & Tukey, 1999). The FDR procedure used in NAEP has been described by Benjamini and Hochberg (1994). These methods for controlling error rates in multiple comparisons are described in Chapter 13.

A third component contributing to drawing correct inferences is limiting comparisons to those for which there are adequate data. In NAEP reports and data summaries, estimates of quantities such as composite and content area scale score means, percentages of students at or above the achievement levels, and percentages of students indicating particular levels of background variables (as measured in the student, teacher, and school questionnaires) are reported for the total population as well as for key subgroups determined by the background variables. In some cases, sample sizes were not large enough to permit accurate estimation of scale score or background variable results for one or more of the categories of these variables.

For results to be reported for any subgroup in NAEP, a minimum sample size of 62 is required. This number was arrived at by determining the sample size required to detect an effect size of 0.5 with a probability of .8 or greater. The effect size of 0.5 pertains to the “true” difference in mean scale score between the subgroup in question and the total population, divided by the standard deviation of scale score in the total population. In addition, subgroup members must represent at least five primary sampling units (PSUs).

A fourth component contributing to drawing correct inferences is limiting comparisons to those comparing statistics with standard errors that are estimated well. Standard errors of mean proficiencies, proportions, and percentiles play an important role in interpreting subgroup results and comparing the performances of two or more subgroups. The jackknife standard errors reported by NAEP are statistics whose quality depends on certain features of the sample from which the estimate is obtained. In certain cases, typically when the number of students upon which the standard error is based is small or when this group of students all come from a small number of participating schools, the mean squared error associated with the estimated standard errors may be quite large. In the summary reports, estimated standard errors subject to large mean squared errors are followed by the symbol "!".

The magnitude of the mean squared error associated with an estimated standard error for the mean or proportion of a group depends on the coefficient of variation *(CV)* of the estimated size of the population group, denoted as *N*. The coefficient of variation is estimated by:

$$CV(\hat{N}) = \frac{SE(\hat{N})}{\hat{N}}$$

where $\hat{N}$ is a point estimate of *N* and $SE\left(\hat{N}\right)$ is the jackknife standard error of $\hat{N}$.

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples for this

group may be quite large. Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are followed by "!" in the tables of all summary reports. These standard errors, and any confidence intervals or significance tests involving them, should be interpreted with caution. (Further discussion of this issue can be found in Johnson & Rust, 1993.)

A final component contributing to drawing correct inferences pertains to comparisons involving extreme proportions. When proportions are close to zero or one, their distributions differ greatly from t- or *z*-distributions. For this reason, hypothesis tests of the sort used by NAEP are not appropriate in these cases. Under these conditions, no test is made. Chapter 13 includes the specific definition of extreme proportion used in the analysis of 1998 data.

9.4 OVERVIEW OF CHAPTERS 10 THROUGH 24

The remaining chapters of this report are as follows:

Chapters 10 and 11: The 1998 national assessment used a stratified multistage probability sampling design that provided for sampling certain subpopulations at higher rates (see Chapters 3 and 4). Because probabilities of selection are not the same for all assessed students, sampling weights must be used in the analysis of NAEP data. Also, in NAEP's complex sample, observations are not independent. As a result, conventional formulas for estimating the sampling variance of statistics are inappropriate. Chapters 10 and 11 describe the weighting procedures and methods for estimating sampling variance that are necessitated by NAEP's sample design. Further detail on sampling and weighting procedures is provided in the *NAEP 1994 Sampling and Weighting Report* (Wallace & Rust, 1996), published by Westat, the NAEP contractor in charge of sampling.

Chapter 12: A major NAEP innovation introduced by ETS is the reporting of subject-area results in terms of IRT-based scales. Scaling methods can be used to summarize results even when students answer different subsets of items. For purposes of summarizing item responses, NAEP developed a scaling technique that has its roots in IRT and in the theories of imputation of missing data. Chapter 12 describes this scaling technique, the underlying theory, and the application of these methods to 1998 NAEP data. The final section of Chapter 12 gives an overview of the NAEP scales that were developed for the 1998 assessment.

Chapter 13: The 1998 assessment analyses included changes in the methods, procedures, and conventions used in making group comparisons. Chapter 13 highlights these changes and provides details about which results were reported.

Chapter 14: The 1998 reading assessment was based on a framework developed by the National Assessment Governing Board for the 1992 reading assessment. This framework was used in the 1994 and 1998 assessments. Chapter 14 discusses the framework and assessment instruments used in the 1998 assessment.

Chapters 15, 16, and 17 describe analyses of the reading data for national and state assessments. This analysis included a study of the cognitive variables and student background variables. At grades 4 and 8, background information and data on instructional methods were collected from teachers, and the relation of these variables to reading scale scores was examined. The reading results appear in the *NAEP 1998 Reading Report Card for the Nation and the States* (Donahue et al., 1999).

Chapter 18: The 1998 writing assessment was based on a new framework developed by the National Assessment Governing Board for the 1998 assessment. Chapter 18 discusses the framework and assessment instruments used in the 1998 assessment.

Chapters 19, 20, and 21 describe analyses of the writing data for national and state assessments. This analysis included a study of the cognitive variables and student background variables. At grade 8, background information and data on instructional methods were collected from teachers and the relation of these variables to writing data was examined. The writing results appear in the *NAEP 1998 Writing Report Card for the Nation and the States* (Greenwald et al., 1999).

Chapter 22: The 1998 civics assessment was based on a new framework developed by the National Assessment Governing Board for the 1998 assessment. Chapter 22 discusses the framework and assessment instruments used in the 1998 assessment.

Chapters 23 and 24 describe analyses of the civics assessment. This analysis included a study of the cognitive variables and student background variables. At grades 4 and 8, background information and data on instructional methods were collected from teachers and the relation of these variables to civics scale scores was examined. The civics results appear in the *NAEP 1998 Civics Report Card for the Nation* (Lutkus et al., 1999).

Chapter 10

WEIGHTING PROCEDURES AND ESTIMATION OF SAMPLING VARIANCE FOR THE NATIONAL ASSESSMENT[1]

Jiahe Qian, Bruce A. Kaplan, and Eugene G. Johnson
Educational Testing Service

Tom Krenzke and Keith F. Rust
Westat

10.1 INTRODUCTION

As in previous assessments, the 1998 national assessment used a complex sample design with the goal of securing a sample from which estimates of population and subpopulation characteristics could be obtained with reasonably high precision (as measured by low sampling variability). At the same time, it was necessary that the sample be economically and practically feasible to obtain. The resulting sample had certain properties that had to be taken into account to ensure valid analyses of the data from the assessment.

The 1998 NAEP sample was obtained through a stratified multistage probability sampling design that included provisions for sampling certain subpopulations at higher rates (see Chapter 3). To account for the differential probabilities of selection, and to allow for adjustments for nonresponse, each student was assigned a sampling weight. Section 10.2 discusses the procedures used to derive these sampling weights.

Section 10.3 discusses other weighting procedures in the NAEP samples. These procedures include generating modular weights, which would allow analysts to compare results between sample types. National linking (NL)[2] weights were generated so that national and state-by-state assessments could be equated for national and state results to be reported on a common scale. School weights were created so that school-level data could be analyzed. Also, reporting weights for samples with accommodations were processed for possible use in 2002 when reporting trend from 1998. Section 10.4 discusses the potential bias due to nonresponse.

Another consequence of the NAEP sample design is its effect on the estimation of sampling variability. Because of the effects of cluster selection (cluster of elements: students within schools, schools within primary sampling units) and because of the effects of certain adjustments to the sampling weights (nonresponse adjustment and poststratification), observations made on different students cannot be assumed to be independent of one another. In particular, as a result of clustering, ordinary formulas for the estimation of the variance of sample statistics based on assumptions of independence will tend to underestimate the true sampling variability. Section 10.5 discusses the jackknife technique used by NAEP to estimate sampling variability.

---

[1] Keith F. Rust and Tom Krenzke were responsible for the design and implementation of the weighting process for the 1998 NAEP national assessment. Jiahe Qian, with the assistance of Bruce Kaplan and in consultation with Eugene G. Johnson, was responsible for the planning, specification, and coordination of the national weighting at ETS.

[2] Note that in previous NAEP state assessments, the weights for national linking samples were called the state aggregate comparison, or SAC, weights. Many people thought this was easy to confuse with state weights, so the term ‘national linking’ will be used in this report.

10.2 WEIGHTING PROCEDURES FOR ASSESSED AND EXCLUDED STUDENTS IN THE NATIONAL SAMPLES

Since the sample design determines the derivation of the sampling weights and the estimation of sampling variability, it will be helpful to note the key features of the 1998 national sample design. A description of the design appears in the first four sections of this report.

The 1998 sample was a multistage probability sample consisting of four stages. The first stage of selection, the primary sampling units (PSUs), consisted of counties or groups of counties. The second stage of selection consisted of elementary and secondary schools. The assignment of sessions and sample types to sampled schools (see Chapter 3) comprised the third stage of sampling, and the fourth stage involved the selection of students within schools and their assignment to sessions.

The probabilities of selection of the first-stage sampling units were proportional to measures of their size, while the probabilities for subsequent stages of selection were such that the overall probabilities of selection of students were approximately uniform, with exceptions for certain subpopulations that were oversampled by design. Schools with relatively high concentrations of Black students, Hispanic students, or both, were deliberately sampled at a higher than normal rate to obtain larger samples of respondents from those subpopulations, in order to increase the precision in the estimation of the characteristics of these subpopulations. Nonpublic-school students were sampled at three times the normal rate, again to increase the precision of estimates for this population subgroup. For all assessment components, students from schools with smaller numbers of eligible students received lower probabilities of selection, as a means of enhancing the cost efficiency of the sample.

The 1998 national assessment includes three student cohorts: students in grades 4, 8, and 12. The national assessment of all grades was conducted in the spring of 1998 to provide a cross-sectional view of students' abilities in reading, writing, and civics.

The full 1998 national assessment thus includes a number of different samples from several populations. Each of these samples has its own set of weights that are to be used to produce estimates of the characteristics of the population addressed by the sample (the target population). Each sample has an additional set of weights to accommodate the reporting requirements. The various samples and their target populations are as follows. The target population for each of these samples (one for each grade) consisted of all students who were in the specified grade and were deemed assessable by their school. There were three distinct session types at each grade: writing/civics, reading, and civics special trend. Each session type was conducted as one or more distinct sessions within a school. Administration of each session type was always conducted separately from other session types. Within the writing/civics sessions, students in grade 4 received either a 25-minute writing booklet or a civics booklet, while in grades 8 and 12 students received a 25-minute writing booklet, a 50-minute writing booklet, or a civics booklet.

To facilitate analyses, two kinds of weights were produced. "Reporting weights" were produced separately by grade and assessment type for analyses of the reporting samples that were defined for each assessment. Several of the reporting samples included students from multiple sample types. "Modular weights," as discussed in Section 10.3.1, were produced separately by grade and sample type for the reading assessment. They are applied for analyses involving any one sample type, or for comparing one sample type with another. Thus, across grades, session types, and sample types, there were 14 sets of reporting weights, and there were 6 sets of modular weights for students in reading assessments.

10.2.1 Base Weights

As indicated earlier, to enhance the precision of estimates of characteristics of these oversampled subgroups, NAEP deliberately oversampled certain subpopulations to obtain larger samples of respondents from those subgroups by using differential sampling rates. Because of the oversampling public schools with high concentrations of Black and/or Hispanic students and the oversampling of nonpublic schools, these subpopulations are overrepresented. As a result of oversampling students, subpopulations to Black and/or Hispanic students from public schools with low concentrations of Black and/or Hispanics, and corresponding to SD/LEP students in schools assigned reading sessions, are also overrepresented in the sample. Lower sampling rates were introduced also for very small schools (those schools with only 1 to 19 eligible students). This reduced level of sampling from small schools was undertaken in a near optimal manner as a means of reducing variances per unit of cost (since it is relatively costly to administer assessments in these small schools). Appropriate estimation of population characteristics must take disproportionate representation into account. This is accomplished by assigning a weight to each respondent, where the weights approximately account for the sample design and reflect the appropriate proportional representation of the various types of individuals in the population.

Two sets of weights were computed for the 1998 samples. “Modular weights” were computed for analyses involving students of reading assessments in one sample type, or for comparing results between sample types. Each reading assessment type, by grade and sample type, weights up separately to the target population. “Reporting weights” were computed for analyses of the reporting samples defined in Table 10-1. The reading reporting samples include students from more than one sample type. For reporting samples that include only one sample type (i.e., writing/civics and civics special trend), the reporting weights are identical to the modular weights. The steps for computing these two sets of weights are identical, up to and including the step of “trimming” the weights. The trimmed weights were poststratified separately by sample type to create the modular weights. In a parallel procedure, the trimmed weights were scaled back using a “reporting factor” so that the sample types included in each reporting sample, when combined, would weight up to the target population. The resulting weights were poststratified (but not separately by sample type) to create the reporting weights.

Table 10-1
Reporting Samples for 1998 National Assessments

Subject	Grade Assessed	Reporting Samples*
Civics	4, 8, 12	A3+B3
Civics Special Trend	4, 8, 12	A3+B3
Reading	4, 8, 12	A2+A3+B2
25-Minute Writing	4, 8, 12	A3+B3

* **A** indicates assessed non SD/LEP students; **B** indicates assessed SD/LEP students; and 2 or 3 indicates the sample type.

The weighting procedures for 1998 included computing the student’s base weight, the reciprocal of the probability that the student was selected for a particular subject type. Such weights are those appropriate for deriving estimates from probability samples via the standard Horvitz-Thompson estimator (see Cochran, 1977). These base weights were adjusted for nonresponse and then subjected to a trimming algorithm to reduce a few excessively large weights. The weights were further adjusted by a student-level poststratification procedure to reduce the sampling error. The poststratification was performed by adjusting the weights of the sampled students so that the resulting estimates of the total number of students in a set of specified subgroups of the population corresponded to population totals, which were based on information from the Current Population Survey and U.S. Census Bureau estimates of the

population. The subpopulations were defined in terms of race, ethnicity, geographic region, grade, and age relative to grade. The distribution of the various weighting factors is presented in Westat's report entitled *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

The base weight assigned to a student is the reciprocal of the probability that the student was selected for a particular assessment. That probability is the product of six factors:

1. The probability that the PSU was selected
2. The probability that a Catholic, religious-affiliated, or other nonpublic school was selected for the PSS file
3. The conditional probability, given the PSU, that the school was selected
4. The conditional probability, given the sample of schools in a PSU, that the school was allocated to the specified session type
5. The conditional probability, given the sample of schools in a PSU, that the sample type was assigned to the school
6. The conditional probability, given the school, that the student was selected for the specified subject type

Thus, the base weight for a student may be expressed as the product

$$W_B = PSUWGT\_M \bullet QSCHWT \bullet SCH\_WT \bullet STYWT \bullet SA\_WT \bullet STUSA\_WT$$

where *PSUWGT_M, QSCHWT, SCH_WT, STYWT, SA_WT,* and *STUSA_WT* are, respectively, the reciprocals of the preceding probabilities.

Variations across the various 1998 assessments in probabilities of selection, and consequently of weights, were introduced by design, either to increase the effectiveness of the sample in achieving its goals of reporting for various subpopulations, or to achieve increased efficiency per unit of cost.

The PSU weight, *PSUWGT_M*, is the reciprocal of the probability of selection for the PSU. Of the 94 PSUs selected, 22 were certainty PSUs and have a PSU weight of 1.0. For the remaining 72 PSUs, the probability of selection was calculated to account for the initial selection of one PSU per stratum.

The PSS weight, *QSCHWT*, is the reciprocal of the probability of selection of the Catholic, religious-affiliated, and other nonpublic schools from the PSS area frame. *QSCHWT*= 1 for schools on the PSS list frame. See Section 3.2.4.1 for more information about the PSS list and area frames.

The school weight, *SCH_WT*, is the reciprocal of the probability of selection of the school conditional on the PSU.

The session allocation weight, *SA_WT*, is the reciprocal of the probability that the particular session was allocated to the school. This is a function of the session type and the number of sessions allocated to the school. Session allocation weights were calculated separately for each session type. The values for the session allocation weights are summarized in Table 10-2. The session allocation weights were adjusted for smaller-than-expected schools to account for one or more session types that were

dropped. The adjustment factor was computed as the number of sessions assigned divided by the number of retained sessions assigned for the session type.

Table 10-2
Session Allocation Weights Used in the 1998 National Assessment

	Writing/Civics		Reading		Civics Special Trend	
Grade	Session Allocation Weight	Number of Sessions Assigned	Session Allocation Weight	Number of Sessions Assigned	Session Allocation Weight	Number of Sessions Assigned
4	18/13	1	18/4	1	18	1
	1	2	18/8	2	18/2	2
	1	3	18/12	3	18/3	3
	1	4	18/16	4	18/4	4
8	47/34	1	47/11	1	47/2	1
	1	2	47/22	2	47/4	2
	1	3	47/33	3	47/6	3
	1	4	47/44	4	47/8	4
	1	5	1	5	47/10	5
12	49/34	1	49/13	1	49/2	1
	1	2	49/26	2	49/4	2
	1	3	49/39	3	49/6	3
	1	4	49/45	4	49/8	4
	1	5	49/47	5	49/10	5

The sample type weight, STYWT, is the reciprocal of the probability that the sample type was assigned to the school. For reading, the weight is 2, and for other sessions the weight was set to 1.

Cooperating substitute schools received the values of the following weighting components from the original sampled school that it replaced: *PSUWGT_M, QSCHWT, SCH_WT, SA_WT, STYWT*.

For assessed students, the student weight, STUSA_WT, is the reciprocal of the probability that the student was selected for the particular session to which he or she was assigned. This probability is the product of the within-school sampling rate; the proportion of the relevant eligible students assigned to the particular session type within the school, as prescribed by the sampling allocation factor; the proportion of students in the session given a subject-specific assessment booklet (see Table 10-3 for the subject factors); and a factor that adjusts for students in year-round schools that are not in school at the time of assessment. Special attention was given to the writing sample allocation factors for accommodated SD/LEP students and nonaccommodated students. The SD/LEP students in 50-minute writing that were accommodated were given 25-minute writing booklets. Therefore, the accommodated students have a higher chance of being assigned the 25-minute writing booklet than the nonaccommodated students. A special poststratification procedure was done for the 50-minute writing sample, as described in Section 10.2.5.1.

Excluded students were weighted with assessed students for each assessment. This was done because the exclusion criteria did not depend on session type. For excluded students, STUSA_WT is computed the same way as assessed and absent students.

Table 10-3
1998 National Assessment Writing and Civics Sample Allocation

Subject	Grade 4	Grade 8	Grade 12
25-Minute Writing Nonaccommodated	13/10	17/10	17/10
25-Minute Writing Accommodated	13/10	17/13	17/13
50-Minute Writing	N/A	17/3	17/3
Civics	13/3	17/4	17/4

10.2.2 Adjustment of the Base Weights for Nonresponse

The base weight for a student was adjusted by two nonresponse factors: SF_WT, to adjust for noncooperating schools and schools that did not conduct all of their assigned sessions (i.e., a session nonresponse); and STUNRADJ, to adjust for students who were invited to the assessment but did not appear either in the scheduled or a makeup session. Thus the nonresponse adjusted weight for a student was of the form:

STUAWT = PSUWGT_M • QSCHWT SCH_WT • SA_WT • STYWT • STUSA_WT • SF_WT • STUNRADJ

The nonresponse adjustment factors were computed as described below.

10.2.2.1 Session Nonresponse Adjustment (SESNRF)

Sessions were assigned to schools before cooperation status was final. The session nonresponse adjustment was intended to compensate for session type nonresponse due to refusing schools or individual session types not conducted. The first three digits of PSU stratum, called subuniverse (formed by crossing the PSU major stratum and the first socioeconomic characteristic used to define the final PSU stratum; see Chapter 3 for more detail) were used in calculating nonresponse adjustments. The adjustment factors were computed separately within classes formed by subuniverse within sample type for reading, and by subuniverse for the other assessment types. Occasionally, additional collapsing of classes was necessary to improve the stability of the adjustment factors, especially for the smaller assessment components. Most classes needing collapsing contained small numbers of cooperating schools. Occasionally, classes with low-response rates were collapsed.

In subuniverse *s* in session type *h*, the session nonresponse adjustment factor SF_WT_{hs} was given by

$$SF\_WT_{hs} = \frac{\sum_{B_{hs}} PSUWGT\_M_i \bullet QSCHWT_i \bullet SCH\_WT_i \bullet SA\_WT_{hi} \bullet STYWT_{hi} \bullet G_i}{\sum_{C_{hs}} PSUWGT\_M_i \bullet QSCHWT_i \bullet SCH\_WT_i \bullet SA\_WT_{hi} \bullet STYWT_{hi} \bullet G_i}$$

where

$PSUWGT\_M_i$ = the PSU weight for the PSU containing school *i*,

$QSCHWT_i$ = the PSS school weight for school *i*,

SCH_WT_i = the school weight for school *i*,

SA_WT_{hi} = the session allocation weight for session type *h* in school *i*,

$STYWT_i$ = the sample type weight for school *i*,

G_i = the estimated number of grade-eligible students in school *i* (the values of G_i were based on QED or PSS data or updated grade enrollment values from field operations),

set B_{hs} = consists of all in-scope originally sampled schools allocated to session type *h* in subuniverse *s* (excluding substitutes), and

set C_{hs} = consists of all schools allocated to session type *h* in subuniverse *s* that ultimately participated (including substitutes).

It should be noted that the nonresponse adjustments assume that nonresponse occurs at random within the categories within which adjustments are made (see Little & Rubin, 1987). Some degree of bias could result to the extent that this assumption is false. It should also be noted that the adjustment accounts for the difference between the substitute's estimated grade enrollment and its corresponding original school's estimated grade enrollment. For the state assessments, a separate weighting factor is used to account for the difference in estimated grade enrollments (see Section 11.2.4).

10.2.2.2 Student Nonresponse Adjustment (STUNRADJ)

Student nonresponse adjustment factors were computed separately for each subject type. The adjustment classes were based on sample type (for reading only), subuniverse, modal age status, and race class (White or Asian/Pacific Islander, other). In some cases, two or more nonresponse classes were collapsed into one to improve the stability of the adjustment factors. For each class c in subject type k, the student nonresponse adjustment factor $STUNRADJ_{kc}$ is computed by

$$STUNRADJ_{kc} = \frac{\sum_{A_{kc}} PSUWGT\_M_j \bullet QSCHWT_j \bullet SCH\_WT_j \bullet SA\_WT_{hj} \bullet STYWT_{hj} \bullet SF\_WT_{hj} \bullet STUSA\_WT_{kj}}{\sum_{B_{kc}} PSUWGT\_M_j \bullet QSCHWT_j \bullet SCH\_WT_j \bullet SA\_WT_{hj} \bullet STYWT_{hj} \bullet SF\_WT_{hj} \bullet STUSA\_WT_{kj}}$$

where,

$PSUWGT\_M_j$ = the PSU weight for the PSU containing student *j*,

$QSCHWT_j$ = the PSS school weight for school containing student *j*,

SCH_WT_j = the school weight for the school containing student *j*,

SA_WT_{hj} = the session allocation weight for the school containing student *j* in session type *h*,

$STYWT_{hj}$ = the sample type weight for the school containing student *j* in session type *h*,

SF_WT_{hj}	=	the session nonresponse adjustment factor for the school containing student *j* in session type *h*,
$STUSA\_WT_{hj}$	=	the within-school student weight for student *j* in subject type *k*,
Set A_{kc}	=	consists of the students in class *c* who were sampled for subject type *k* and not excluded, and
Set B_{kc}	=	consists of the students in class *c* who were assessed in subject type *k*.

Excluded students received nonresponse adjustments of 1.0.

10.2.3 Variation in Weights

As mentioned earlier, the basic sampling design was to select students with uniform selection probability except for planned oversampling in certain types of schools to improve estimates for certain subgroups. However, additional variation in weights was caused by a number of factors. Variation arose from undersampling schools with fewer than six expected students eligible for the grade category. Variation also arose from limiting the number of students selected from large schools. Inaccurate school measures of size also contributed to variability. When the measures of size were off by more than 20 percent, within-school sampling intervals were changed in order to meet the target sample size in the school. In these cases the self-weighting sample design was abandoned in order to meet the target sample size. In addition, the process of session assignment added variability to the weights. The number of sessions was assigned to the school first, and then specific session types were assigned. Thus, the number of sessions of any one type assigned to a school was a random variable. More oversampling within schools, as discussed in Chapter 3, than in 1996 may have caused an increased variation in weights. Finally, adjustment for nonresponse at the school and student levels added to the variation in weights.

Such variability in weights contributed to the variance of overall estimates from the survey by approximately a factor of $F = 1 + V_W^2$, where V_W^2 denotes the coefficient of variation of the student weights. The calculated factors are displayed in Table 10-4.

By design, the use of poststratification factors, to be discussed in Section 10.2.5, also added to weight variation. However, poststratification presumably reduced the variance of overall estimates by reducing the variability in the relative contribution to the overall estimates of subclasses that respond differently.

Table 10-4
Value of Factor F for Sample Subjects Used in the 1998 National Assessment

Grade	Subject	*F*
4	Reading	1.41
	25-Minute Writing	1.41
	Civics	1.41
	Civics Special Trend	1.25
8	Reading	1.42
	25-Minute Writing	1.37
	50-Minute Writing	1.36
	Civics	1.38
	Civics Special Trend	1.31
12	Reading	1.45
	25-Minute Writing	1.34
	50-Minute Writing	1.34
	Civics	1.36
	Civics Special Trend	1.32

10.2.3.1 Trimming the Weights for Outliers

In a number of cases, students were assigned relatively large weights[3]. One cause of large weights was underestimation of the number of eligible students in some schools, leading to inappropriately low probabilities of selection for those schools. A second major cause is the presence of large schools (high schools in particular) in PSUs with small selection probabilities. In such cases, the maximum permissible within-school sampling rate (determined by the maximum sample size allowed per school—see Chapter 3) could well be smaller than the desired overall within-PSU sampling rate for students. Large weights arose also because very small schools were, by design, sampled with low probabilities. Other large weights arose as the result of high levels of nonresponse coupled with low to moderate probabilities of selection, and the compounding of nonresponse adjustments at various levels.

Students with notably large weights have an unusually large impact on estimates such as weighted means. As discussed in the previous section, the variability in weights contributes to the variance of an overall estimate by an approximate factor $(1+V_W^2)$, where V_W is the coefficient of variation of the weights. An occasional unusually large weight is likely to produce large sampling variances of the statistics of interest, especially when the large weights are associated with students with atypical performance characteristics.

To reduce the effect of large contributions to variance from a small set of sample schools, the weights of such schools were reduced, that is, trimmed. The trimming procedure introduces a bias but is expected to reduce the mean square error of sample estimates.

[3] Trimming of small weights was not an issue in national and state NAEP assessments. The distribution of weights for NAEP assessment samples is usually positively skewed. The size of the student groups with relatively small weights is usually relatively large. Thus small weights are usually not outliers and would not contribute to a large coefficient of variation of weights.

The trimming algorithm was identical to that used since 1996 and had the effect, approximately, of trimming the weight of any school that contributed more than a specified proportion, θ, to the estimated variance of the estimated number of students eligible for assessment. The details of the algorithm of trimming weights are given in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

The trimming procedure was done separately within sample type for reading, and overall for 25-minute writing, 50-minute writing, civics, and civics special trend. The number of schools where weights were trimmed was no more than 13 in any one assessment. The most extreme trimming factors applied were of the order of 0.41; trimming affects the weights of only a very small proportion of the assessed and excluded students.

Table 10-5 shows the distributions of eligible students based on the trimmed weights of assessed students for the 25-minute writing samples for each grade. The distributions are similar to those before trimming shown later in the section. To the extent that the characteristics in the table are related to student performance on the 25-minute writing assessment, there is a small bias introduced in the assessment by trimming.

Table 10-5

Distribution of Populations of Eligible Students Based on Trimmed Weights of Assessed Students in Participating Schools, 1998 National 25-Minute Writing Samples

Population	Grade 4	Grade 8	Grade 12
Total Population	3,430,090	3,440,089	2,533,413
Age Category			
At modal age or younger	63.8	59.4	64.1
Older than modal age	36.2	40.6	35.9
Race/Ethnicity Category			
White	58.9	62.1	67.6
Black	13.8	13.1	11.3
Hispanic	20.1	18.5	13.7
Other	7.2	6.4	7.4
Gender*			
Male	50.6	50.0	47.9
Female	49.4	50.0	52.0
SD			
Yes	7.5	7.0	4.3
No	92.5	93.0	95.7
LEP			
Yes	3.5	2.7	2.2
No	96.5	97.3	97.8
SD, LEP			
SD yes, LEP yes	0.2	0.3	0.1
SD yes, LEP no	7.3	6.8	4.2
SD no, LEP yes	3.3	2.5	2.1
SD no, LEP no	89.2	90.5	93.6

* For a very small percentage of students at grades 4, 8, and 12, gender is unknown.

10.2.4 Reporting Factors

Each set of trimmed weights for a given sample type in the reading assessment sums to the target population. Reporting factors were assigned to students in order to scale back the trimmed weights so that final student (reporting) weights within each reporting sample (which may combine students from different sample types) sum to the target population. The reporting factors assigned to students are specific to the reporting samples defined in Table 10-1. Each assessed and excluded student in the reporting sample for reading assessment received a reporting factor as shown in Table 10-6. Students that were assessed or excluded in 25-minute writing, 50-minute writing, civics, and civics special trend, were assigned a reporting factor equal to 1.0, since all students are part of the reporting sample.

Table 10-6
1998 National Reading Assessment
Reporting Factors for Assessed and Excluded Students

Sample Type	Non SD/LEP Students	SD/LEP Students
2	0.5	1
3	0.5	—

10.2.5 Poststratification

As in most sample surveys, the respondent weights are random variables that are subject to sampling variability. Even if there were no nonresponse, the respondent weights would at best provide unbiased estimates of the various subgroup proportions. However, since unbiasedness refers to average performance over a conceptually infinite number of replications of the sampling, it is unlikely that any given estimate, based on the achieved sample, will exactly equal the population value. Furthermore, the respondent weights have been adjusted for nonresponse and a few extreme weights have been reduced in size.

To reduce the mean squared error of estimates using the sampling weights, these weights were further adjusted so that estimated population totals for a number of specified subgroups of the population, based on the sum of weights of students of the specified type, were the same as presumably better estimates based on composites of estimates from the 1995 and 1996 Current Population Survey and 1997 population projections made by the U.S. Census Bureau. For details of the method used to derive these independent estimates, see Appendix C in the *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

This adjustment, called poststratification, is intended especially to reduce the mean squared error of estimates relating to student populations that span several subgroups of the population, and thus also to reduce the variance of measures of changes over time for such student populations.

The poststratification in 1998 was done for all subjects and grades. Within each grade and assessment type group, poststratification adjustment cells were defined in terms of race, ethnicity, and Census region as shown in Tables 10-7. Note that NAEP region was used in years prior to 1996 instead of Census region. This change was made because the data from the Current Population Survey and Census Projections are more reliable for Census regions than for NAEP regions.

These subgroups were used as adjustment cells at grade 12. For grades 4 and 8, each of the seven subgroups was further divided into two eligibility classes: of modal age and not of modal age.

Table 10-7
Major Subgroups for Poststratification in the 1998 National Assessment

Race	Ethnicity	Census Region
Black	Not Hispanic	All
Any	Hispanic	All
Other	Not Hispanic	All
White	Not Hispanic	Northeast
White	Not Hispanic	Midwest
White	Not Hispanic	South
White	Not Hispanic	West

The procedure used at grade 12 was adopted because the independent estimates of the numbers of students in the population did not provide consistent data on the numbers of twelfth-grade students by age. Specifically, the counts of twelfth-grade students age 18 and older are not reliable because they include adult education students. This procedure has been used since 1988. (See Rust, Bethel, Burke, & Hansen, 1990, and Rust, Burke, & Fahimi, 1992, for further details.)

Thus, there were 7 or 14 cells for poststratification. The poststratified weight for each student within a particular cell was the student's base weight, with adjustments for nonresponse and trimming, and the reporting factor from Section 10.2.4, times a poststratification factor. For each cell, the poststratification factor is a ratio whose denominator is the sum of the weights (after adjustments for nonresponse and trimming) of assessed and excluded students, and whose numerator is an adjusted estimate, based on more reliable data, of the total number of students in the cell. The poststratification factor for student j in subject type k and poststratification adjustment class c is given by

$$RPTPS\_AD_{kc} = \frac{TOTAL_c}{\sum_{C_{hc}} W_{Bj} \bullet SF\_WT_j \bullet STUNRADJ_j \bullet TRIMFCTR_j \bullet RPT\_FCTR_j}$$

where

W_{Bj} = the base weight for student j (see Section 10.2.1);

$TOTAL_c$ = the total number of grade-eligible students in class c, from the October 1995 and 1996 Current Population Surveys and 1997 population projections;

SF_WT_j = the session nonresponse adjustment factor for the school containing student j in subject type k;

$STUNRADJ_j$ = the student nonresponse adjustment for student j;

$TRIMFCTR_j$ = the trimming factor for student j;

RPT_FCTR_j = the reporting factor for student j;

Set C_{kc} = consists of the students in class c who were assessed in subject type k, except those at grade 12 who were age 18 or older.

The major subgroups for poststratification in 1998 assessments are shown in Tables 10-7. The poststratification factors can be found in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

10.2.5.1 The 50-Minute Writing Session

The accommodated SD/LEP students sampled in the 50-minute writing session were given a 25-minute writing booklet. Therefore, the set of assessed 50-minute writing students did not contain accommodated students. To allow for comparisons between nonaccommodated students assessed in 25-minute writing to students (all nonaccommodated) in the 50-minute writing session, a special poststratification procedure was used for the weighting of students assessed in the 50-minute writing session. The poststratification adjustment factors for the 50-minute writing session were computed using the set of accommodated students in 25-minute writing, along with the set of students assessed in the 50-minute writing session. After poststratification, the estimated nonaccommodated universe sizes for grade 8 25-minute and 50-minute writing sessions were 3,572,375 and 3,570,306, respectively. For grade 12, the estimated nonaccommodated universe sizes for grade 12 25-minute and 50-minute writing sessions were 3,139,073 and 3,172,348, respectively.

10.2.6 Final Student Reporting Weights

NAEP estimates of student characteristics are based on final student weights, that is, the weight resulting after adjusting the student base weight for nonresponse, trimming, reporting sample factor, and poststratification. The student final weight, FSTUWT, is given by

$$FSTUWT = STUAWT \bullet TRIMFCTR \bullet RPT\_FCTR \bullet PSFCTR$$

where

STUAWT = nonresponse adjusted student base weight, (as defined in Section 10.2.2),

TRIMFCTR = trimming factor (as discussed in Section 10.2.3.1),

RPT_FCTR = reporting sample factor (as defined in Section 10.2.4), and

PSFCTR = poststratification factor (as discussed Section in 10.2.5).

The student full-sample reporting weight, FSTUWT, was used to derive all estimates of population and subpopulation characteristics that have been presented in the various NAEP reports, including simple estimates such as the proportion of students of a specified type who would respond in a certain way to an item and more complex estimates such as mean scale score levels. The distributions of the final student reporting weights are given in Table 10-8. The sample types contained in each reporting sample of the assessment can be found in Table 10-1.

As indicated earlier, under some simplifying assumptions the factor $1 + V_w^2$ indicates the approximate relative increase in variance of estimates resulting from the variability in the weights. The factor V_w^2 for each sample is readily derivable from Table 10-8 by squaring the ratio of the standard deviation to the mean weight. These factors, resulting from the combined effect of the variations in weights introduced by design and from other causes, are discussed in Section 10.2.3.

Table 10-8
Distributions of Final Student Weights for 1998 National Reporting Samples

Grade	Subject	n	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
4	25-Minute Writing	21,266	186	119	26	102	150	220	1,195
	Reading	8,217	480	308	70	269	373	631	2,707
	Civics Special Trend	2,264	1,742	867	401	1,098	1,519	2,242	6,585
	Civics	6,355	621	399	90	340	489	759	4,140
8	25-Minute Writing	21,463	171	104	17	102	137	207	1,075
	Reading	11,674	315	203	29	175	259	388	2,493
	Civics Special Trend	2,148	1,710	945	159	1,033	1,388	2,199	5,705
	Civics	8,553	430	265	47	254	345	526	2,370
	50-Minute Writing	6,275	569	344	61	338	457	698	3,856
12	25-Minute Writing	20,163	158	93	25	94	130	194	1,266
	Reading	13,123	241	161	35	129	194	297	1,373
	Civics Special Trend	2,296	1,399	790	273	870	1,153	1,693	4,809
	Civics	8,010	401	242	64	236	328	501	3,060
	50-Minute Writing	6,006	528	309	86	312	432	648	4,972

10.3 OTHER WEIGHTING PROCEDURES IN THE NATIONAL SAMPLES

10.3.1 Modular Weights

As discussed in Section 10.2, modular weights were computed for the reading assessment to facilitate analyses involving students from a single sample type. The same procedures were used to derive modular and reporting weights up through the weight trimming step described in Section 10.2.3.1. After trimming, weighting continued in two parallel processes. Final student reporting weights were the result of one of these processes, and modular weights were the result of the other.

Modular weights differ from reporting weights for reading in two ways. First, they did not contain the reporting factor described in Section 10.2.4. The second difference lies in the manner in which the weights were poststratified. Since the number of students in the reading reporting samples are nearly twice the number of students in each sample type (type 2 or type 3), the mean of the modular weights is about twice the mean of reporting weights for reading.

The modular weights were poststratified as described in Section 10.2.5, except that each sample type within each grade for reading was poststratified separately. The same initial adjustment cells were used: 7 cells based on race/region for each sample type at grade 12, and 14 cells based on race/region and eligibility class (of modal age, not of modal age) for each sample type at grades 4 and 8. Some adjustment factors were quite variable for the same adjustment cell across different sample types for the same grade and session. This indicates that the individual samples by sample type may not be particularly stable.

The modular weight is the student's base weight after the application of the various adjustments described in Section 10.2, with the exception of applying a reporting factor, and the new poststratification factor described above. The distributions of the modular weights are given in Table 10-9. Note that except for the reading subject, modular weights are identical to reporting weights for a particular grade/subject/sample type combination when that sample type is the only one included in the reporting sample for that grade.

Table 10-9
Distribution of Modular Weights Used in the 1998 National Assessment

Grade	Subject	n	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
4	Reading/2*	4,593	859	510	127	462	721	1,113	3,460
	Reading/3	4,597	858	567	155	481	679	1,034	5,224
8	Reading/2*	6,848	537	344	61	338	457	698	3,856
	Reading/3	6,078	604	409	43	336	514	751	5,977
12	Reading/2*	7,048	444	317	45	224	348	594	2,303
	Reading/3	7,050	453	313	53	236	373	543	2,615

* 2 refers to sample type 2 and 3 refers to sample type 3.

10.3.2 Linking Weights

Linking (NL) weights were generated so that national NAEP and state-by-state assessments could be equated for national and state results to be reported on a common scale. Therefore, the results of each participating jurisdiction would be meaningfully compared with those from the nation samples. Technical details of the 1996 state assessments can be found in *the Technical Report for the NAEP 1996 State Assessment Program in Mathematics* (Allen, Jenkins, Kulick, and Zelenak, 1997) and in the *Technical Report for the NAEP 1996 State Assessment Program in Science* (Allen, Swinton, Isham, and Zelenak, 1998).

The fourth-grade reading and eighth-grade reading and writing assessments conducted in February 1998 in the NAEP 1998 state assessment consisted of identical assessment material to that administered in the corresponding national sample sessions. The guiding principles in the process of linking state and national results were similar to those used for the 1996 assessments. (Technical details of the NAEP 1996 state assessments are given in Allen, Jenkins, Kulick, and Zelenak (1997) and Allen, Swinton, Isham, and Zelenak (1998).) The national and state-by-state assessments were equated so that state and national results could be reported on a common scale. The equating was achieved by using from each assessment that part of the sample representing a common population. For the national samples, this consisted of those fourth-grade or eighth-grade public-school students from a participating state (including the District of Columbia) who were assessed in the national reading or (for grade 8) writing assessment reporting samples.

Although each sample of students received appropriate weights from the weighting procedure used for the national assessment, in an effort to increase the precision of the equating process, an additional weighting adjustment was developed and applied to each subsample by grade and subject, solely for use in equating. For each subsample, the distributions of the national sample reporting weights for three categorical variables were adjusted to agree closely with those obtained from the weighted aggregate sample from the state assessments in the participating states. The first two variables were NAEP region (Northeast, Southeast, Central, and West) and race/ethnicity (White non-Hispanic, Black non-Hispanic, Hispanic, and other). For fourth- and eighth-grade reading, the third variable was reading skill (very good, good, other). For eighth-grade writing, the third variable was the student's writing skill ("I am good at writing."). This variable was based on a writing background item that asks how much a student agrees with the statement "I am good at writing." The categorical variables and control totals for each of the assessed grades and subjects are presented in Tables 10-10 and 10-11.

Table 10-10

*First and Second Categorical Variables Used for Raking**

Raking Dimensions		Fourth Grade Reading Control Total	Eighth Grade Reading Control Total	Eighth Grade Writing Control Total
First Dimension	*NAEP Region*			
	Northeast	427,412	383,213	400,534
	Southeast	731,635	717,450	730,862
	Central	478,480	347,368	318,990
	West	975,015	960,961	971,641
	Total	2,612,532	2,408,992	2,422,027
Second Dimension	*Race/Ethnicity*			
	White non-Hispanic	1,573,388	1,452,593	1,430,992
	Black non-Hispanic	418,533	372,219	375,766
	Hispanic	445,567	427,097	454,611
	Other	175,043	157,082	160,658
	Total	2,612,532	2,408,992	2,422,027

*Due to rounding, the sum of values within categorical variables may not equal the corresponding totals.

Table 10-11

Third Categorical Variable Used for Raking

Grade	Skill		Control Totals*
4	Reading Skill	1. Very Good	1,105,087
		2. Good	965,306
		3. Other	542,139
		Total	2,612,532
8	Reading Skill	1. Very Good	596,581
		2. Good	845,194
		3. Other	967,216
		Total	2,408,992
8	Writing Skill	1. Agree	1,206,813
	("I am good at writing.")	2. Undecided	708,624
		3. Other	506,590
		Total	2,422,027

*Due to rounding, the sum of skill values may not equal the corresponding totals.

The equating of each weight distribution was achieved using a procedure known as iterative proportional fitting, or raking (described by Little & Rubin, 1987). In raking, the marginal population totals, $N_{i.}$ and $N_{.j}$ are known (i.e., age and gender population counts); however, the interior cells of the

cross-tabulation N_{ij} (the age by gender cells) are estimated from the sample by $\hat{N}_{ij}$, where these are the sum of weights in the cells.

The raking algorithm proceeds by proportionally scaling the $\hat{N}_{ij}$, such that the following relations are satisfied:

$$\sum_j \hat{N}_{ij} = N_{i.}$$

and

$$\sum_i \hat{N}_{ij} = N_{.j.}$$

At the completion of the fitting, adjustment factors were derived. The national sample weights for each subgroup were multiplied by these adjustment factors to force their distribution to agree with those from the aggregated state samples for each of these three variables in turn. This process was then repeated, and the final set of adjusted weights was compared with the state sample weights on all three distributions, and found to be in very close agreement. Table 10-12 shows the distribution of the adjustment factors for each of the grades and subjects assessed.

Table 10-12
Percentiles of Raking Adjustments

Distribution	Grade 4 Reading	Grade 8 Reading	Grade 8 Writing
Minimum	0.805	0.885	0.832
10th Percentile	0.816	0.901	0.851
25th Percentile	0.837	0.912	0.899
Median	0.955	1.008	0.987
75th Percentile	1.121	1.026	1.076
90th Percentile	1.150	1.196	1.237
Maximum	1.640	1.523	1.570

10.3.3 School Weights

The sampling procedures used to obtain national probability samples of assessed students also gave rise indirectly to several national probability samples of schools (from which the students were subsequently sampled). So that the school samples can be utilized for making national estimates about schools, appropriate nonresponse adjusted survey weights have been developed.

The school weights were computed separately by session within grade. The school weights were a direct by-product of the student weighting process. The weight for school *i* in session *h* is given by

$$SW_{hi} = PSUWGT\_M_i \bullet QSCHWT_i \bullet SCH\_WT_i \bullet SA\_WT_{hi} \bullet STYWT_{hi} \bullet SF\_WT_{hi}$$

where

$PSUWGT\_M_i$, $QSCHWT_i$, SCH_WT_i , SA_WT_{hi} , $STYWT_{hi}$, and SF_WT_{hi} are defined in Section 10.2.

The school weights for the reading samples are modular weights. Each sample defined by sample type weights up separately to the population. Different school weights are required for analyses involving schools from both sample types. The weights in such cases can be developed by dividing the modular weights by two.

Twelve samples of schools were weighted to be nationally representative. For each grade, the samples include writing/civics, civics special trend, reading sample type 2, and reading sample type 3.

10.3.4 Reporting Weights with Accommodations

Reporting weights were generated using accommodated students in the 1998 reading samples as part of the reporting sample. The weights may be useful in the year 2002 when reporting trend from 1998. These weights will also be used in looking into issues dealing with accommodation. The procedure began with the trimmed weights (Section 10.2.3.1), and proceeded to the application of the reporting factors as shown in Table 10-13. The reporting factors relating to the reporting sample with accommodated students were set to 1.0, while the reporting factors for non-SD/LEP students in the 1998 national reporting sample were 0.5. Thus nonzero weights were produced for the SD/LEP students in sample type 3, while not including the SD/LEP students in sample type 2.

Table 10-13
Reporting Factors for the Reporting Weights with Accommodations for the 1998 National Reading Assessment

Sample Type	Non SD/LEP Students	SD/LEP Students
2	.5	—
3	.5	1

Poststratification was done on the accommodated reporting weights. The resulting final accommodated reporting weights are summarized in Table 10-14.

Table 10-14
Distribution of Accommodated Reporting Weights for the 1998 National Reading Assessment

Grade	n	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
4	8,205	480.80	306.97	74.22	275.84	366.67	624.37	4,662.20
8	11,561	317.77	223.43	29.09	177.33	260.62	389.67	4,887.60
12	13,087	241.76	162.09	35.34	130.09	191.88	295.97	1,424.57

10.3.5 Jackknife Replicate Weights

In addition to the weights that were used to derive all estimates of population and subpopulation characteristics, other sets of weights, called jackknife replicate weights, were derived to facilitate the estimation of sampling variability by the jackknife variance estimation technique. These weights and the jackknife estimator are discussed in Section 10.5.

10.4 POTENTIAL FOR BIAS DUE TO NONRESPONSE

Although school and student nonresponse adjustments are intended to reduce the potential for nonparticipation to bias the assessment results, they cannot completely eliminate this potential bias with certainty. The extent of bias remains unknown, of course, since there are no assessment data for the nonparticipating schools and students. Recently, some studies related with this issue had been done, such as on the effects of excluded students in reporting results (see Donoghue, 2000).

Some insight can be gained about the potential for residual nonresponse bias, however, by examining the weighted school- and student-level distributions of characteristics known for both participants and nonparticipants, especially for those characteristics known or thought likely to be related to achievement on the assessment. If the distributions for the full sample of schools (or students) without the use of nonresponse adjustments are close to those for the participants with nonresponse adjustments applied, there is reason to be confident that the bias from nonparticipation is small.

There are several school-level characteristics available for both participating and nonparticipating schools. The tables below show the combined impact of nonresponse and of the nonresponse adjustments on the distributions of schools (weighted by the estimated number of eligible students enrolled) and students, by the type of school (public, Catholic, other nonpublic), the size of the school as measured by the estimated number of eligible students enrolled, and the urban/rural nature of the place where the school is located. Three size classes have been defined for each grade. The data in the tables that follow are for the 25-minute writing assessment because it is the largest assessment at each grade. It is assumed that other large assessments would behave similarly. More of these types of data are available for other grades and subjects in Appendix A.

Several student-level characteristics are available for both absent and assessed students. The tables that follow show the impact of school nonresponse and nonresponse adjustments, and student nonresponse and nonresponse adjustments on the distributions of eligible students for each grade. This discussion also focuses on the writing/civics session for school-level summaries, and 25-minute writing assessment for student-level tables. The distributions are presented by age category (at or below modal age, and above modal age), race category (White, Black, Hispanic, and other), gender, SD, and LEP.

Table 10-15 shows the weighted marginal distributions of students for each of the three classification variables for each grade, using weighted eligible schools. The distributions before school nonresponse adjustments are based on the full sample of in-scope schools for the writing/civics session—those participating, plus those refusals for which no substitute participated. The distributions after school nonresponse adjustments are based only on participating schools for writing/civics, with school nonresponse adjustments applied to them.

It can be seen from Table 10-15 that even though the level of school nonparticipation is as high as 18 percent after substitution for grade 12 (see Table 3-7) and somewhat lower for the other grades, for the most part, the distributions for the three characteristics considered remain similar. Exceptions may be rural schools in grades 4 and 12, and large grade 12 schools.

Table 10-15

Distribution of Populations of Eligible Students Based on Full Weighted Sample of Eligible Schools, Before and After School Nonresponse Adjustments, 1998 National 25-Minute Writing Samples

	Grade 4		Grade 8		Grade 12	
Population	**Before**	**After**	**Before**	**After**	**Before**	**After**
Total Population	3,775,102	3,775,102	3,714,224	3,714,224	2,856,379	2,856,379
School Type						
Catholic	6.0%	6.8%	4.9%	5.8%	5.3%	6.4%
Other Nonpublic	4.5%	3.7%	4.4%	4.3%	3.8%	2.7%
Public*	89.5%	89.5%	90.6%	89.9%	90.9%	90.9%
School Size†						
1	17.8%	18.1%	9.7%	11.1%	5.3%	6.1%
2	43.7%	42.5%	53.2%	52.4%	67.9%	69.3%
3	38.5%	39.5%	37.1%	36.5%	26.8%	24.6%
School Location						
Large City	18.5%	17.4%	16.5%	17.2%	14.2%	14.3%
Midsize City	19.8%	19.4%	18.5%	17.4%	18.6%	17.3%
Urban Fringe/Large City	26.9%	26.6%	27.1%	27.2%	29.1%	28.7%
Urban Fringe/Midsize City	7.8%	8.0%	10.3%	10.5%	9.5%	10.4%
Large Town	1.1%	0.9%	1.7%	1.2%	1.1%	1.0%
Small Town	11.4%	11.2%	12.9%	11.7%	15.4%	13.8%
Rural	14.5%	16.5%	13.0%	14.7%	12.1%	14.6%

* The term "public schools" extends to state-run, Department of Defense Education Activity (DoDEA), and Bureau of Indian Affairs (BIA) schools.

† Distributions by school size are only comparable to 1996 assessments, since students were eligible by grade only, instead of by grade or age before 1996. School size = number of eligible students enrolled:

	1	2	3
Grade 4	1–49	50–99	100 +
Grade 8	1–49	50–299	300 +
Grade 12	1–49	50–399	400 +

Table 10-16 shows the distributions of the same three classification variables, plus additional distributions of student-level characteristics, using weighted eligible students. The distributions before student nonresponse adjustments are based on assessed and absent science students (with base weights adjusted for school nonparticipation). The distributions after student nonresponse adjustments are based on assessed science students only, with the student nonresponse adjustments also applied to them.

Table 10-16

Distribution of Populations of Eligible Students Before and After Student Nonresponse Adjustments, 1998 National 25-Minute Writing Samples

Population	Grade 4		Grade 8		Grade 12	
	Before	After	Before	After	Before	After
Total Population	3,447,973	3,447,973	3,477,714	3,477,714	2,598,835	2,598,835
School Type						
Catholic	7.1%	7.1%	6.0%	6.3%	6.9%	7.8%
Other Nonpublic	3.8%	3.9%	4.2%	4.3%	2.7%	3.2%
Public*	89.1%	89.0%	89.9%	89.4%	90.4%	88.9%
School Location						
Large City	16.6%	16.5%	17.2%	17.0%	14.4%	14.0%
Midsize City	19.6%	19.6%	17.0%	16.9%	17.6%	17.3%
Urban Fringe/Large City	27.2%	27.3%	28.1%	28.2%	28.9%	28.9%
Urban Fringe/Midsize	7.7%	7.6%	10.6%	10.7%	10.3%	10.4%
City	0.8%	0.8%	1.1%	1.2%	0.8%	0.8%
Large Town	11.5%	11.5%	11.4%	11.5%	13.7%	14.0%
Small Town	16.7%	16.7%	14.5%	14.5%	14.3%	14.6%
Rural						
Age Category						
At Modal Age or Younger	63.8%	63.7%	59.2%	59.4%	63.6%	64.0%
Older than Modal Age	36.2%	36.3%	40.8%	40.6%	36.4%	36.0%
Race/Ethnicity Category						
White	59.2%	59.0%	62.4%	62.3%	68.6%	68.1%
Black	14.1%	13.8%	13.2%	13.0%	11.5%	11.1%
Hispanic	19.7%	20.0%	18.1%	18.3%	13.2%	13.4%
Other	7.0%	7.2%	6.3%	6.4%	6.7%	7.4%
Gender†						
Male	50.5%	50.6%	50.2%	50.0%	48.4%	47.9%
Female	49.4%	49.3%	49.8%	50.0%	51.6%	52.0%
SD						
Yes	7.5%	7.5%	7.3%	7.0%	4.7%	4.3%
No	92.5%	92.5%	92.7%	93.0%	95.3%	95.7%
LEP						
Yes	3.5%	3.5%	2.7%	2.7%	2.1%	2.2%
No	96.5%	96.5%	97.3%	97.3%	97.9%	97.8%
SD, LEP						
SD yes, LEP yes	0.2%	0.2%	0.3%	0.3%	0.1%	0.1%
SD yes, LEP no	7.4%	7.4%	7.0%	6.8%	4.6%	4.2%
SD no, LEP yes	3.3%	3.3%	2.4%	2.5%	2.0%	2.1%
SD no, LEP no	89.2%	89.2%	90.3%	90.5%	93.3%	93.6%

* The term "public schools" extends to state-run, Department of Defense Education Activity (DoDEA), and Bureau of Indian Affairs (BIA) schools.

† Gender is unknown for a small percentage of students.

The rates of student nonparticipation for 25-minute writing were 5.1 percent for grade 4, 7.8 percent for grade 8, and 20.3 percent for grade 12 (see Table 3-16). Table 10-17 shows that for the distributions of type of school attended and place where the school is located, the combined effect of student nonparticipation and the subsequent nonresponse adjustments have resulted in very little change in distribution.

When comparing the distributions in Table 10-16 before and after student nonresponse adjustments, distributions by age category and race/ethnicity are expected to be similar because these variables were used to determine student nonresponse adjustment classes. However, the distributions by

gender, SD, and LEP are also similar. To the extent that nonrespondents would perform like respondents with the same characteristics (defined by the classification variables in the tables), the bias in the assessment data is small.

Table 10-17 shows the weighted distributions of eligible students in participating schools, using the base weights of assessed and absent students unadjusted for school-level nonresponse. Tables 10-16 and 10-17 show that both school and student-level nonresponse and nonresponse adjustments have little effect on the distributions of eligible students by age, race/ethnicity, gender, SD and LEP. All of the distributions in the tables are similar.

Table 10-17

Distribution of Populations of Eligible Students Before School and Student Nonresponse Adjustments, 1998 National 25-Minute Writing Samples

Population	Grade 4	Grade 8	Grade 12
Total Population	3,065,866	2,946,000	2,598,835
Age Category			
At Modal Age or Younger	64.2%	59.3%	63.6%
Older than Modal Age	35.8%	40.7%	36.4%
Race/Ethnicity Category			
White	58.4%	61.9%	68.6%
Black	14.5%	13.6%	11.5%
Hispanic	20.0%	18.3%	13.2%
Other	7.0%	6.2%	6.7%
Gender*			
Male	50.5%	50.2%	48.4%
Female	49.4%	49.8%	51.6%
SD			
Yes	7.6%	7.2%	4.7%
No	92.4%	92.8%	95.3%
LEP			
Yes	3.6%	2.8%	2.1%
No	96.4%	97.2%	97.9%
SD, LEP			
SD yes, LEP yes	0.2%	0.3%	0.1%
SD yes, LEP no	7.4%	7.0%	4.6%
SD no, LEP yes	3.4%	2.5%	2.0%
SD no, LEP no	89.0%	90.2%	93.3%

* Gender is unknown for a small percentage of students.

Further information about potential nonresponse bias can be gained by studying the absent students. NAEP scale score estimates are biased to the extent that assessed and absent students within the same weighting class differ in their distribution of scale scores. It seems likely that the assumption that absent students are similar in proficiency to assessed students is reasonable for some absent students namely, those whose absence can be characterized as random. Conversely, it seems likely that students with longer and more consistent patterns of absenteeism, such as truants, dropouts, near dropouts, and the chronically ill, are unlikely to be as proficient as their assessed counterparts.

In the 1998 assessments, schools were asked to classify each absent student into one of nine categories. The results of this classification for the 25-minute writing assessment are shown in Table 10-18. The discussion focuses on the 25-minute writing assessment because it is the largest. It is assumed that the other large assessments would behave similarly.

Table 10-18 shows that, as anticipated, the majority of absence from the assessment was the result of an absence from school of a temporary and unscheduled nature. The table shows that absence among twelfth-graders occurs at about four times the rate of absence among fourth-graders, and two-and-a-half times that of eighth-graders. The proportion of absence classified as temporary differs somewhat by grade, but is of the same magnitude for grades 8 and 12. These two facts taken together suggest strongly that a substantial proportion of the temporary absences among twelfth-grade students is not a result of illness, because such absences are occurring at almost three times the rate that they do among fourth- or eighth-grade students. Whereas it might be reasonable to regard temporary absence due to illness as independent of proficiency, for other temporary absences, this appears less tenable. The data in the table give support to the contention that, at grade 4, student absences are unlikely to introduce any significant bias into NAEP estimates. The absentee rate is low; most absences are temporary, and a third of the remaining absences are a result of parental refusal.

Table 10-18
Weighted Distribution of Absent Students by Nature of Absenteeism for All Grades, 1998 National 25-Minute Writing Samples

Nature of Absenteeism	Grade 4	Grade 8	Grade 12
Temporary Absence*	87.4%	74.6%	71.9%
Long-Term Absence†	0.7%	2.2%	0.8%
Chronic Truant	0.2%	1.6%	0.8%
Suspended or Expelled	0.9%	3.7%	0.4%
In School, Did Not Attend	0.2%	1.4%	8.3%
Disruptive Behavior	0.0%	0.4%	0.1%
Parent Refusal	4.1%	9.5%	3.5%
Student Refusal	0.2%	1.7%	7.4%
Missing	0.0%	0.0%	0.0%
Other, Specify on Cover	0.8%	2.0%	5.5%
Incorrectly Coded as Excluded	5.3%	2.8%	1.2%
Total Absentee Sample	1,067	1,731	5,017
Total Sample Size of Invited Students	20,883	22,317	24,522
Overall Absentee Rate, Unweighted	5.1%	7.8%	20.5%

* Absent less than two weeks due to illness, disability, or excused absence.
† Absent more than two weeks due to illness or disability.

At grades 8 and 12, however, a significant component of absenteeism is not temporary or due to parental refusal. Chronic truants, those suspended, and those in school but did not attend, and disruptive behavior constitute the obvious candidates for potential bias. These groups comprise 7.1 percent of absent students at grade 8 (or 0.6% of the total sample) and 9.6 percent of absent students at grade 12 (or 2.0% of the total sample). Thus their potential for introducing significant bias under the current procedures is minor.

10.5 VARIANCE ESTIMATION

A major source of uncertainty in the estimation of the value in the population of a variable of interest exists because information about the variable is obtained on only a sample from the population. To reflect this fact, it is important to attach to any statistic (e.g., a mean) an estimate of the sampling variability to be expected for that statistic. Estimates of sampling variability provide information about how much the value of a given statistic would be likely to change if the statistic had been based on another, equivalent, sample of individuals drawn in exactly the same manner as the achieved sample.

Another important source of variability is that due to imprecision in the measurement of individual scale scores. For the 1998 assessment, scale scores in all subject areas were summarized through item response theory (IRT) models, but not in the way that these models are used in standard applications where each person responds to enough items to allow for precise estimation of that person's scale score. In NAEP, each individual responds to relatively few items so that individual scale score values are not well determined. Consequently, the variance of any statistic based on scale score values has a component due to the imprecision in the measurement of the scale scores of the sampled individuals in addition to a component measuring sampling variability. The estimation of the component of variability due to measurement imprecision and its effect on the total variability of statistics based on scale score values are discussed in Chapter 12.

The estimation of the sampling variability of any statistic must take into account the sample design. In particular, because of the effects of cluster selection (students within schools, schools within PSUs) and because of effects of nonresponse and poststratification adjustments, observations made on different students cannot be assumed to be independent of each other (and are, in fact, generally positively correlated). Furthermore, to account for the differential probabilities of selection (and the various adjustments), each student has an associated sampling weight, which should be used in the computation of any statistic and is itself subject to sampling variability. Ignoring the special characteristics of the sample design and treating the data as if the observations were independent and identically distributed, will generally produce underestimates of the true sampling variability, due to the clustering and unequal sampling weights.

10.5.1 Procedure to Estimate Sampling Variability

The proper estimation of the sampling variability of a statistic based on the NAEP data is complicated and requires techniques beyond those commonly available in standard statistical packages. Fortunately, the jackknife procedure (see, e.g., Kish & Frankel, 1974; Rust, 1985; Wolter, 1985) provides good quality estimates of the sampling variability of most statistics, at the expense of increased computation, and can be used in concert with standard statistical packages to obtain a proper estimate of sampling variability.

The jackknife procedure used by NAEP has a number of properties that make it particularly suited for the analysis of NAEP data. When properly applied, a jackknife estimate of the variability of a linear estimator (such as a total) will be the same as the standard textbook variance estimate specified for the sample design (if the first-stage units were sampled with replacement and approximately so otherwise). Additionally, if the finite sampling corrections for the first-stage units can be ignored, the jackknife produces asymptotically consistent variance estimates for statistics such as ratios, regression estimates, or weighted means and for any other nonlinear statistic that can be expressed as a smooth function of estimated totals of one or more variables (Krewski & Rao, 1981).

Through the creation of student replicate weights (defined below), the jackknife procedure allows the measurement of variability attributable to the use of poststratification and other weight adjustment factors that are dependent on the observed sample data. Once these replicate weights are derived, it is a straightforward matter to obtain the jackknife variance estimate of any statistic.

The jackknife procedure in this application is based on the development of a set of jackknife replicate weights for each assessed student (or school depending on the file involved). The replicate weights are developed in such a way that, when utilized as described below, approximately unbiased estimates of the sampling variance of an estimate result, with an adequate number of degrees of freedom to be useful for purposes of making inferences about the parameter of interest.

The estimated sampling variance of a parameter estimator *t* is the sum of *M* squared differences (where *M* is the number of replicate weights developed):

$$\hat{Var}(t) = \sum_{i=1}^{M} (t_i - t)^2$$

where t_i denotes the estimator of the parameter of interest, obtained using the i^{th} set of replicate weights, $SRWT_i$, in place of the original sample of full sample estimates *FSTUWT*.

There were 62 replicate weights developed using the procedures outlined below. Full details of the generation of replicate weights for all samples are given in *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

Of the 62 replicate weights formed for each record from a national assessment sample, 36 act to reflect the amount of sampling variance contributed by the noncertainty strata of PSUs, with the remaining 26 replicate weights reflecting the variance contribution of the certainty PSU samples.

The derivation of the 36 replicate weights reflecting the variance of the noncertainty PSUs involves first defining pairs of PSUs in a manner that models the design as one in which two PSUs are drawn with replacement per stratum. This definition of pairs is undertaken in a manner closely reflective of the actual design, in that PSUs are pairs that are drawn from strata within the same subuniverse, and with similar stratum characteristics. The same definition of pairs was used for each of the age/grade classes in the national assessment, since all were drawn from the same sample of noncertainty PSUs. The 72 noncertainty PSUs, drawn one from each of 72 strata, were formed into 36 pairs of PSUs, where the pairs were composed of PSUs from adjacent strata within each subuniverse (thus the strata were relatively similar on socioeconomic characteristics such as proportion minority population, population change since 1980, per capita income, civilian unemployment rate, educational attainment, and unemployment rate). Whereas the actual sample design was to select one PSU with probability proportional to size from each of 72 strata, for variance estimation purposes the design is regarded as calling for the selection of two PSUs with probability proportional to size with replacement from each of 36 strata. This procedure likely gives a small positive bias to estimates of sampling error.

The student replicate weight for the i^{th} pair of noncertainty PSUs, for the 36 pairs corresponding to values of *i* from 1 to 36, is computed as follows:

1. Let W_B be the base weight of a student, as described in Section 10.2, which accounts for the various components of the selection probability for the student.

2. At random, one PSU in each pair is denoted as PSU number 1, while the other is denoted as PSU number 2. The i^{th} replicate base weight W_{Bi} is given by:

$$W_{Bi} = \begin{cases} 0 & \text{if the student belongs to PSU number 1 of pair } i \\ 2 \times W_B & \text{if the student belongs to PSU number 2 of pair } i \\ W_B & \text{if the student is from neither PSU in pair } i \end{cases}$$

3. The i^{th} student replicate weight $SRWT_i$ is obtained by applying the various school and student nonresponse adjustments, the weight trimming, and the poststratification to the i^{th} set of replicate base weights, using procedures identical to those used to obtain the final student weights WT from the set of base weights W_B.

In brief, the procedure for deriving the sets of W_{Bi} values from the W_B values reflects the sampling of PSUs, schools, sessions, and students. By repeating the various weight adjustment procedures in each set of replicate base weights, the impact of these procedures on the sampling variance of the estimator, t, is appropriately reflected in the variance estimator $\hat{V}ar(t)$ defined above.

The procedure for obtaining the 26 sets of replicate weights to estimate the sampling variance from the certainty PSUs is analogous, but somewhat more complex. The first stage of sampling in this case is at the school level, and the derivation of replicate weights must reflect appropriately the sampling of schools within certainty PSUs. Since each of the three grade classes in the national assessment involved different samples of schools, the procedure for forming replicate base weights was individualized to each of these sample components. In common across these three samples were the 22 certainty PSUs used, and the fact that 26 replicate weights were formed in each case.

For each grade, within the 22 certainty PSUs, a sample of schools was drawn systematically within each. Using the schools listed in order of sample selection within each of eight "combinations" of NAEP region and type of school (public, nonpublic), successive schools were grouped (i.e., PAIR). The number of variance groups within a combination depended on the number of schools in the combination, or indirectly assigned in proportion to the relative size of the combination. Thus, generally speaking, the largest combination were assigned the largest numbers of replicates (or pairs). When splitting the combinations, the schools were split into groups of (as close as possible) equal size, based on the ordering at the time of sample selection. One group was assigned to each replicate. Within each group in each combination, schools were alternately numbered 1 or 2 starting randomly. When, however, there were exactly three schools sampled in the variance group, the schools were randomly numbered 1, 2, or 3. The method of forming replicate base weights in variance groups (i.e., PAIR) where there were not exactly three schools was the same as for the noncertainty strata. If a variance group (PAIR) contained three schools, students in these schools had their weights perturbed for two sets of replicates, say i_1 and i_2, as follows:

$$W_{Bi_1} = \begin{cases} 0 & \text{if the student in school number 1 of a PSU in set } i \\ 1.5 \times W_B & \text{if the student in school number 2 or 3 of a PSU in set } i \\ W_B & \text{if the student does not belong to a PSU in set } i \end{cases}$$

$$W_{Bi_2} = \begin{cases} 1.5 \times W_B & \text{if the student in school number 1 or 2 of a PSU in set } i \\ 0 & \text{if the student in school number 3 of a PSU in set } i \\ W_B & \text{if the student does not belong to a PSU in set } i \end{cases}$$

The actual pattern of replicate base weight assignment used for each of the samples is given in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

The nonresponse, trimming, and poststratification adjustments were applied to each set of replicate base weights to derive the final replicate weights in each case, exactly as in the noncertainty PSUs. In fact, these procedures were applied to the full set of weights from all parts of the given sample together, just as for the full sample weights. That is, for example, poststratification factors were derived from the full set of data for each replicate, not separately for certainty and noncertainty PSUs.

This estimation technique was used by NAEP to estimate all sampling errors presented in the various reports. A further discussion of the variance estimation procedure used by NAEP, including a discussion of alternative jackknife estimators that were also considered, appears in Johnson (1989).

As stated above, a separate estimate of the contribution to variance due to the imprecision in the measure of individual proficiencies is made and added to the jackknife estimate of variance. That variance component could have been approximately reflected in the jackknife variance estimates simply by separately applying the IRT computations to each jackknife replicate. Because of the heavier IRT computational load, this was not done. Less work was involved by the simple procedure of making separate estimates of this component to be added to the jackknife variance estimates. Also, a separate measure of this component of variance is then available, which would not be so if it were reflected in the jackknife variance estimate.

10.5.2 Approximating the Sampling Variance Using Design Effects

In practical terms, the major expenditure of resources in the computation of a jackknife variance estimate occurs in the preparation of estimates for each of the pseudo-replicates. In the 1998 assessment, this implies that the statistic of interest has to be recomputed up to 63 times, once for the overall estimate t, and once for each of the up to 62 pseudo-replicates t_i. Because this is a considerable increase in the amount of computation required, relative to a conventional variance estimate, it is of interest to see how much the jackknife variance estimates differ from their less computationally intensive, simple random sampling based, analogues.

The comparison of the conventional and the jackknife methods of variance estimation will be in terms of a statistic called the *design effect*, which was developed by Kish (1965) and extended by Kish and Frankel (1974). The design effect for a statistic is the ratio of the actual variance of the statistic (taking the sample design into account) over the conventional variance estimate based on a simple random sample with the same number of elements. The design effect is the inflation factor to be applied to the conventional variance estimate in order to adjust error estimates based on simple random sampling assumptions to account approximately for the effect of the sample design. The value of the design effect depends on the type of statistic computed and the variables considered in a particular analysis as well as the combined clustering, stratification, and weighting effects occurring among sampled elements. While stratification drives down the sampling variance, the effects of clustering and weighting that drive variances up are generally sufficient to produce variance estimates that are larger than variances based on simple random sampling assumptions. Consequently, the design effects will be greater than one. In NAEP, the underestimates are the result of ignoring the effects of clustering and unequal probabilities of selection in the variance calculations.

Since most of the analyses conducted by NAEP are based on the results of scaling models that summarize performance of students across a learning area, design effects are expected for analyses based on these scale scores. For reasons given in Chapter 12, NAEP provides each individual with a set of "plausible values," each of which is a random draw from the distribution of the potential scale scores for

that individual. Since NAEP's current interest is on the effect of the sampling design on estimation and inference, attention is restricted to a single measure of an individual's scale score, the first plausible value of the individual's scale score.

A key statistic of interest is the estimated mean scale score of a subgroup of the population. An estimate of the subgroup mean scale score is the weighted mean of the first plausible values of scale score of the sampled individuals who belong to the subpopulation of interest. Let $\overline{Y}$ be the weighted mean of the plausible values of the sampled members of the subpopulation. The conventional estimate of the variance of $\overline{Y}$ is

$$Var_{con}\left(\overline{Y}\right) = \frac{\sum_{i=1}^{N} w_i \left(y_i - \overline{Y}\right)^2}{N \cdot W_+},$$

where N is the total number of sampled individuals in the subpopulation for which plausible values are available, w_i is the weight of the i^{th} individual, y_i is a plausible value from the distribution of potential proficiencies for that individual, and W_+ is the sum of the weights across the N individuals.

The design effect for the subgroup mean scale score estimate is

$$deff(\overline{Y}) = Var_{JK}(\overline{Y}) / Var_{con}(\overline{Y})$$

where $Var_{JK}(\overline{Y})$ is the jackknife variance of $\overline{Y}$ (As has been pointed out previously, $Var_{JK}(\overline{Y})$ as computed does not measure the variability of $\overline{Y}$ due to imprecision in the measurement of the proficiencies of the sampled individuals. The estimation of this very important source of variability is discussed in Chapter 12.) Of the factors that determine $deff(\overline{Y})$, the effects of stratification are usually less than one, which means the efficiency of a stratified sampling is better than a simple random sampling; whereas the clustering effects are always larger than one. The clustering effects can be approximated by

$$1 + (\overline{m} - 1)\rho,$$

where $\overline{m}$ is the average cluster size and ρ is the intracluster correlation (Cochran, 1977, p. 209). Therefore, the large cluster size or large intercluster correlation will inflate the clustering effects.

Values of the design effects for subgroup mean proficiencies are displayed, by grade, in Tables 10-19 through 10-21, for the 1998 national assessments of reading, writing, and civics, respectively. Design effects are shown for the population as a whole (Total) as well as for a variety of demographic subgroups: gender; race/ethnicity (White, Black, Hispanic, Asian American, other); type of location (central city, urban fringe/large town, rural/small town); parental education (did not graduate high school, graduated high school, post-high school, graduated college, unknown); and type of school (public, nonpublic). These particular demographic variables were selected because (1) they are major variables in NAEP reports and (2) they reflect different types of divisions of the population that might have different levels of sampling variability.

The tables show that the design effects are predominantly larger than 1, indicating that standard variance estimation formulas will be generally too small, usually markedly so. Although the design effects appear somewhat different for certain subgroups of the population, they are, perhaps, similar enough (at least within a subject and grade) to select an overall composite value that is adequate for most purposes. In choosing a composite design effect, some consideration must be made about the relative

consequences of overestimating the variance as opposed to underestimating the variance. For example, if an overestimate of the variance is viewed as severe an error as an underestimate, the composite design effect should be near to the center of the distributions of the design effects. Possible composites of this type are the mean and median design effects across the combined distribution of all design effects. Larger design effects should be used if it is felt that it is a graver error to underestimate the variability of a statistic than to overestimate it. For example, Johnson and King (1987) examine estimation of variances using design effects (among other techniques) under the assumption that the consequences of an underestimate are three times as severe as those of an overestimate of the same magnitude. Adopting a loss function that is a weighted sum of absolute values of the deviations of predicted from actual with underestimates receiving three times the weight of overestimates, produces the upper quartile of the design effects as the composite value. This assumes that the distribution of design effects is roughly independent of the jackknife estimates of variance, so that the size of a design effect does not depend on the size of the variance.

To compare Table 10-21 with Tables 10-19 and 10-20, the design effects for mean civics proficiencies are smaller than those of reading and writing. The reading reporting samples consist of non-SD/LEP students in sample types 2 and 3, and SD/LEP students in sample types 2. The intracluster correlation is larger for reading reporting samples that contain large groups of non-SD/LEP students. Therefore, the clustering effects for the reading reporting samples become larger than those of civics, which only used students in sample type 3.

Table 10-19
*Design Effects by Demographic Subgroup and Grade for Mean Reading Scale Scores**

	Grade 4	Grade 8	Grade 12
Total	3.15	5.30	3.98
Male	2.95	3.69	3.86
Female	1.38	3.14	2.09
White	2.55	4.55	2.96
Black	2.31	2.55	3.62
Hispanic	3.01	7.23	3.08
Asian American	1.35	7.62	4.53
Other race/ethnicity	1.50	2.30	1.57
Urban	6.12	7.81	8.11
Suburban	4.72	6.52	3.98
Rural	2.24	4.80	3.70
PARED < HS	1.00	2.22	1.74
PARED = HS	1.41	2.96	1.69
PARED > HS	0.92	2.47	1.77
PARED = College	2.68	2.72	2.15
PARED = Unknown	1.40	2.17	1.51
Public school	2.92	4.64	4.09
Nonpublic school	6.37	6.59	3.68

* Design effects are based on the conventional and jackknife variances of subgroup means of the first plausible values of scale score.

Table 10-20

*Design Effects by Demographic Subgroup and Grade for Mean Writing Scale Scores**

	Grade 4	Grade 8	Grade 12
Total	5.42	6.42	6.60
Male	3.48	5.11	4.14
Female	3.11	3.26	3.99
White	3.95	5.57	4.90
Black	1.88	2.53	5.01
Hispanic	5.76	5.45	3.02
Asian American	3.06	9.58	6.89
Other race/ethnicity	2.04	1.66	2.06
Urban	6.90	10.40	10.92
Suburban	5.95	12.95	8.88
Rural	6.48	4.74	2.42
PARED < HS	6.07	3.45	1.87
PARED = HS	1.65	1.40	1.71
PARED > HS	2.12	2.51	2.62
PARED = College	4.21	5.12	3.70
PARED = Unknown	1.45	1.14	1.38
Public school	5.80	5.71	7.09
Nonpublic school	4.59	5.33	5.60

* Design effects are based on the conventional and jackknife variances of subgroup means of the first plausible values of scale score.

Table 10-21

*Design Effects by Demographic Subgroup and Grade for Mean Civics Scale Scores**

	Grade 4	Grade 8	Grade 12
Total	2.34	3.23	3.70
Male	1.82	2.57	2.83
Female	1.48	1.95	2.36
White	2.24	3.25	3.39
Black	0.82	1.33	2.95
Hispanic	2.79	1.42	1.54
Asian American	0.94	8.44	6.41
Other race/ethnicity	1.41	1.02	1.78
Urban	2.15	3.67	4.52
Suburban	2.65	3.75	3.74
Rural	4.32	3.88	3.15
PARED < HS	1.35	3.66	1.19
PARED = HS	1.94	1.75	0.97
PARED > HS	1.34	1.84	2.07
PARED = College	1.83	2.16	2.5
PARED = Unknown	1.67	1.67	1.53
Public school	2.13	2.84	3.85
Nonpublic school	4.05	12.31	2.71

* Design effects are based on the conventional and jackknife variances of subgroup means of the first plausible values of scale score.

Table 10-22 gives the composite values of mean, median, and upper quartile of the distribution of design effects for mean scale score by grade for the reading, writing, and civics assessments, and across those assessments.

Table 10-22
Within-Grade Mean, Median, and Upper Quartile of the Distribution of Design Effects for 1998 National Assessments by Subject Area and Across Subject Areas

	Statistic	Grade 4	Grade 8	Grade 12
Distribution Across Demographic Subgroups				
Mean Reading Proficiencies				
	Upper Quartile	3.00	6.22	3.95
	Mean	2.67	4.40	3.23
	Median	2.43	4.12	3.35
Mean Writing Proficiencies				
	Upper Quartile	5.79	5.68	6.35
	Mean	4.11	5.13	4.60
	Median	4.08	5.12	4.07
Mean Civics Proficiencies				
	Upper Quartile	2.32	3.67	3.62
	Mean	2.07	3.37	2.84
	Median	1.89	2.71	2.77
Distribution Across Subject Areas and Demographic Subgroups				
Across Subject Areas				
	Upper Quartile	4.03	5.42	4.07
	Mean	2.95	4.30	3.56
	Median	2.33	3.56	3.12

* Design effects are based on the conventional and jackknife variances of subgroup means of the first plausible values of scale score.

The $Var_{con}(\bar{Y})$ as defined above is an estimate of S^2/N where S^2 represents the unit variance for a simple random sample for the population of students from which the sample is also drawn. This is an appropriate estimate of the increase in variance over simple random sampling from that population due to the effects of weighting. However, the computer packages used for estimating the variance may not reflect the weights in estimating the unit variance, as given above, but instead may provide an estimate of a unit variance of the form

$$\frac{1}{N(N-1)} \sum_{i=1}^{N} (y_i - \bar{Y})^2 .$$

In this case, the unweighted estimate of unit variance would be appropriate for the denominator of a design effect measure of the increase in variance over the unit variance as estimated by the computer package. If there is no correlation between the wi and yi, there would be little difference between the two.

Chapter 11

STATE WEIGHTING PROCEDURES AND VARIANCE ESTIMATION[1]

Jiahe Qian, Bruce A. Kaplan, and Eugene G. Johnson
Educational Testing Service

Ibrahim S. Yansaneh and Keith F. Rust
Westat

11.1 OVERVIEW

The 1998 state assessment program included samples of fourth- and eighth-grade students in public and nonpublic schools. The samples of students were selected using a complex multistage design involving the sampling of students from participating schools within each state. See Chapter 4 for a detailed description of the state sample design. Tables providing weighted counts of assessed and excluded students appear in this chapter. Supplemental data is provided in Appendix B tables.

The weighting process involved the development of survey weights for students, using data from a periodic assessment of students for each participating school in each of the states, territories, and military jurisdictions of the U.S. Following the collection of assessment and background data from and about assessed and excluded students, the processes of deriving sampling weights and associated sets of replicate weights were carried out. The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn. Replicate weights are used in the estimation of sampling variance, through a procedure known as jackknife repeated replication.

Weights were developed for students sampled at grades 4 and 8 for the state assessment in reading and at grade 8 for the state assessment in writing. Each student was assigned a weight to be used for making inferences about each state's students. This weight is known as the full-sample or overall sample weight. The full-sample weight contains five components. First, a base weight is established that is the inverse of the overall probability of selecting the sampled student. The base weight incorporates the probability of selecting a school and the student within a school. This weight is then adjusted for two sources of nonparticipation—school level and student level. These weighting adjustments seek to reduce the potential for bias from such nonparticipation by increasing the weights of students from schools similar to those schools not participating, and by increasing the weights of students similar to those students from within participating schools who did not attend the assessment session (or makeup session) as scheduled. Furthermore, the weights reflect the trimming of extremely large weights at each stage in the weighting process. For more detail on the implementation of these weighting steps, see Sections 11.2 and 11.3.

Section 11.4 addresses the effectiveness of the adjustments made to the weights using the procedures described in Section 11.3, examining characteristics of nonresponding schools and students, and investigating the extent to which nonrespondents differ from respondents in ways not accounted for

[1] Ibrahim Yansaneh and Keith F. Rust were responsible for the design and implementation of the weighting process for the 1998 NAEP state assessments. Jiahe Qian, with the assistance of Bruce Kaplan and in consultation with Eugene G. Johnson, was responsible for the planning, specification, and coordination of the state weighting at ETS. The statistical programming for this chapter was overseen by Bruce Kaplan and provided by Phillip Leung, Michael Narcowich, and Youn-Hee Lim.

in the weight adjustment procedures. Section 11.5 considers the distributions of the final student weights in each jurisdiction, and whether there were outliers that called for further adjustment.

In addition to the full-sample weights, a set of replicate weights was provided for each student. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method. Full details of the method of using these replicate weights to estimate sampling errors are contained in the *Technical Report of the NAEP 1994 Trial State Assessment Program in Reading* (Mazzeo, Allen, & Kline, 1995) and in earlier NAEP state technical reports. Section 11.6 of this report describes how the sets of replicate weights were generated for the 1998 state assessment data. The methods of deriving these weights were aimed at reflecting the features of the sample design appropriately in each jurisdiction, so that when the jackknife variance estimation procedure is implemented, approximately unbiased estimates of sampling variance are obtained.

As detailed in Chapter 5, two different sets of administration rules indicated by the sample type field were used in the 1998 state assessment program for reading. ETS raked the student weights for each subset to force agreement with the totals estimated using both subsets combined. This raking process is detailed in Section 11.7. The process of trimming extremely large raked student weights is also described.

11.2 CALCULATION OF BASE WEIGHTS

11.2.1 Calculation of School Base Weights

Base weights were assigned to schools separately by grade and subject. The base weight assigned to a school was calculated as the reciprocal of the overall probability of selection of that school. For the grade 8 samples, the school base weight depended on the assessment subject, because some schools were so small that students were tested in only one subject. For "new" schools selected using the supplemental new school sampling procedures (see Chapter 4), the school base weight reflected the combined probability of selection of the district, and school within district.

Thus the base weight for school *i* was calculated as

$$w_i^{sch} = \begin{cases} \dfrac{1}{Min\{EHIT,\ 1\}} & \text{for originally sampled schools; and} \\ \dfrac{1}{DISTPROB \times TCPNEW} & \text{for new schools} \end{cases}$$

where EHIT denotes the expected number of hits during sample selection; DISTPROB denotes the selection probability assigned to each sampled school district for updating purposes; and TCPNEW denotes the school probability of selection of new and newly eligible schools.

In each jurisdiction, all schools included in the sample with certainty were assigned school base weights of unity. Schools sampled with certainty were sometimes selected more than once in the systematic sampling process. For example, a school that was selected twice was allocated twice the usual number of students for the assessments, or two sessions; a school that was selected three times was allocated three times the usual number of students for the assessments, or three sessions. All schools at grade 8 with less than 20 students were assigned one subject (see Chapter 4). For these schools, the base weight included a factor of 2. Additional details about the weighting process are given in the sections below.

11.2.2 Weighting New Schools

New public schools were identified and sampled through a two-stage sampling process, involving the selection of districts, and then of new schools within selected districts. This process is described in Chapter 4. There were two distinct processes used depending upon the size of the district.

Within each jurisdiction, public school districts were partitioned into "small" districts—those having at most three schools on the aggregate frame and no more than one fourth-, one eighth-, and one twelfth-grade school. The remainder of the districts were denoted as "large" districts. For the larger districts (i.e., those having multiple schools in at least one of grades 4, 8, and 12), a sample of districts was selected in each jurisdiction. Districts in the sample were asked to identify schools having grade 4 or grade 8 that were not included on the school frame. A sample of these newly identified schools was then selected. The base weight for these schools reflected the probability of two factors: (i) that the district was selected for this updating process; and (ii) that the school was included in the NAEP sample, having been identified as new by the district. If the school was in grade 8 but was only large enough to assess one subject, the base weight included a factor of 2, as described in Section 11.2.1. There were no schools identified in small districts (see Tables 4-8 and 4-9).

11.2.3 Trimming School Base Weights for New Schools

The base weights for new schools were evaluated for possible trimming. The process involved computing a hypothetical school base weight for the new schools as though they had been selected as part of the original sample. The hypothetical base weight was then compared to the actual base weight. Those schools with actual base weights greater than three times the hypothetical base weights had their base weights trimmed to three times their hypothetical base weights.

The trimming factor was computed as

$$f_i = \begin{cases} \dfrac{3}{RSCHBWT} & \text{for new schools with } RSCHBWT > 3; \text{ and} \\ 1 & \text{for other new schools and for non-new schools;} \end{cases}$$

where RSCHBWT denotes the ratio of the school base weight to the hypothetical base weight.

The trimmed school base weight, denoted by w_i^{tsch}, was then defined as the product of the school base weight and the trimming factor. That is,

$$w_i^{tsch} = f_i \times w_i^{sch} .$$

Two schools had their weights trimmed as a result of this process. One of these schools is in a state that dropped out of the assessment. The other school has a trimming factor very close to 1, and therefore is not expected to have a significant impact on the weights.

11.2.4 Treatment of Substitute Schools

A school that replaced a refusing school (i.e., a substitute school) was assigned the weight of the refusing school. Thus the substitute school was treated as though it were the original school that it

replaced, for purposes of obtaining school base weights. The base weight was adjusted by a factor of 2 for grade 8 schools that were only large enough to assess one subject.

11.2.5 Calculation of Student Base Weights

Within the sampled schools, eligible students were sampled for assessment using the procedures described in Chapter 4. The within-school probability of selection for each subject therefore depended on the number of grade-eligible students in the school and the number of students selected for the assessment (usually 30). The within-school weights for sampled schools were adjusted to account for the fact that some schools operate twelve months per year and have only a proportion of their total enrollment attending school at any one time. For substitute schools, the within-school weights were further adjusted to compensate for differences in the grade enrollments of the substitute and the originally sampled (replaced) schools. In the case of eighth-grade schools, the within-school weight also incorporated a factor to account for (i) cases in which small schools were assigned at random to do one subject (reading or writing); and (ii) the random assignment of students to subjects. Thus, in general, the within-school student weight for the j^{th} student in school i was equal to:

$$W_{ij}^{within} = \frac{N_i}{n_i} \cdot K_{1i} \times K_{2i}$$

where

N_i = the number of grade-eligible students enrolled in the school, as reported at the time of student sampling; and

n_i = the number of students selected for the given subject.

The factors K_{1i} and K_{2i} in the formula for the within-school student weight generally apply to only a few schools in each jurisdiction. The factor K_{1i} adjusts the count of grade-eligible students in a substitute school to be consistent with the corresponding count of the originally sampled (replaced) school. Specifically, for substitute schools,

$$K_{1i} = \frac{E_i}{E_i^s}$$

with

E_i = the grade enrollment of the originally sampled (replaced) school; and

E_i^s = the grade enrollment of the substitute school.

For nonsubstitute schools, $K_{1i} = 1$.

The factor K_{2i}, which was applied to schools determined to be year-round schools, is defined as:

$$K_{2i} = \frac{1}{1 - p_{off}}$$

where p_{off} is the percentage of students enrolled in the school who were not scheduled to attend school at the time of assessment. For schools that are not year-round schools (the great majority), $K_{2i} = 1$.

The overall student base weight for a student *j* selected for the assessment for a given subject (reading or writing) in school *i* was obtained by multiplying the trimmed school base weight by the within-school student weight and therefore was computed as:

$$W_{ij}^{base} = W_i^{tsch} \times W_{ij}^{within} \quad .$$

11.3 ADJUSTMENTS FOR NONRESPONSE

As mentioned earlier, the base weight for a student was adjusted by two factors: one to adjust for nonparticipating schools for which no substitute participated, and another to adjust for students who were invited to the assessment but did not attend the scheduled sessions (original or makeup).

11.3.1 Defining Initial School-Level Nonresponse Adjustment Classes

School-level nonresponse adjustment classes were created separately for public and nonpublic schools within each jurisdiction. For each set, these classes were defined as a function of their sampling strata as follows.

Public Schools. For each jurisdiction, except Virgin Islands, DoDEA/DDESS[2], and DoDEA/DoDDS[3], the initial school nonresponse adjustment classes were formed by cross classifying the level of urbanization and minority status (see Chapter 4 for definitions of these characteristics). Where there was only one minority status category within a particular level of urbanization, a categorized version of median household income was crossed with the urbanization category. For this purpose within each level of urbanization, public schools were sorted by the median household income, and then divided into three groups of about equal size, representing low, middle, and high income areas. In Virgin Islands, there was no information on minority status or median household income. Thus, for Virgin Islands, at grade 4 a categorized version of estimated grade enrollment was used, and at grade 8, due to the small number of schools, all schools were placed in the same initial nonresponse adjustment cell. In all cases, for schools with SD/LEP students, sample type (whether accommodations were offered or not) was used in addition to the variables described above.

Department of Defense Education Activity/Department of Defense Domestic Elementary Schools (DoDEA/DDESS) and Department of Defense Education Activity/Department of Defense Dependents Schools (DoDEA/DoDDS). For the jurisdictions comprising DoDEA/DDESS and DoDEA/DoDDS schools, urbanization, median income, and metro status were not available. Therefore, the initial school nonresponse adjustment classes were defined by the state or district code, except for DoDEA/DDESS grade 8, which had only one adjustment cell due to the small number of schools. Again, sample type was used in addition to the variables described above.

Nonpublic Schools. For each jurisdiction (excluding Virgin Islands nonpublic schools), initial nonresponse adjustment classes were formed by cross classifying school type (Catholic and non-Catholic) and metropolitan status (the urban/rural nature of the place where the school is located). For Virgin Islands, urban/rural status was not available, so only school type was used. For schools with SD/LEP students, sample type was used in addition to the variables described above.

[2] Department of Defense Education Activity/Department of Defense Domestic Elementary and Secondary Schools
[3] Department of Defense Education Activity/Department of Defense Dependents Schools

11.3.2 Constructing the Final Nonresponse Adjustment Classes

The objective in forming the nonresponse adjustment classes is to create as many classes as possible that are internally as homogeneous as possible, but such that the resulting nonresponse adjustment factors are not subject to large random variation. Consequently, all initial nonresponse adjustment classes deemed unstable were collapsed with suitable neighboring classes so that: (i) the combined class contained at least six sessions, and (ii) the resulting nonresponse adjustment factor did not exceed 1.35. (In a few cases, a factor in excess of 1.35 was permitted). When 100 percent of the public schools in a jurisdiction responded, no action was taken for a public-school adjustment class that contained fewer than six sessions. The same approach was used for nonpublic schools where 100 percent of the schools participated. Although there is clearly no adjustment for school nonresponse in these cases, this procedure could have an effect on the final definition of the student nonresponse adjustment classes (see Section 11.3.4).

Public Schools. For public schools, inadequate nonresponse adjustment classes were reinforced by collapsing adjacent levels of minority status (or median household income level if minority information was missing). Metropolitan and non-metropolitan schools were combined together in cases where there were less than six cooperating schools after collapsing across all levels of minority status (or median household income levels, if minority status information was missing) that were not mixed. No collapsing was done across sample type.

Nonpublic Schools. For nonpublic schools in all states except Virgin Islands, inadequate classes were reinforced by collapsing adjacent levels of metropolitan-area status within school type. Catholic and non-Catholic schools were kept apart to the extent possible, particularly when the only requirement to combine such schools was as a means of reducing the adjustment factors below 1.35. For nonpublic schools in Virgin Islands, Catholic and non-Catholic schools were collapsed together in order to form a stable nonresponse adjustment class.

11.3.3 School Nonresponse Adjustment Factors

The school-level nonresponse adjustment factor for the i^{th} school in the h^{th} class was computed as:

$$F_h^{(1)} = \frac{\sum_{i \in C_h} W_{hi}^{sch} \times E_{hi}}{\sum_{i \in C_h} W_{hi}^{sch} \times E_{hi} \times \delta_{hi}}$$

where

C_h = the subset of school records in class h,

W_{hi}^{sch} = the base weight of the ith school in class h,

E_{hi} = the grade enrollment for the ith school in class h,

$$\delta_{hi} = \begin{cases} 1 & \text{if the } i\text{th school in adjustment class } h \text{ participated in the assessments; and} \\ 0 & \text{otherwise.} \end{cases}$$

Both the numerator and denominator of the nonresponse adjustment factor contained only schools that were determined to have eligible students enrolled.

In the calculation of the above nonresponse adjustment factors, a school was said to have participated if:

- it was selected for the sample from the frame or from the lists of new schools provided by participating school districts, and student assessment data were obtained from the school; or
- the school participated as a substitute school and student assessment data were obtained (so that the substitute participated in place of the originally selected school).

The nonresponse-adjusted weight for the ith school in class *h* was computed as:

$$W_{hi}^{adj} = F_h^{(1)} \times W_{hi}^{sch} \ .$$

11.3.4 Student Nonresponse Adjustment Classes

The initial student nonresponse classes for assessed students were formed based on several variables. These variables are based on information from the sample design, age of the student, final collapsed school nonresponse cells, and the actual monitor status (or assigned monitor status, if the actual monitor status is not available; see Chapter 4) at the session level. The first of these was public/nonpublic strata and an indicator of whether or not a student was excluded from the assessment. Public/nonpublic strata were then cross classified by a variable created from combining SD/LEP status and the sample type for the student.

Within these categories, the initial student nonresponse adjustment classifications were defined further depending on the SD/LEP status of a student. For all schools except DoDEA/DDESS and DoDEA/DoDDS, if a student was SD or LEP, then the class was formed by urbanization cross classified by student age. Age was used to classify students into two groups (for grade 4, those born in September 1987 or earlier and those born in October 1987 or later, and for grade 8, those born in September 1983 or earlier and those born in October 1983 or later). If a student was neither SD nor LEP, then the initial nonresponse adjustment class was formed by urbanization cross classified by student age (as defined above), by the quality control monitoring status (see Chapter 4), then finally by minority status as collapsed for the school nonresponse. For the DoDEA/DDESS and DoDEA/DoDDS schools, the nonresponse adjustment classes for SD and LEP students was student age cross classified by the minority status variable as defined for the school nonresponse adjustment classes.

Following creation of these student nonresponse adjustment classes, all unstable classes were identified for possible collapsing with other classes. A class was considered to be unstable when either of the following conditions was true for the given class:

- number of responding eligible students was fewer than 20, or
- nonresponse adjustment factor exceeded 1.5.

All classes deemed unstable in the previous step were collapsed with other classes using the following rules:

- Do not collapse across public and nonpublic.
- Do not collapse across SD/LEP and non-SD–non-LEP.
- If within cells defined by the cross classification of public/nonpublic and SD-LEP/non-SD–non-LEP status, and sample type within the SD/LEP categories, all of the adjustments are one, no adjustments are made.
- Collapse across the last variable of the nonresponse adjustment cell only (i.e., collapse across geography for SD/LEP students in Department of Defense Education Activity (DoDEA) schools).

More collapsing was necessary only if the resulting classes had fewer than 15 responding eligible students. Collapsing then continued within the successive variables until the class size was no longer deficient or until a "set" boundary that could not be crossed was reached. In the case of SD or LEP students, more collapsing was done to eliminate the rare situation in which all students in a class were nonrespondents.

11.3.5 Student Nonresponse Adjustments

As described above, the student-level nonresponse adjustments for the assessed students were made within classes defined by the SD/LEP status, sample type, final school-level nonresponse adjustment classes, monitoring status of the school, and age group of the students. Subsequently, in each jurisdiction, the final student weight for the j^{th} student of the i^{th} school in class k was then computed as:

$$W_{kij}^{final} = W_i^{adj} \times W_{ij}^{within} \times F_k$$

where

W_i^{adj} = the nonresponse-adjusted school weight for school i;

W_{ij}^{within} = the within-school weight for the jth student in school i; and

$$F_k = \frac{\sum_j W_{ij}}{\sum_j W_{ij}\delta_{kj}} \quad .$$

In the above formulation, the summation included all students, j, in the k^{th} final (collapsed) nonresponse class. The indicator variable δ_{kj} had a value of 1 when the jth student in adjustment class k participated in the assessment; otherwise, $\delta_{kj} = 0$.

For excluded students, no nonresponse adjustment procedures were applied because excluded students were not required to complete an assessment. In effect, all excluded students were considered respondents. Weights are provided for excluded students so as to estimate the size of this group and its population characteristics. Tables 11-1 through 11-6 summarize the unweighted and final weighted counts of assessed and excluded students in public and nonpublic schools for each jurisdiction, grade and subject.

Table 11-1

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 4 Public Schools, 1998 Reading State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Total	109,148	2,646,973	9,186	260,558	118,334	2,907,530
Alabama	2,559	56,372	239	4,922	2,798	61,294
Arizona	2,602	55,867	318	6,349	2,920	62,216
Arkansas	2,656	30,773	144	1,613	2,800	32,386
California	1,898	372,225	384	65,127	2,282	437,352
Colorado	2,656	49,221	195	3,309	2,851	52,530
Connecticut	2,607	38,543	379	4,971	2,986	43,514
Delaware	2,483	8,171	127	381	2,610	8,552
District of Columbia	2,464	4,691	284	504	2,748	5,194
DoDEA/DDESS	2,693	2,821	128	128	2,821	2,949
DoDEA/DoDDS	2,670	6,310	105	234	2,775	6,545
Florida	2,658	154,056	224	12,220	2,882	166,276
Georgia	2,733	96,499	179	6,058	2,912	102,557
Hawaii	2,742	13,548	144	676	2,886	14,224
Illinois	2,264	124,291	200	10,148	2,464	134,439
Iowa	2,339	33,263	171	2,324	2,510	35,587
Kansas	1,922	32,925	104	1,657	2,026	34,582
Kentucky	2,508	41,123	233	3,661	2,741	44,784
Louisiana	2,701	51,743	308	5,741	3,009	57,484
Maine	2,464	15,635	231	1,294	2,695	16,929
Maryland	2,344	57,644	204	4,894	2,548	62,538
Massachusetts	2,478	70,290	188	5,222	2,666	75,512
Michigan	2,416	116,655	179	8,068	2,595	124,723
Minnesota	2,425	61,069	94	2,179	2,519	63,248
Mississippi	2,591	36,430	118	1,565	2,709	37,995
Missouri	2,599	60,008	206	4,488	2,805	64,496
Montana	1,936	11,065	67	360	2,003	11,425
Nevada	2,732	20,105	388	2,652	3,120	22,757
New Hampshire	1,908	15,509	91	671	1,999	16,180
New Mexico	2,550	21,238	330	2,521	2,880	23,759
New York	2,318	192,009	196	16,046	2,514	208,055
North Carolina	2,628	87,078	265	8,222	2,893	95,300
Oklahoma	2,647	43,087	303	4,366	2,950	47,453
Oregon	2,550	36,836	192	2,597	2,742	39,433
Rhode Island	2,698	11,139	221	844	2,919	11,983
South Carolina	2,518	43,925	273	4,493	2,791	48,418
Tennessee	2,735	66,272	120	2,737	2,855	69,009
Texas	2,443	249,823	383	37,861	2,826	287,684

(continued)

Table 11-1 (continued)

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 4 Public Schools, 1998 Reading State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Utah	2,784	31,657	185	1,903	2,969	33,560
Virgin Islands	1,485	1,552	95	95	1,580	1,647
Virginia	2,723	76,981	228	6,123	2,951	83,104
Washington	2,491	67,261	137	3,662	2,628	70,923
West Virginia	2,568	19,137	271	1,868	2,839	21,005
Wisconsin	2,183	55,418	245	5,548	2,428	60,966
Wyoming	2,779	6,708	110	257	2,889	6,965

Table 11-2

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Public Schools, 1998 Reading State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Total	93,223	2,441,495	6,068	151,260	99,291	2,592,754
Alabama	2,490	54,366	177	3,718	2,667	58,084
Arizona	2,529	53,001	183	3,376	2,712	56,377
Arkansas	2,489	32,855	170	2,056	2,659	34,911
California	2,182	364,480	159	23,908	2,341	388,388
Colorado	2,673	49,634	133	2,270	2,806	51,904
Connecticut	2,617	35,939	214	2,655	2,831	38,594
Delaware	2,081	8,220	122	399	2,203	8,618
District of Columbia	1,589	3,967	142	306	1,731	4,273
DoDEA/DDESS	630	1,324	28	56	658	1,380
DoDEA/DoDDS	2,221	4,746	61	122	2,282	4,868
Florida	2,545	147,121	145	7,863	2,690	154,984
Georgia	2,600	95,969	146	4,870	2,746	100,839
Hawaii	2,602	12,468	163	715	2,765	13,183
Illinois	2,148	127,567	117	6,459	2,265	134,026
Kansas	1,932	34,261	105	1,574	2,037	35,835
Kentucky	2,342	44,684	105	1,943	2,447	46,627
Louisiana	2,585	50,192	228	3,982	2,813	54,174
Maine	2,474	15,471	164	963	2,638	16,434
Maryland	2,178	54,030	123	2,738	2,301	56,768
Massachusetts	2,306	60,590	148	3,546	2,454	64,136
Minnesota	2,039	63,573	61	1,669	2,100	65,242
Mississippi	2,332	33,909	173	2,363	2,505	36,272
Missouri	2,632	63,890	142	3,288	2,774	67,178

(continued)

Table 11-2 (continued)

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Public Schools, 1998 Reading State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Montana	1,946	12,021	82	412	2,028	12,433
Nevada	2,564	18,154	200	1,319	2,764	19,473
New Mexico	2,365	21,623	239	1,885	2,604	23,508
New York	1,923	181,223	208	17,019	2,131	198,242
North Carolina	2,595	81,637	222	6,317	2,817	87,954
Oklahoma	2,234	42,355	236	4,081	2,470	46,436
Oregon	2,294	38,419	105	1,498	2,399	39,917
Rhode Island	2,513	10,591	160	596	2,673	11,187
South Carolina	2,509	45,583	169	2,765	2,678	48,348
Tennessee	2,245	58,759	122	2,975	2,367	61,734
Texas	2,500	248,845	175	16,047	2,675	264,892
Utah	2,601	34,340	133	1,548	2,734	35,888
Virgin Islands	643	1,464	54	108	697	1,572
Virginia	2,592	73,995	187	4,824	2,779	78,819
Washington	2,323	69,342	104	2,856	2,427	72,198
West Virginia	2,537	20,565	239	1,756	2,776	22,321
Wisconsin	1,997	62,606	152	4,234	2,149	66,840
Wyoming	2,626	7,716	72	183	2,698	7,899

Table 11-3

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Public Schools, 1998 Writing State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Total	91,996	2,429,504	4,872	124,329	96,868	2,553,832
Alabama	2,449	53,997	169	3,521	2,618	57,518
Arizona	2,499	53,315	162	2,992	2,661	56,307
Arkansas	2,462	32,430	162	1,945	2,624	34,375
California	2,157	359,589	155	23,418	2,312	383,007
Colorado	2,697	50,662	117	1,914	2,814	52,576
Connecticut	2,592	36,138	221	2,786	2,813	38,924
Delaware	2,119	8,265	80	269	2,199	8,533
District of Columbia	1,592	4,007	130	276	1,722	4,283
DoDEA/DDESS	650	1,362	19	38	669	1,400
DoDEA/DoDDS	2,182	4,704	34	68	2,216	4,772
Florida	2,574	150,236	130	7,085	2,704	157,321
Georgia	2,605	96,368	138	4,599	2,743	100,967

(continued)

Table 11-3 (continued)

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Public Schools, 1998 Writing State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Hawaii	2,647	12,619	123	522	2,770	13,141
Illinois	2,145	129,782	95	5,263	2,240	135,045
Kentucky	2,341	44,823	66	1,145	2,407	45,968
Louisiana	2,653	51,962	158	2,882	2,811	54,844
Maine	2,508	15,659	148	860	2,656	16,519
Maryland	2,263	55,675	55	1,216	2,318	56,891
Massachusetts	2,399	62,177	131	3,091	2,530	65,268
Minnesota	1,980	63,353	65	1,884	2,045	65,237
Mississippi	2,401	35,008	130	1,708	2,531	36,716
Missouri	2,621	63,703	79	1,747	2,700	65,450
Montana	2,024	12,492	62	319	2,086	12,811
Nevada	2,553	18,325	181	1,167	2,734	19,492
New Mexico	2,426	22,277	192	1,476	2,618	23,753
New York	1,981	189,995	123	10,306	2,104	200,301
North Carolina	2,669	83,857	127	3,673	2,796	87,530
Oklahoma	2,258	42,418	239	4,054	2,497	46,472
Oregon	2,323	38,838	90	1,251	2,413	40,089
Rhode Island	2,516	10,584	129	488	2,645	11,072
South Carolina	2,469	45,294	160	2,619	2,629	47,913
Tennessee	2,275	59,184	104	2,536	2,379	61,720
Texas	2,530	250,733	169	15,518	2,699	266,251
Utah	2,588	34,091	117	1,355	2,705	35,446
Virgin Islands	614	1,412	59	118	673	1,530
Virginia	2,605	74,518	131	3,392	2,736	77,910
Washington	2,286	68,730	96	2,637	2,382	71,367
West Virginia	2,611	21,219	157	1,127	2,768	22,346
Wisconsin	2,006	62,152	105	2,895	2,111	65,047
Wyoming	2,726	7,551	64	169	2,790	7,720

Table 11-4

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 4 Nonpublic Schools, 1998 Reading State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Total	8,101	210,902	64	2,131	8,165	213,033
Arkansas	166	2,386	0	0	166	2,386
Colorado	225	4,599	2	54	227	4,653
Connecticut	263	4,214	2	26	265	4,241
Florida	274	20,284	1	67	275	20,351
Georgia	270	6,631	6	113	276	6,744
Hawaii	379	2,000	0	0	379	2,000
Illinois	355	25,870	3	194	358	26,064
Iowa	330	4,257	1	17	331	4,274
Louisiana	425	10,462	4	120	429	10,582
Maine	131	917	0	0	131	917
Maryland	297	8,750	3	115	300	8,865
Massachusetts	284	8,951	5	156	289	9,106
Michigan	265	15,375	3	160	268	15,535
Minnesota	338	8,426	1	22	339	8,448
Mississippi	224	3,763	0	0	224	3,763
Missouri	320	9,621	2	74	322	9,695
Montana	102	466	1	4	103	471
Nebraska	478	3,063	3	21	481	3,083
Nevada	150	962	1	6	151	968
New Mexico	249	2,350	8	83	257	2,433
New York	377	36,271	5	398	382	36,669
North Carolina	236	6,773	0	0	236	6,773
Rhode Island	382	1,506	0	0	382	1,506
South Carolina	227	3,951	2	31	229	3,983
Utah	107	681	0	0	107	681
Virgin Islands	426	461	0	0	426	461
Washington	175	4,965	0	0	175	4,965
West Virginia	125	973	0	0	125	973
Wisconsin	426	11,710	10	463	436	12,173
Wyoming	95	266	1	4	96	271

Table 11-5

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Nonpublic Schools, 1998 Reading State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Total	5,554	182,810	43	1,000	5,597	183,810
Arkansas	133	1,754	2	33	135	1,787
Arizona	176	6,072	6	223	182	6,294
California	295	44,862	0	0	295	44,862
Colorado	154	2,310	0	0	154	2,310
Connecticut	371	5,143	3	50	374	5,192
Florida	190	14,159	1	45	191	14,204
Georgia	185	7,090	0	0	185	7,090
Illinois	289	20,787	1	78	290	20,865
Louisiana	459	10,267	2	47	461	10,314
Massachusetts	185	5,986	0	0	185	5,986
Maryland	329	8,021	0	0	329	8,021
Maine	78	535	0	0	78	535
Missouri	297	7,199	0	0	297	7,199
Mississippi	0	0	0	0	0	0
Montana	147	646	0	0	147	646
North Carolina	238	5,032	3	75	241	5,107
Nebraska	366	2,950	4	33	370	2,982
New Mexico	170	1,471	9	67	179	1,539
Nevada	130	943	1	11	131	954
New York	351	29,209	3	244	354	29,453
Rhode Island	403	1,507	5	19	408	1,527
Virgin Islands	228	394	0	0	228	394
Washington	230	5,284	3	76	233	5,360
West Virginia	99	1,041	0	0	99	1,041
Wyoming	51	149	0	0	51	149

Table 11-6

Unweighted and Final Weighted Counts of Assessed and Excluded Students by Jurisdiction, Grade 8 Nonpublic Schools, 1998 Writing State Samples

	Assessed		Excluded		Assessed and Excluded	
Jurisdiction	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**	**Unweighted**	**Weighted**
Total	5,593	173,497	27	960	5,620	174,457
Arkansas	140	2,143	1	13	141	2,155
Arizona	130	3,234	11	306	141	3,540
California	224	30,585	0	0	224	30,585
Colorado	137	2,916	0	0	137	2,916
Connecticut	240	4,151	2	30	242	4,180
Florida	213	13,409	1	42	214	13,451
Georgia	144	6,246	1	35	145	6,281
Illinois	314	23,623	0	0	314	23,623
Louisiana	580	11,449	0	0	580	11,449
Massachusetts	263	8,395	1	28	264	8,423
Maryland	350	9,168	0	0	350	9,168
Maine	95	831	0	0	95	831
Missouri	303	9,843	0	0	303	9,843
Montana	206	853	1	5	207	858
North Carolina	248	6,142	3	50	251	6,192
Nebraska	354	2,835	0	0	354	2,835
New Mexico	204	1,842	2	12	206	1,854
Nevada	108	730	0	0	108	730
New York	380	27,993	4	439	384	28,432
Rhode Island	434	1,680	0	0	434	1,680
Virgin Islands	193	383	0	0	193	383
Washington	155	3,824	0	0	155	3,824
West Virginia	117	977	0	0	117	977
Wyoming	61	246	0	0	61	246

11.4 CHARACTERISTICS OF NONRESPONDING SCHOOLS AND STUDENTS

In the previous section, procedures were described for adjusting the survey weights so as to reduce the potential bias of nonparticipation of sampled schools and students. To the extent that the characteristics of nonresponding schools or students are different from those of respondents in the same nonresponse adjustment class, potential for nonresponse bias remains. Recently, some studies related with this issue have been done, such as on the effects of excluded students in reporting results (see Donoghue, 2000).

This section examines the potential for remaining nonresponse bias in two related ways. First, weighted distributions for each grade and subject within each jurisdiction of certain characteristics of schools and students, both for the full sample and for respondents only, are discussed. This analysis is of necessity limited to those characteristics that are known for both respondents and nonrespondents, and hence, cannot directly address the question of nonresponse bias. The approach taken does reflect the reduction in bias obtained through the use of nonresponse weighting adjustments. As such, it is more

appropriate than a simple comparison of the characteristics of nonrespondents with those of respondents for each subject and jurisdiction.

The second approach involves modeling the probability that a school is a respondent, as a function of the nonresponse adjustment class to which the school belongs, together with other school characteristics. This was achieved using linear logistic regression models, with school response status as the dependent variable. By testing to see if the school characteristics add any predictive ability to the model over using the membership of the nonresponse adjustment class to make this prediction, researchers can obtain some insight into the remaining potential for nonresponse bias. If these factors are substantially marginally predictive, there is danger that significant nonresponse bias will remains. See Section 11.4.2 for details on how this approach was implemented.

11.4.1 Weighted Distributions of Schools Before and After School Nonresponse

To study the potential for nonresponse bias, Westat analysts compared the school characteristics before and after school nonresponse for public schools. For public schools, the variables for which means are presented are the percentage of Black students in the school, the percentage of Hispanic students, the median household income (1989) of the ZIP code area where the school is located, and the type of location. The first two variables were obtained from the sample frame, and hence from Quality Education Data, Inc., (QED) as described in Chapter 4. Median income was obtained from the 1990 Donnelly File. The variable designating type of location was derived for each sampled school using U.S. Bureau of Census data. The type of location variable has seven possible levels, which are defined in Chapter 4. Although this variable is not interval-scaled, the mean value does give an indication of the degree of urbanization of the population represented by the school sample (lower values for type of location indicate a greater degree of urbanization).

For public schools, the mean values of the variables, both before and after nonresponse, were calculated for all jurisdictions in reading grades 4 and 8, and writing grade 8. The means are weighted appropriately to reflect whether nonresponse adjustments have been applied (i.e., to respondents only) or not (to the full set of in-scope schools). The tables are presented in Appendix B. For each grade and subject, two sets of means are presented for these four variables. The first set shows the weighted mean derived from the full sample of in-scope schools selected for each subject, that is, respondents and nonrespondents (for which there was no participating substitute). The weight for each sampled school is the product of the school base weight and the grade enrollment. This weight therefore represents the number of students in the state represented by the selected school. The second set of means is derived from responding schools only, after school substitution. In this case the weight for each school is the product of the nonresponse-adjusted school weight and the grade enrollment of the original school, and therefore indicates the number of students in the jurisdiction represented by the responding school.

The characteristics of interest for nonpublic schools were the proportion of Catholic schools and the proportion of schools that are located in urban districts. As was done for public schools, two sets of means are presented: the means for the full sample and for the responding sample.

For both public and nonpublic schools, the differences between these sets of means give an indication of the potential for nonresponse bias that has been introduced by nonresponding schools with no participating substitute. For example, for grade 4 reading in Illinois, the mean percentage Black enrollment, estimated from the original sample of public schools, is 20.92 percent. The estimate from the responding schools is 26.33 percent. Thus there may be a slight bias in the results for Illinois because these two means differ. Note, however, that the differences in the two sets of mean values are generally very slight, at least in absolute terms, suggesting that it is unlikely that substantial bias has been introduced by schools that did not participate and for which no substitute participated. Of course in a

number of states there was no nonresponse at the school level (weighted participation rate is 100%), so that these sets of means are identical. Even in those jurisdictions where school nonresponse was relatively high (such as in New Hampshire grade 4 reading, Minnesota grade 8 writing, and Wisconsin grade 8 reading and writing), the absolute differences in means are slight. Occasionally the relative difference is large, for instance, the "Percent Black" in Illinois for both grade 4 and grade 8 reading (for public schools), or West Virginia grade 4 reading, Wyoming grade 4 reading, and New York grade 8 reading (for nonpublic schools). However, these are for small population subgroups, and thus are very unlikely to have a large impact on results for the jurisdiction as a whole.

11.4.2 Characteristics of Schools Related to Response

In an effort to evaluate the possibility that substantial bias remains as a result of school nonparticipation, following the use of nonresponse adjustments, a series of analyses were conducted on the response status for public schools. These analyses were restricted to those jurisdictions with a participation rate of below 90 percent (after substitution), because these are the jurisdictions where the potential for nonresponse bias was likely to be the greatest. Jurisdictions with an initial public-school response rate below 70 percent were not included, since NAEP does not report results for these jurisdictions because of concern about nonresponse bias. Information about this can be found in Chapters 17 and 21. Nonpublic schools were omitted from these analyses as well because of the small sample sizes involved, meaning that it is difficult to assess whether a potential for bias exists. Table 11-7 gives each participating states' participation rate as included in the analysis for each grade and subject.

Table 11-7
Jurisdictions Included in Logistic Regression Analysis of the NAEP 1998 State Assessment

Grade	Subject	Jurisdiction	Participation Rate
4	Reading	CA	80%
		IL	84%
		IA	84%
		KS	70%
		MD	88%
		MA	88%
		MN	86%
		MT	78%
		NH	70%
		NY	84%
		WA	89%
		WI	82%
8	Reading	CA	84%
		IL	81%
		KS	71%
		KY	87%
		MD	85%
		MA	89%
		MN	74%

(continued)

Table 11-7 (continued)
Jurisdictions Included in Logistic Regression Analysis of the NAEP 1998 State Assessment

Grade	Subject	Jurisdiction	Participation Rate
8	Writing	MT	78%
		NY	77%
		OR	88%
		TN	89%
		WA	86%
		WI	73%
		CA	83%
		IL	80%
		KY	87%
		MD	86%
		MA	89%
		MN	74%
		MT	78%
		NY	77%
		OR	88%
		TN	89%
		WA	87%
		WI	73%

The approach used was to develop a logistic regression model to predict the probability of participation as a function of the nonresponse adjustment classes and other school characteristics. These models were developed for public schools in each of the jurisdictions and for each grade and subject specified in the above table. For the three grade-subject combinations, this resulted in the development of 37 models, which differ only in the number of nonresponse class levels that are included in the model. The number of final nonresponse adjustment classes varied by state. The logistic regression analysis was used to determine whether the response rates are significantly related to school characteristics, after accounting for the effect of the nonresponse class. Thus, "dummy" variables were created to indicate nonresponse class membership.

If there are *k* nonresponse classes within a jurisdiction, for nonresponse class $i = 1, ..., k-1$, let

X_{ij} = 1 if the school *j* is classified in nonresponse class *i*,

= 0 otherwise.

Within each jurisdiction, a logistic model was fitted to the data on public-school participation. In the model, the indicator variables for nonresponse class, and additional variables available for participating and nonparticipating schools alike were included. These variables are denoted as Y_{ij}, for *i* from 1 to 4 of school j. They were the percentage of Black students (Y_{1j}), the percentage of Hispanic students (Y_{2j}), the estimated enrollment for grades 4 and 8 of the school (Y_{3j}), and the median household income of the ZIP code area in which the school was located (Y_{4j}).

Let P_j denote the probability that school j is a participant, and let L_j denote the logit of P_j. That is,

$$L_j = \ln\left(\frac{P_j}{1-P_j}\right).$$

The model fitted in each jurisdiction was the following:

$$L_j = A + \sum B_i X_{ij} + \sum C_i Y_{ij},$$

where A, B_i, and C_i are the coefficients of the logistic regression model.

Note that this model cannot be estimated if there are nonresponse classes in which all schools participated (so that no adjustments for nonresponse were made for those schools). Even though this analysis was restricted to those jurisdictions with relatively poor response, unestimatable cases occurred in a number of instances. When this happened, those (responding) schools in such classes were dropped from the analyses. Tables 11-8, through 11-10 show the proportion of the state public-school student population that is represented in the sample by schools from classes with less than 100 percent response for each grade and subject. Thus in grade 4 reading for Illinois, Kansas, and New Hampshire, there was some nonresponse within every adjustment class, whereas for the other nine states in grade 4, some portion of the population is not represented because schools were dropped from classes with no nonresponse. The states in which the entire student population is represented in the sample by schools from classes with less than 100% response are Illinois, Kansas, Minnesota, New York, and Wisconsin for grade 8 reading; and Illinois, Minnesota, New York, and Wisconsin for grade 8 writing. For the rest of the states, in both grades, some portion of the student population is not represented because schools were dropped from classes with no nonresponse.

The tables show that only three of the 37 models that contained all of the variables were significant. These were the models for grade 8 reading and writing for Illinois and Minnesota, all with p-values ranging from 0.0013 to 0.0184. Furthermore, the variables designating median household income and percent of Hispanic students were not significant for any of the 37 models. For the models for Minnesota grade 8 reading and writing, the only individual variable that was significant was the estimated grade enrollment, with p-values of 0.0009 and 0.0007 respectively. The only significant variable in the model for Illinois grade 8 writing was the percent of Black students, with a p-value of 0.0064. For some states, the overall model was not significant, but had individual variables that were significant. Examples of such states are Kansas grade 4, where the significant individual variable was the dummy variable corresponding to nonresponse class 4, which indicates for this state that the nonresponse classes significantly explain the variation in the response rates. In fact, Kansas was the only state in which the nonresponse class turned out to be a significant individual variable in the model. There were two models, for grade 8 reading and writing in the state of Wisconsin, in which the percent of Black students was significant even though the overall model was not.

As mentioned before, the variable designating the percent of Black students was clearly significant in the models for Wisconsin grade 8 reading and writing, and for Illinois grade 8 writing. This variable was used in forming nonresponse adjustment classes in these states. Note that the percent of Black students in Wisconsin is 7.99 for the grade 8 reading fill sample (see Table B-2 in Appendix B), and 9.56 for the respondents. This indicates that the final sample is somewhat over-representative of schools with relatively high proportion of Black students. Similar results hold for Illinois and Wisconsin grade 8 writing (see Table B-3 in Appendix B).

Table 11-8

Results of Logistic Regression Analysis of School Nonresponse - Grade 4, 1998 Reading State Samples

			Model with All Variables			Test: Y_{ij}'s = 0	
Jurisdiction	**School Participation Rate (%)**	**Percent of Population Covered by Model**	**Degrees of Freedom**	**Significance**	**Significant Variables**	**Degrees of Freedom**	**Significance**
California	79.92	92.74	7	p=0.279	none	4	p=0.069
Iowa	83.94	80.13				4	
Illinois	84.13	100.00	12	p=0.309	none	4	p=0.839
Kansas	70.42	100.00	8	p=0.237	nonresponse cell 4, p=0.0390	4	p=0.309
Massachusetts	88.15	56.93				4	
Maryland	88.42	73.21				4	
Minnesota	85.82	55.45					
Montana	78.48	91.37				4	
New Hampshire	70.48	100.00	7	p=0.564	none	4	p=0.954
New York	83.92	82.25				4	
Washington	89.25	88.51				4	
Wisconsin	82.04	80.15				4	

Table 11-9

Results of Logistic Regression Analysis of School Nonresponse – Grade 8, 1998 Reading State Samples

Jurisdiction	School Participation Rate (%)	Percent of Population Covered by Model	Model with All Variables: Degrees of Freedom	Model with All Variables: Significance	Significant Variables	Test: Y_{ij}'s = 0: Degrees of Freedom	Test: Y_{ij}'s = 0: Significance
California	83.74	79.87					
Illinois	81.12	100.00	9	p=0.001	none	4	p=0.126
Kansas	70.60	100.00	9	p=0.748	none	4	p=0.353
Kentucky	87.32	72.63				4	
Massachusetts	89.20	77.59				4	
Maryland	85.45	81.62				4	
Minnesota	73.73	100.00	7	p=0.009	estimated grade enrollment, p=0.0009	4	p=0.003
Montana	77.81	79.74				4	
New York	77.27	100.00	8	p=0.198	none	4	p=0.282
Oregon	87.53	86.66				4	
Tennessee	89.03	60.09	8	p=0.203	none	4	p=0.083
Washington	86.13	95.22	11	p=0.701	none	4	p=0.897
Wisconsin	73.18	100.00	8	p=0.331	percent Black, p=0.0134	4	p=0.075

Table 11-10

Results of Logistic Regression Analysis of School Nonresponse – Grade 8, 1998 Writing State Samples

Jurisdiction	School Participation Rate (%)	Percent of Population Covered by Model	Model with All Variables: Degrees of Freedom	Model with All Variables: Significance	Significant Variables	Test: Y_{ij}'s = 0: Degrees of Freedom	Test: Y_{ij}'s = 0: Significance
California	83.15	85.83				4	
Illinois	80.28	100.00	9	p=0.003	Percent of Black students, p=0.0064	4	p=0.067
Kentucky	87.14	73.23				4	
Massachusetts	89.28	77.42				4	
Maryland	86.42	81.62				4	
Minnesota	73.51	100.00	7	p=0.018	Estimated grade enrollment, p=0.0007	4	p=0.010
Montana	77.60	82.51				4	
New York	77.27	100.00	8	p=0.099	none	4	p=0.588
Oregon	87.53	86.66				4	
Tennessee	89.03	60.07	8	p=0.354	none	4	p=0.140
Washington	86.59	95.16	11	p=0.506	none	4	p=0.852
Wisconsin	72.91	100.00	8	p=0.246	Percent of Black students, p=0.0068	4	p=0.044

The only models in which the estimated grade-specific enrollment is significant are those for grade 8 reading and writing in the state of Minnesota. For public schools, this variable was not used in forming nonresponse adjustment classes in these states (it was used only for Virgin Islands). This variable is not shown in Tables B-1 through B-3 in Appendix B. However, the near-zero value of the coefficient for this variable in the logistic model indicates that small schools have as much chance of participating as larger schools, after controlling for the other predictor variables.

To determine if the variables other than the nonresponse adjustment class variables added explanatory power to the model, all variables except the nonresponse adjustment class variables were tested collectively to see if the estimates of the parameters were equal to zero. This evaluates whether, taken as a group, the Y variables are significantly related to the response probability, after accounting for nonresponse class. The results are shown in the last columns of Tables 11-8 through 11-10. Only three of the 37 tests were significant. The rest of the tests were not significant, which suggests that the variables did not add to the model after accounting for the nonresponse adjustment classes, even though on occasion an individual variable was significant. These results hold for Kansas grade 4 reading, where the full model was not significant, but the dummy variable representing nonresponse class 4 was significant. This seems to indicate for Kansas, the nonresponse adjustment classes alone explain the significant variations in the probability of participation in the grade 4 assessments.

The results of the analysis indicate that on occasion there were differences between the originally sampled schools and those that participated, that were not fully removed by the process of creating nonresponse adjustments. Although these effects were not dramatic, they were sometimes statistically significant, and in these instances, this was reflected in noticeable differences in population characteristics between respondents against those who were originally sampled. However, the evidence presented here does not permit valid speculation about the likely size or even direction of the bias in achievement results in reading and writing for the few states where these sample differences are noticeable. The results and details of the logistic regression analysis are given in Westat's *Sampling Activities and Field Operations for 1998 NAEP* (Gray, et al., 2000).

11.4.3 Weighted Distributions of Students Before and After Student Absenteeism

To check the difference between the full sample and the assessed samples, Westat analysts studied weighted distributions of students before and after student absenteeism. For the public schools in each jurisdiction, subject, and grade, Westat calculated the weighted sampled percentages of students by gender (male) and race/ethnicity (White, not Hispanic; Black, not Hispanic; Hispanic), as well as SD/LEP status for the full sample of students (after student exclusion), and for the assessed sample. See tables in Appendix B. The mean student age in months is also computed on each basis. In those jurisdictions having adequate school response rates to permit reporting of combined results for public- and nonpublic-school students, these statistics were calculated for both grades and subjects for all students, public and nonpublic.

The weight used for the full sample was the adjusted student base weight, defined in Section 11.2.5. The weight for the assessed students was the final student weight, defined in Section 11.3.5. The difference between the estimates of the population subgroups is an estimate of the bias in estimating the size of the subgroup, resulting from student absenteeism.

Care must be taken in interpreting these results. First, note that there is generally little difference in the proportions estimated from the full sample and those estimated from the assessed students. While this is encouraging, it does not eliminate the possibility that bias exists within the state as a whole, within the results for gender and race/ethnicity subgroups, or within other subgroups. Second, when differences do exist, they

cannot be used to indicate the likely magnitude or direction of the bias with any reliability. For example, in Illinois the percentages of White and Black students in the full sample are respectively 56.87 and 22.24 percent. For assessed students, these percentages are 61.97 for White students and 18.61 for Black students. This indicates that White students are overrepresented and Black students are underrepresented in the sample of assessed students. While these differences raise the possibility that some bias exists, it is not appropriate to speculate on the magnitude of this bias by considering the assessment results for White or Black students in comparison to other students in the state. The reason is that the overrepresented White students or underrepresented Black students may not be typical of students that were included in the sample. Similarly, White students who are disproportionately underrepresented or Black students who are disproportionately overrepresented may not be typical either, because not all students within the same race/ethnicity group receive the same student nonresponse adjustment.

One other feature to note is that, for assessed students, information about the student's gender and race/ethnicity is provided by the student, whereas for absent students, it is provided by the school. Evidence from past NAEP assessments (see, for example, Rust & Johnson, 1992) indicates that there can be substantial discrepancies between those two sources, particularly for grade 4 Hispanic students.

11.5 VARIATION IN WEIGHTS

After computing the full-sample weights, an analysis was conducted on the distribution of the final student weights for each grade-subject combination in each jurisdiction. The analysis was intended to (1) check that the various weight components had been derived properly in each jurisdiction, and (2) examine the impact of variability in the sample weights on the precision of the sample estimates, both for the jurisdiction as a whole and for major subgroups within the jurisdiction.

The analysis was conducted by looking at the distribution of the final student weights for the assessed students in each jurisdiction, grade, and subject separately by public and nonpublic schools. Two key aspects of the distribution were considered in each case: the coefficient of variation (equivalently, the relative variance) of the weight distribution, and the presence of outliers—cases whose weights were several standard deviations away from the median weight.

It was important to examine the coefficient of variation of the weights, because a large coefficient of variation reduces the effective size of the sample. Assuming that the variables of interest for individual students are uncorrelated with the weights of the students, the sampling variance of an estimated average or aggregate is approximately $(1+V_W^2)$ times as great as the corresponding sampling variance based on a self-weighting sample of the same size, where V_W is the coefficient of variation of the weights. Outliers, or cases with extreme weights, were examined because the presence of such outliers was an indication of the possibility that an error was made in the weighting procedure, and because it was likely that a few extreme cases would contribute substantially to the size of the coefficient of variation.

In most jurisdictions, the coefficients of variation were 35 percent or less, both for the whole sample and for all subgroups. This means that the variation in sampling weights had little impact on the precision of sample estimates.

A few relatively large student weights were observed in some jurisdictions for reading at both grades 4 and 8. An evaluation was made of the impact of trimming these largest weights back to a level consistent with the largest remaining weights found in the state and grade. Such a procedure produced an appreciable reduction in the size of the coefficient of variation for these weights, and hence this trimming was implemented. Westat

judged that this procedure had minimal potential to introduce bias, while the reduction in the coefficient of variation of the weights gave rise to an appreciable decrease in sampling error for all jurisdictions, grades, and subjects.

11.6 CALCULATION OF REPLICATE WEIGHTS

A replication method known as jackknife was used to estimate the variance of statistics derived from the full sample. The process of replication involves repeatedly selecting portions of the sample (replicates) and calculating the desired statistic (replicate estimates). The variability among the calculated replicate estimates is then used to obtain the variance of the full-sample estimate.

In each jurisdiction, replicates were formed in two steps. First, each school was assigned to one of a maximum of 62 replicate groups, each group containing at least one school. In the next step, a random subset of schools (or, in some cases, students within schools) in each replicate group was excluded. The remaining subset and all schools in the other replicate groups then constituted one of the 62 replicates. The process of forming these replicate groups, core to the process of variance estimation, is described below.

11.6.1 Defining Replicate Groups and Forming Replicates for Variance Estimation

Replicate groups were formed separately for public and nonpublic schools. Once replicate groups were formed for all schools, students were then assigned to their respective school replicate groups. The formation of replicate groups was done separately for SD/LEP and non-SD/LEP students. For SD/LEP students, there was an additional set of replicate group assignments for reading at each grade for states with certainty schools. Different replicate group assignments were needed for SD/LEP students in reading because only SD/LEP students that were not offered accommodations will be used in reporting for reading. This essentially meant that certainty schools were treated as noncertainty schools for replication of SD/LEP students in reading.

In general, public schools (except schools in Virgin Islands and DoDEA/DDESS grade 8) were assigned to replicates as follows: Noncertainty schools were first paired and then each pair was assigned to its own replicate group. Large certainty schools were assigned to two replicate groups each, and small certainty schools were assigned to one replicate group each.

For nonpublic schools, the assignment of replicate groups was as follows: If the sample of noncertainty schools was small, each noncertainty school was randomly assigned to its own replicate group. If the sample of noncertainty schools was large enough, this procedure was implemented separately for Catholic and non-Catholic noncertainty schools. Then, large certainty schools were assigned to two replicate groups each, and small certainty schools were assigned to one replicate group each.

Replicate group assignments for schools in Virgin Islands and DoDEA/DDESS grade 8 were handled differently because of small sample sizes. Nonpublic schools in Virgin Islands were assigned to replicate groups using the procedure described in the preceding paragraph for nonpublic schools. For public schools in Virgin Islands and DoDEA/DDESS grade 8, schools were assigned to a number of replicate groups proportional to the estimated grade-specific enrollment.

The details about the replicate group assignments for all schools are given below.

11.6.1.1 Replicate Group Assignments for Non-SD/LEP Students

All Public Schools, Except Schools in Virgin Islands and DoDEA/DDESS Grade 8. Noncertainty schools were sorted by jurisdiction according to sample type. Then within sample type, the schools were sorted by new school status and the order in which they were selected from the sampling frame. The schools were then grouped in pairs. Where there was an odd number of schools, the last replicate group contained three schools instead of two. If a jurisdiction had more than 62 pairs, the pair numbering would have gone up to 62 and then from 62 backwards as needed; however, this did not happen in 1998.

Each of the certainty public schools was assigned to one replicate group or to more replicate groups if its size was large. If a school was selected three or more times in the sampling process, then it was assigned to two replicate groups. Here, schools were sorted by the estimated grade enrollment prior to group assignments. Again, depending on the jurisdiction, a maximum of 62 certainty groups was formed. The group numbering resumed from the last group number used for the noncertainty schools if the total number of public-school groups was less than 62. Otherwise, the numbering started from 62 down to the number needed for the last certainty public school. In jurisdictions where all schools were certainty schools and the total number of public schools (that is, certainty schools) exceeds 62, the numbering of the groups started at 62 and went downward to 1, and then from 1 up to the number needed for the last certainty school. For instance, in the District of Columbia grade 4 reading, which had only 114 certainty schools (no noncertainty schools), group numbers started at 62 and continued down to 1 and then from 1 up to 52. In the District of Columbia grade 8 reading, which had only 37 certainty schools, the group numbers went from 1 to 55. Eighteen of the 37 certainty schools in the District of Columbia were selected three or more times and thus were assigned to two replicate groups. A replicate was formed by randomly deleting one half of the students in a certainty school from the sample. For certainty schools that were assigned to two replicate groups, the students were split equally between four "halves," two halves in each of the two replicate groups. This process was repeated for each certainty school.

The purpose of this scheme was to assign as many replicates to a jurisdiction's public schools as permitted by the design, to a maximum of 62. When more than 62 replicates were assigned, the procedure ensured that no subset of the replicate groups (pairs of noncertainty schools, individual certainty schools, or groups of these) was substantially larger than the other replicate groups. The aim was to maximize the degrees of freedom available for estimating variances for public-school data.

A single replicate estimate was formed by dropping one member assigned to a particular replicate group. This process was repeated successively across replicate groups, giving up to 62 replicate estimates.

Nonpublic Schools. Replicate groups for noncertainty nonpublic schools were formed in one of the two methods described below. It depends on the number of nonpublic noncertainty schools, such as the number of available noncertainty Catholic or non-Catholic schools. If any of the following conditions was true for a given jurisdiction, then the subsequent steps were taken to form replicate groups. Here, the numbering started at 62 down to the last needed number.

Conditions for Method 1:

- fewer than 11 nonpublic noncertainty schools; or
- fewer than 2 Catholic noncertainty schools; or
- fewer than 2 non-Catholic noncertainty schools.

Steps for Method 1:

- all schools were grouped into a single replicate group;
- schools were randomly sorted; and
- starting with the second school, replicates were formed by consecutively leaving out one of the remaining *n* - 1 schools; each replicate included the first school.

When a given jurisdiction did not match conditions of the first method (i.e., when all of the following conditions were true), then the preceding steps were repeated separately for two groups, one consisting of Catholic schools and one consisting of non-Catholic schools.

Conditions for Method 2:

- more than 10 nonpublic noncertainty schools; and
- more than 1 Catholic noncertainty school; and
- more than 1 non-Catholic noncertainty school.

For jurisdictions with certainty nonpublic schools (Hawaii and Virgin Islands for reading at grade 4; Rhode Island, Virgin Islands, and Wyoming for both reading and writing at grade 8) each school was assigned to one or more groups. If a school was selected three or more times in the sampling, it was assigned to two groups. Prior to this assignment, schools were sorted in descending order of the estimated grade enrollment. The group numbering started at the last number where the noncertainty nonpublic schools ended. A replicate was formed by randomly deleting one half of the students in a certain school from the sample. For the certainty schools that were assigned to two replicate groups, the students were split equally between four "halves," two halves in each of two replicate groups. This was repeated for each certainty school.

Again, the aim was to maximize the number of degrees of freedom for estimating sampling errors for nonpublic schools (and indeed for public and nonpublic schools combined) within the constraint of forming 62 replicate groups. Where a jurisdiction had a significant contribution from both Catholic and non-Catholic schools, Westat ensured that the sampling error estimates reflected the stratification on this characteristic.

Virgin Islands. For Virgin Islands, where all schools were selected with certainty, nonpublic schools were assigned in the usual way, and public schools were assigned to a number of replicate groups proportional to their estimated grade enrollment.

DoDEA/DDESS Grade 8. Schools in the DoDEA/DoDDS grade 8 sample were assigned to a number of replicate groups proportional to their estimated grade enrollment. Schools in all other Department of Defense Domestic Dependent Elementary and Secondary Schools (DoDEA/DDESS) and DoDEA/DoDDS samples were assigned to replicate groups following the general rules described above for all public schools. In grade 8 writing, the one noncertainty school was treated like a certainty school.

11.6.1.2 Replicate Group Assignments for SD/LEP Students in Reading

For reading certainty schools with non-SD/LEP students were reassigned to replicate groups. The replicate group assignments for all other schools remained the same. As mentioned before, there were no certainty schools for SD/LEP replication for reading (certainty schools were treated as noncertainty schools). The reassignment of replicate groups for certainty schools was implemented as follows.

All Public Schools, Except those in Virgin Islands and DoDEA/DDESS Grade 8. The assignment of schools to replicate groups was done separately for various subgroups of the reading SD/LEP sample. For public noncertainty schools, the schools were first sorted by jurisdiction according to sample type. Within each sample type, the schools were sorted by their new school status and sample selection order. In those jurisdictions where the number of replicate groups for public schools did not exceed 62, the schools in the sorted list were assigned group numbers, two to a group, beginning where the previous assignments for the public non-certainty schools with non-SD/LEP students stopped. If the number of schools was odd, then the last three schools were assigned to the same replicate group. If the number of public noncertainty schools exceeded 62, then the group numbering started at 62 and proceeded backwards, assigning pairs of schools to the same replicate group. If the number of public noncertainty schools to be assigned was odd, the last three schools were assigned to the same replication group. For Arkansas, Illinois, and Mississippi grade 4; and Florida, North Carolina, and Tennessee grade 8, there was only one public noncertainty school with SD/LEP students assessed in reading. This school was assigned to the last replicate group used for the public noncertainty schools with non-SD/LEP students. If there was an odd number of such schools, then the triple was broken up into two doubles and the school in question was assigned to the last double.

Nonpublic Schools. Nonpublic schools were assigned to replicate groups as follows. For noncertainty schools, the replicate group assignments were the same for Catholic and non-Catholic schools, and used one of the two methods described below.

Method 1. If the conditions for Method 1 for non-SD/LEP replication were met, then the first school in the sorted list was not assigned to any group. The second and subsequent schools were assigned to one replicate group each, beginning where the numbering for nonpublic noncertainty schools in the non-SD/LEP replication stopped. The numbering then proceeded backwards.

Method 2. If the conditions for Method 2 for non-SD/LEP replication were met, then the procedure for Method 1 was implemented for Catholic and non-Catholic schools separately. Catholic schools were assigned first, starting from where the numbering for nonpublic noncertainty non-Catholic schools in the non-SD/LEP replication stopped. The numbering for the non-Catholic schools started from where that for the Catholic schools stopped.

Virgin Islands. In Virgin Islands, nonpublic schools were assigned to replicate groups in the usual way, and the public schools were assigned in the same way as nonpublic schools.

DoDEA/DDESS Grade 8. In the DoDEA/DDESS grade 8, schools were assigned to replicate groups in exactly the same way as for nonpublic schools.

11.6.2 School-Level Replicate Weights

As mentioned above, each replicate sample had to be reweighted to compensate for the dropped unit(s) defining the replicate. This reweighting was done in two stages. At the first stage, the i^{th} school included in a particular replicate r was assigned a replicate-specific school base weight defined as:

$$W_{ri}^{sch} = K_r \times W_i^{sch},$$

where W_i^{sch} is the full-sample base weight for school i, and, for public schools,

$$K_r = \begin{cases} 1.5 & \text{if school } i \text{ was contained in a "pair" consisting of 3 units} \\ & \text{from which the complimentary member was dropped to form replicate } r, \\ 2 & \text{if school } i \text{ was contained in a pair consisting of 2 units} \\ & \text{from which the complimentary member was dropped to form replicate } r, \\ 0 & \text{if school } i \text{ was dropped to form replicate } r, \text{ and} \\ 1 & \text{if school } i \text{ was not assigned to replicate } r, \text{ or if school } i \text{ was a certainty.} \end{cases}$$

For nonpublic schools, Method 1:

$$K_r = \begin{cases} \dfrac{n}{n-1} & \text{if school } i \text{ was not dropped in forming replicate } r, \text{ and} \\ 0 & \text{if school } i \text{ was dropped to form replicate } r. \end{cases}$$

For nonpublic schools, Method 2 (with n_1 Catholic schools and n_2 non-Catholic schools):

$$K_r = \begin{cases} \dfrac{n_1}{n_1 - 1} & \text{if school } i \text{ was Catholic not dropped from replicate } r, \\ & \text{and replicate } r \text{ was formed by dropping a Catholic school;} \\ 1 & \text{if school } i \text{ was Catholic and replicate } r \text{ was formed by dropping a non-Catholic school;} \\ \dfrac{n_2}{n_2 - 1} & \text{if school } i \text{ was non-Catholic not dropped from replicate } r, \\ & \text{and replicate } r \text{ was formed by dropping a non-Catholic school;} \\ 1 & \text{if school } i \text{ was dropped to form replicate } r. \end{cases}$$

Using the replicate-specific school base weights, W_{ri}^{sch}, the school-level nonresponse weighting adjustments were recalculated for each replicate *r*. That is, the school-level nonresponse adjustment factor for schools in replicate *r* and adjustment class *k* was computed as:

$$F_{rk} = \frac{\sum_{i \in C_k} (W_{rki}^{sch} \times E_{ki})}{\sum_{i \in C_k} (W_{rki}^{sch} \times E_{ki} \times \delta_{rki})}$$

where

C_k = the subset of school records in adjustment class *k*,

W_{rki}^{sch} = the replicate-*r* base weight of the i^{th} school in class *k*, and

E_{ki} = the grade enrollment for the i^{th} school in class k.

In the above formulation, the indicator variable δ_{rki} had a nonzero value only when the i^{th} school in replicate r and adjustment class k participated in the assessment. The replicate-specific nonresponse-adjusted school weight for the i^{th} school in replicate r in class k was then computed as:

$$W_{rki}^{adj} = F_{rk} \times W_{rki}^{sch} \times \delta_{rki} \ .$$

11.6.3 Student-Level Replicate Weights

The replicate-specific adjusted student base weights were calculated by multiplying the replicate-specific adjusted school weights as described above by the corresponding within-school student weights. That is, the adjusted student base weight for the j^{th} student in adjustment class k in replicate r was initially computed as:

$$W_{rkij} = W_{rki}^{adj} \times W_{ij}^{within}$$

where

W_{rki}^{adj} = the nonresponse-adjusted school weight for school i in school adjustment class k and replicate r, and

W_{ij}^{within} = the within-school weight for the j^{th} student in school i.

The final replicate-specific student weights were then obtained by applying the student nonresponse adjustment procedures to each set of replicate student weights. Let F_{rk} denote the student-level nonresponse adjustment factor for replicate r and adjustment class k. The final replicate r student weight for student j in school i in adjustment class k was calculated as:

$$W_{rkij}^{final} = F_{rk} \times W_{rki}^{adj} \times W_{ij}^{within}$$

Finally, estimates of the variance of sample-based estimates were calculated as:

$$Var_{JK}(\hat{x}) = \sum_{r=1}^{62} (\hat{x}_r - \hat{x})^2$$

where

$$\hat{x}_r = \sum_{i,j} W_{rkij}^{final} \times x_{rkij}$$

denotes an estimated total based on replicate *r* (one of 62 replicates), and $\hat{x}$ denote the corresponding estimate based on the full sample. The standard error of an estimate $\hat{x}$ is estimated by taking the square root of the estimated variance, $\mathrm{Var}_{JK}(\hat{x})$.

11.7 RAKING OF WEIGHTS

Raking (also known as *iterative proportional fitting*) is done in place of poststratification. Unlike poststratification, it is performed iteratively to two or more different distributions of a population total (i.e., gender and age). It is typically used in situations in which the interior cells of a cross-tabulation are either unknown, or some sample sizes in the cells are too small for efficient estimation. In raking, the marginal population totals, $N_{i.}$ and $N_{.j}$ are known (i.e., age and gender population counts); however, the interior cells of the cross-tabulation N_{ij} (the age by gender cells) are estimated from the sample by $\hat{N}_{ij}$, where these are the sum of weights in the cells.

The raking algorithm proceeds by proportionally scaling the $\hat{N}_{ij}$, such that the following relations are satisfied:

$$\sum_j \hat{N}_{ij} = N_{i.}$$

and

$$\sum_i \hat{N}_{ij} = N_{.j}.$$

The 1998 state NAEP assessment program used two different sets of administration rules indicated by sample type 2 and sample type 3 (see Chapter 4). To enable ETS to analyze the reading assessment omitting the SD/LEP students with sample type 3, the SD/LEP student weights were raked separately for the two subsets as defined by sample type. Note that only the weights of SD/LEP students in public schools were raked. Agreement was forced with totals estimated using both of the subsets combined for each of the sample types. The purpose of this was to enhance the reliability (i.e., reduce the sampling error) of estimates produced by using information about student characteristics from the whole sample to enhance the estimates. Because of small sample sizes, the weights of nonpublic SD/LEP students were not raked but were assigned a crude raking factor of 2. Non-SD/LEP students were assigned dummy raking factors of 1.

11.7.1 Raking Dimensions for Full Sample Student Weights

Public Schools. Five variables were used for the raking dimensions. These variables included two levels of SD (SD/non-SD), two levels of LEP (LEP/non-LEP), two levels of gender, five levels of race (White and other; Black; Hispanic; Asian or Pacific Islander; and American Indian or Alaskan Native), and two levels of age. The age variable was defined as follows: for grade 4, those born in August 1987 or earlier and those born in September 1987 or later; and for grade 8, those born in August 1983 or earlier and those born in September 1983 or later. Collapsing of levels was done so that no level of a single dimension contained fewer than 30 students for a state and grade.

Control totals were obtained by summing the trimmed nonresponse-adjusted student weights for each level of the collapsed raking dimension. The final collapsed levels that were used for the raking dimensions, for each jurisdiction and grade, can be found in Tables B-13 and B-14 in Appendix B. An "X" indicates that the variable was not collapsed for raking. A dash indicates that all levels were combined, and thus, the variable was not used as a raking dimension. An asterisk for the race variable indicates that all other levels of the dimension were combined into one level. For example in fourth grade for Florida, there are three levels of race: White, Hispanic, and all others combined.

Nonpublic Schools. Because of the small numbers of nonpublic-school students, no raking was carried out. A factor of 2 was applied to the weights for the SD/LEP students, since only half the SD/LEP sample was used for analysis.

11.7.2 Raking Student Replicate Weights

The replicate weights for the public SD/LEP students were raked similarly. Control totals for each replicate were calculated based on the totals for the replicate weights. The levels of the raking dimensions that were used for the replicates were the same collapsed levels as used for the full sample student weights. For the nonpublic schools, again a factor of 2 was applied to the replicate weights of the SD/LEP students.

11.8 APPROXIMATING THE SAMPLING VARIANCE USING DESIGN EFFECTS

As in Chapter 10's discussion of variance estimation (see Section 10.5), *design effects* (Kish & Frankel, 1974) of mean proficiencies across the state samples were calculated for demographic subgroups for reading grades 4 and 8, and writing grade 8, respectively. The design effect for a statistic is the ratio of the actual variance of the statistic (taking the sample design into account) over the conventional variance estimate based on a simple random sample with the same number of elements. The design effect is the inflation factor to be applied to the conventional variance estimate in order to adjust error estimates based on simple random sampling assumptions, thus accounting approximately for the effect of the sample design. Design effects provide an approximate approach to compute variance from NAEP data for secondary analysis. Moreover, they provide a measure to analyze the efficiency of a study design.

Since most of the analyses conducted by NAEP are based on the results of scaling models that summarize performance of students across a learning area, the design effects are based on these scale scores. A key statistic of interest is the estimated mean scale score of a subgroup of the population. Table 11-11 gives the average design effects for state-level mean scale score, averaged across all jurisdictions by grade for the 1998 state reading and writing assessments.

The table shows that the design effects are predominantly larger than 1, indicating that standard variance estimation formulas will be generally too small, usually markedly so. Although the design effects appear somewhat different for certain subgroups of the population, they are similar enough (at least within a subject and grade) to select an overall composite value that is adequate for most purposes. In choosing a composite design effect, some consideration must be made about the relative consequences of overestimating the variance as opposed to underestimating the variance. (For details, see descriptions in Section 10.5.2.) Table 11-12 gives the composite values of mean, median, and upper quartile of the distribution of design effects for mean state scale scores by grade for the 1998 state reading and writing assessments.

Table 11-11

Average Design Effects by Demographic Subgroup for 1998 Mean State Reading and Writing Scale Scores Averaged Across State Samples[*]

Subgroup	Grade 4 Reading	Grade 8 Reading	Grade 8 Writing
Total	3.81	3.25	3.21
Male	2.54	2.45	2.29
Female	2.49	2.13	2.28
White	2.74	2.44	2.61
Black	1.87	2.17	2.03
Hispanic	2.06	1.70	1.44
Asian/Pacific Islander	1.48	1.42	1.21
Other race/ethnicity	1.47	1.81	1.34
Urban	5.00	4.44	4.37
Suburban	4.07	3.63	3.02
Rural	3.37	3.12	2.75
PARED < HS	1.28	1.52	1.13
PARED = HS	1.39	1.76	1.28
PARED > HS	1.59	1.49	1.59
PARED = College	2.91	2.18	2.40
PARED = Unknown	1.68	1.43	1.11
Public school	3.84	3.13	2.95

[*] Design effects are based on the conventional and jackknife variances of subgroup means of the first plausible values of scale score.

Table 11-12

Mean, Median, and Upper Quartile of the 1998 Across-State Average Design Effects for Mean State Scale Score (Distribution Across Demographic Subgroups) [*]

Subgroup	Grade 4 Reading	Grade 8 Reading	Grade 8 Writing
Upper Quartile	3.37	3.12	2.75
Mean	2.56	2.36	2.18
Median	2.49	2.17	2.28

[*] Design effects are based on the conventional and jackknife variances of subgroup means of the first plausible values of scale score.

Chapter 12

SCALING PROCEDURES[1]

Nancy L. Allen, James E. Carlson, Eugene G. Johnson, and Robert J. Mislevy
Educational Testing Service

12.1 INTRODUCTION

The primary method by which results from the 1998 National Assessment of Educational Progress (NAEP) were disseminated is scale score reporting. The National Assessment Governing Board (NAGB) provides achievement levels that are used to give judgmental meaning to the scale. With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or series of scales even when different students have been administered different items. This chapter presents an overview of the scaling methodologies employed in the analyses of the data from NAEP surveys in general. Details of the scaling procedures specific to the subject areas of reading, writing, and civics are presented in Chapters 14 through 24.

12.2 BACKGROUND

The basic information from an assessment consists of the responses of students to the items presented in the assessment. For NAEP, these items are constructed to measure performance on sets of objectives developed by nationally representative panels of learning-area specialists, educators, and concerned citizens. Satisfying the objectives of the assessment and ensuring that the tasks selected to measure each goal cover a range of difficulty levels typically require many items. Depending on the subject areas, a mixture of multiple-choice, short constructed-response, and extended constructed-response items were used. To reduce student burden, each assessed student was presented only a fraction of the full pool of items through multiple matrix sampling procedures.

The most direct manner of presenting the assessment results is to report separate statistics for each item. However, because of the vast amount of information, having separate results for each of the items in the assessment pool hinders the comparison of the general performance of subgroups of the population. Item-by-item reporting masks similarities in trends and subgroup comparisons that are common across items.

An obvious summary of performance across a collection of items is the average of the separate item scores. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items that can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with mean item scores. First, the interpretation of these results depends on the selection of the items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, the average

[1] Nancy L. Allen and James E. Carlson shared responsibility for the psychometric and statistical analysis of the 1998 national and state NAEP data with John R. Donoghue. Eugene G. Johnson contributed to the design of NAEP and to discussions of sampling issues. Previously he was responsible for the psychometric and statistical analysis of NAEP data. Robert J. Mislevy is a technical consultant contributing in the area of item response theory.

score is related to the particular items comprising the average, so that direct comparisons in performance between subpopulations require that those subpopulations have been administered the same set of items. Third, because this approach limits comparisons to average scores on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, direct estimates of parameters or quantities such as the proportion of students who would achieve a certain score across the items in the pool are not possible when every student is administered only a fraction of the item pool. While the average score across all items in the pool can be readily obtained (as the average of the individual item scores), statistics that provide distributional information, such as quantiles of the distribution of scores across the full set of items, cannot be readily obtained without additional assumptions.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. These variables include a respondent-specific variable, called *scale score*, which quantifies a respondent's tendency to answer items correctly (or, for multipoint items, to achieve a certain item score) and item-specific variables that indicate characteristics of the item such as its difficulty, effectiveness in distinguishing between individuals with different levels of scale score, and the chances of a very low scale score respondent correctly answering a multiple-choice item. (These variables are discussed in more detail in the next section.) When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale, even though none of the respondents takes all of the items within the pool. Using the common scale, it becomes possible to discuss distributions of scale score in a population or subpopulation and to estimate the relationships between scale score and background variables.

It is important to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. Every item in a NAEP survey is of interest and can provide useful information about what United States students know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose.

The scaling for the national main reading, mathematics, science, U.S. history, geography, and music assessments is carried out separately within purposes of reading, mathematics content strands, fields of science, themes, or content areas as specified in the framework. Originally, this scaling within subareas was done because it was anticipated that different patterns of performance or different trends over time might exist for these essential subdivisions of the subject areas. By creating a separate scale for each of these content areas, potential differences in subpopulation performance between the content areas are preserved.

The creation of a series of separate scales to describe performance within a subject area does not preclude the reporting of a single index of overall performance in the subject area—that is, an overall subject–area composite. A composite is computed as the weighted average of the content–area scales, where the weights correspond to the relative importance given to each content area as defined by the framework. The composite provides a global measure of performance within the subject area, while the constituent content area scales allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

For all other national main assessment subjects the framework documents specify a single (unidimensional) scale. The long-term trend scales for reading, writing, mathematics, and science are also scaled as if they were unidimensional.

12.3 SCALING METHODOLOGY

This section reviews the scaling models employed in the analyses of NAEP data and the multiple imputation or "plausible values" methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (1991) for an introduction to plausible values methods and a comparison with standard psychometric analyses to Beaton and Johnson (1992), Donoghue (1993), and Mislevy, Johnson and Muraki (1992), and for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach. It should be noted that the imputation procedure used by NAEP is a mechanism for providing plausible values for the unobserved proficiencies and not for filling in blank responses to background or cognitive variables.

While the NAEP procedures were developed explicitly to handle the characteristics of NAEP data, they build on other research, and are paralleled by other researchers. See, for example, Andersen (1980); Dempster, Laird, and Rubin (1977); Engelen (1987); Hoijtink (1991); Laird (1978); Lindsey, Clogg, and Grego (1991); Little and Rubin (1983, 1987); Rubin (1987, 1991); Tanner and Wong (1987); and Zwinderman (1991).

12.3.1 The Scaling Models

Three distinct scaling models, depending on item type and scoring procedure, are used in the analysis of NAEP data. Each of the models is based on item response theory (IRT; e.g., Lord, 1980). Each is a "latent variable" model, defined separately for each of the scales, which expresses respondents' tendencies to achieve certain scores (such as correct/incorrect) on the items contributing to a scale as a function of a parameter that is not directly observed, called score *(θ)* on the scale.

A three-parameter logistic (3PL) model is used for the multiple-choice items (which are scored correct or incorrect). The fundamental equation of the 3PL model defines the probability that a person whose score on scale *k* is characterized by the *unobservable* variable θ_k will respond correctly to item *j* as:

$$P\left(x_j = 1 \middle| \theta_k, a_j, b_j, c_j\right) = c_j + \frac{\left(1 - c_j\right)}{1 + \exp\left[-1.7 a_j \left(\theta_k - b_j\right)\right]} \equiv P_{j1}\left(\theta_k\right), \tag{12.1}$$

where

x_j is the response to item *j*, 1 if correct and 0 if not;

a_j where $a_j > 0$, is the slope parameter of item *j*, characterizing its sensitivity to scale score;

b_j is the threshold parameter of item *j*, characterizing its difficulty; and

c_j where $0 \le c_j < 1$, is the lower asymptote parameter of item *j*, reflecting the chances of students of very low scale score selecting the correct option.

Further define the probability of an incorrect response to the item as

$$P_{j0} \equiv P\left(x_j = 0 \middle| \theta_k, a_j, b_j, c_j\right) = 1 - P_{j1}\left(\theta_k\right). \tag{12.2}$$

A two-parameter logistic (2PL) model is used for the short constructed-response items that were scored correct or incorrect. The form of the 2PL model is the same as Equations (12.1) and (12.2), with the c_j parameter fixed at zero.

In addition to the multiple-choice and other two-category items, a number of extended constructed-response items are presented in NAEP assessments. The long-term trend and national main writing assessments include only extended constructed-response items, but most other national main and state assessments include some extended constructed-response items. Each of these items is scored on a multipoint scale with potential scores ranging from 0 to 3, from 0 to 4, or from 0 to 5. For some subjects, short constructed-response items are scored on a three-point scale (0–2) as well as on a two-category scale. Items that are scored on a multipoint scale are referred to as polytomous items, in contrast with the multiple-choice and short constructed-response items, which are scored correct or incorrect and referred to as dichotomous items.

The polytomous items are scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model is the probability that a person with score θ_k on scale k will have, for the j^{th} item, a response x_j that is scored in the i^{th} of m_j ordered score categories:

$$P\left(x_j = i \middle| \theta_k, a_j, b_j, d_{j,1}, \ldots, d_{j,m_j-1}\right) = \frac{\exp\left(\sum_{v=0}^{i} 1.7a_j\left(\theta_k - b_j + d_{j,v}\right)\right)}{\sum_{g=0}^{m_j-1} \exp\left(\sum_{v=0}^{g} 1.7a_j\left(\theta_k - b_j + d_{j,v}\right)\right)} \equiv P_{ji}\left(\theta_k\right) \qquad (12.3)$$

where

m_j is the number of categories in the response to item j;

x_j is the response to item j, with possibilities 0, 1, ... , $m_j - 1$;

a_j is the slope parameter;

b_j is the item location parameter characterizing overall difficulty; and

$d_{j,i}$ is the category i threshold parameter (see below).

Indeterminacies in the parameters of the above model are resolved by setting $d_{j,0} = 0$ and setting $\sum_{i=1}^{m_j-1} d_{j,i} = 0$. Muraki (1992) points out that b_j - $d_{j,i}$ is the point on the θ_k scale at which the plots of $P_{j,i-1}\left(\theta_k\right)$ and $P_{ji}\left(\theta_k\right)$ intersect and so characterizes the point on the θ_k scale at which the response to item j has equal probability of falling in response category i-1 and falling in response category i.

When $m_j = 2$, so that there are two score categories (0,1), it can be shown that $P_{ji}(\theta_k)$ of Equation (12.3) for $i = 0,1$ corresponds respectively to $P_{j0}\left(\theta_k\right)$ and $P_{j1}\left(\theta_k\right)$ of the 2PL model [(Equations (12.1) and (12.2) with $c_j = 0$)].

Close examination of the 3PL and generalized partial credit models indicate that both models have a linear indeterminacy of the theta scale. In other words, if the item parameters are estimated in a

different metric, the value of θ_k could be transformed to make Equations (12.1) and (12.3) true. For the purposes of reporting item parameter estimates and other intermediary estimates, the linear indeterminacies apparent in Equations (12.1) and (12.3) may be resolved by an arbitrary choice of the origin and unit size in a given scale. In most cases, a provisional scale standardizing the theta distribution to have mean 0 and standard deviation 1 is employed. Final results for each content area are linearly transformed from the θ scale to a 0-to-500 or a 0-to-300 scale, as described in the subject area chapters in this report.

A basic assumption of item response theory is the conditional independence of the responses by an individual to a set of items, given the individual's scale score. That is, conditional on the individual's θ_k, the joint probability of a particular response pattern $\underline{x} = (x_1,...,x_n)$ across a set of *n* items is simply the product of terms based on Equations (12.1), (12.2), and (12.3):

$$P\left(\underline{x}\middle|\theta_k, item\ \ parameters\right) = \prod_{j=1}^{n}\prod_{i=0}^{m_j-1} P_{ji}\left(\theta_k\right)^{u_{ji}} \tag{12.4}$$

where $P_{ji}(\theta_k)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_j is equal to 2 for the dichotomously scored items, and u_{ji} is an indicator variable defined by

$$u_{ji} = \begin{cases} 1 \text{ response } x_j \text{ is in category } i \\ 0 \text{ otherwise} \end{cases}$$

It is also typically assumed that response probabilities are conditionally independent of background variables $(\underline{y})$, given θ_k, or

$$P\left(\underline{x}\middle|\theta_k, item\ \ parameters, \underline{y}\right) = p\left(\underline{x}\middle|\theta_k, item\ \ parameters\right). \tag{12.5}$$

After $\underline{x}$ is observed, Equation (12.4) can be viewed as a likelihood function, and provides a basis for inference about θ_k or about item parameters. Estimates of item parameters were obtained by the NAEP BILOG/PARSCALE program, which combines Mislevy and Bock's (1982) BILOG and Muraki and Bock's (1991) PARSCALE computer programs[2], and which concurrently estimates parameters for all items (dichotomous and polytomous). Donoghue (1993) reports on the effect of having both dichotomous and polytomous items within a scale. The NAEP BILOG/PARSCALE program has also been adapted to make use of student sampling weights. The item parameters are then treated as known in subsequent calculations. In NAEP analyses, for subject areas with multiple scales (i.e., national main reading, mathematics, science, U.S. history, geography, and music), the parameters of the items constituting each of the separate scales are estimated independently of the parameters of the other scales. Once items are calibrated in this manner, a likelihood function for the scale score θ_k is induced by a vector of responses to any subset of calibrated items, thus allowing θ_k-based inferences from matrix samples. The likelihood function for the scale score θ_k is called the *posterior distribution of the thetas for each student*.

In almost all NAEP IRT analyses, missing responses at the end of each block of items a student was administered are considered "not reached," and are treated as if they had not been presented to the respondent. Missing responses to dichotomous items before the last observed response in a block are considered intentional omissions, and are treated as fractionally correct at the value of the reciprocal of

[2] See Muraki and Bock (1999) for the current version of PARSCALE.

the number of response alternatives, if the item was a multiple-choice item. These conventions are discussed by Mislevy and Wu (1988). With regard to the handling of not-reached items, Mislevy and Wu found that ignoring not-reached items introduces slight biases into item parameter estimation when not-reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited-information maximum likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly.

Missing responses to polytomous items before the last observed response in a block are also considered intentional omissions and scored so that the response is in the lowest category. Occasionally, extended constructed-response items are the last item in a block of items. Because considerably more effort is required of the student to answer these items, nonresponse to an extended constructed-response item at the end of a block is considered an intentional omission (and scored as the lowest category) unless the student also did not respond to the item immediately preceding that item. In that case, the extended constructed-response item is considered not reached and treated as if it had not been presented to the student. In the case of the main and state writing assessment, there is a single extended constructed-response item in each separately-timed block. In the writing assessment when a student does not respond to the item or when the student provides an off-task response, the response is also treated as if the item had not been administered.

Scaling areas in NAEP are determined a priori by grouping items into content areas for which overall performance is deemed to be of interest, as defined by the frameworks developed by the National Assessment Governing Board (NAGB). A scale score θ_k is defined a priori by the collection of items representing that scale. What is important, therefore, is that the models capture salient information in the response data to effectively summarize the overall performance on the content area of the populations and subpopulations being assessed in the content areas.

The local independence assumption embodied in Equation (12.4) implies that item response probabilities depend only on θ and the specified item parameters, and not on the position of the item in the booklet, the content of items around an item of interest, or the test-administration and timing conditions. However, these effects are certainly present in any application. The practical question is whether inferences concerning aggregate performance in the scaling area that are based on the IRT probabilities obtained via Equation (12.4) are robust with respect to the ideal assumptions underlying the IRT model. Our experience with the 1986 NAEP reading anomaly (Beaton & Zwick, 1990) has shown that for measuring small changes over time, changes in item context and speededness conditions can lead to unacceptably large random error components. These can be avoided by presenting items used to measure change in identical test forms, with identical timings and administration conditions. Thus, we do *not* maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations. Rather, we assume that the parameter estimates are context-bound. This is the reason that the long-term trend booklets and administration procedures have not changed since the early 1980s and only a limited number of blocks of items are released after each national main assessment cycle. It was also the reason we prefer common population equating to common item equating whenever equivalent random samples are available for linking. In common item equating, items are assumed to be measuring exactly the same thing for two or more populations, despite any differences in context or administration. In common population equating, results for two or more samples from the same population are matched to one another when linking the scales. Therefore, the data from the state assessment are calibrated separately from the national NAEP data. In this case, the administration procedures differ somewhat between the state assessment and the national NAEP.

Although the IRT models are employed in NAEP only to summarize performance, a number of checks are made to detect serious violations of the assumptions underlying the models. Checks are made

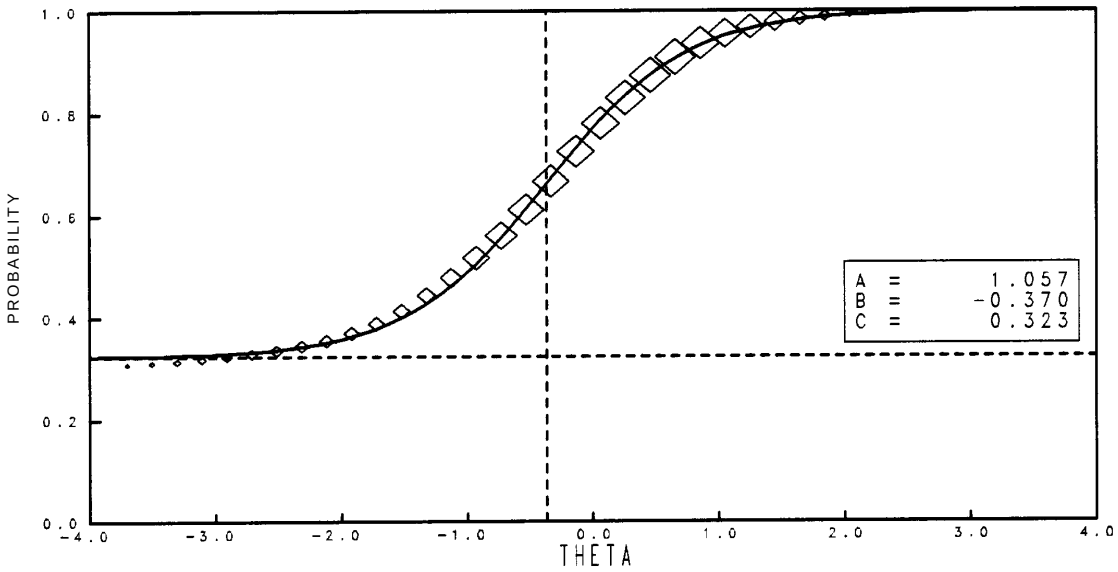
to detect multidimensionality of the construct being measured and certain condition dependencies. DIF analyses are used to examine issues of dimensionality, and what are called χ^2 statistics in the IRT literature are used to flag responses with serious departures from the IRT model. DIF analysis methodologies are discussed in Chapter 9. The latter statistics might better be called item fit statistics since they do not really have χ^2 distributions. These checks include comparisons of empirical and theoretical item response functions to identify items for which the IRT model may provide a poor fit to the data. When warranted, remedial efforts, such as collapsing categories of polytomous items or combining items into a single item, are made to mitigate the effects of such violations on inferences.

In practice, PARSCALE item fit statistics are used as a way to identify items that need further examination. Most of the statistics of this type that are available for use in this setting have distributions that are unknown. Therefore, they cannot be used for final decisions about the fit of the items to the IRT model. Because of the lack of statistical tests for IRT model fit, the fit of the IRT models to the observed data was examined within each scale by comparing the empirical item response functions (IRFs) with the theoretical curves. The primary means of accomplishing this is to generate plots of empirical versus theoretical item response curves. The theoretical curves are plots of the response functions based on the estimates of the item parameters. The empirical proportions are calculated from the posterior distributions of the thetas for each student who received the item. For dichotomous items, the sum of the values of the posterior distributions at a point on the theta scale for each student who answered an item correctly plus the sum of a fractional portion of the values of the posterior distribution at that point on the theta scale for each student who omitted the item is parallel in meaning to the number of students who actually answered the item correctly plus a fraction of the number of students who omitted the item. The sum of the values of the posterior distributions for all students receiving the item at each point on the theta scale is parallel in meaning to the empirical number of students at that point on the theta scale who received the item. The plotted values are sums of these individual posteriors at each point on the theta scale for those who got the item correct plus a fraction of the omitters divided by the sum of the posteriors of those administered the item, in the case of dichotomous items, and for those who scored in the category of interest over the sum for those who received the item, in the case of polytomous items.

As an example, Figure 12-1 contains a plot of the empirical and theoretical IRFs for a dichotomous item from the 1994 NAEP national main reading assessment. In the plot, the horizontal axis represents the theta (score) scale, the vertical axis represents the probability of a correct response. The solid curve is the theoretical IRF based on the item parameter estimates and Equation (12.1). The centers of the diamonds represent the empirical proportions correct as described above. The size of the diamonds are proportional to the sum of the posteriors at each point on the theta scale for all of those who received the item; this is related to the number of students contributing to the estimation of that empirical proportion correct.

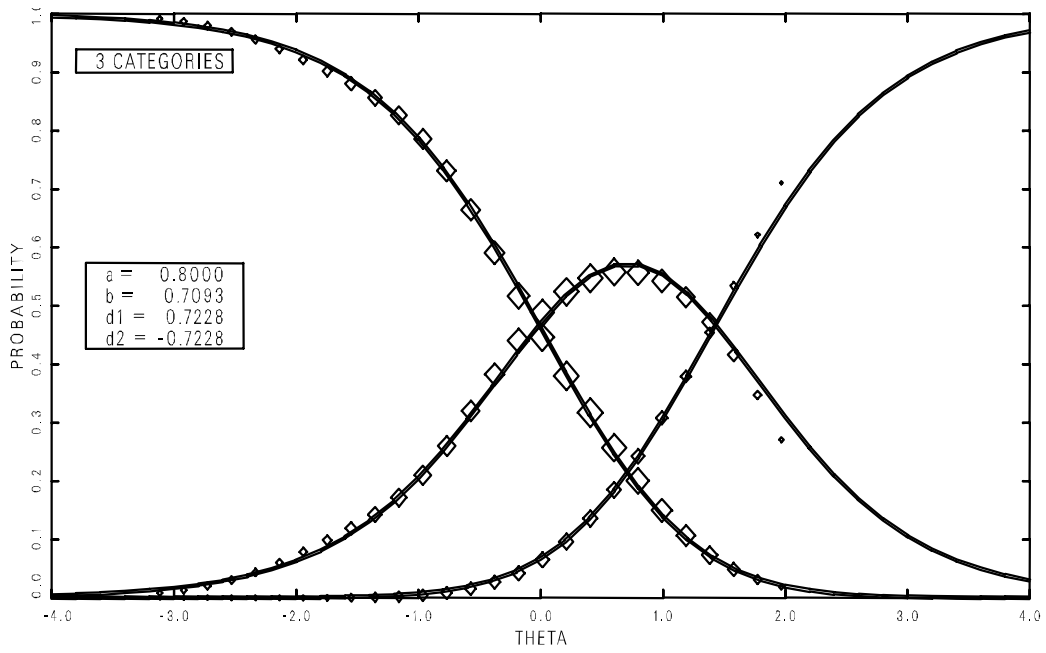
Figure 12-2 contains a plot of the empirical and theoretical IRFs for a polytomous item from the 1997 Arts (Theatre) National Assessment. As for the dichotomous item plot in Figure 12-1, the horizontal axis represents the score scale, but the vertical axis represents the probability of having a response fall in each category. The solid curves are the theoretical IRFs based on the item parameter estimates and Equation (12.3). The centers of the diamonds represent the empirical proportions of students with responses in each category and are proportional to the sum of the posteriors at each point on the theta scale for the students who received the item.

Figure 12-1
*Dichotomous Item (R016102) Exhibiting Good Model Fit**



** Diamonds represent 1994 age 13/grade 8 reading assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) assuming a logistic form.*

Figure 12-2
*Polytomous Item (HC00004) Exhibiting Good Model Fit**

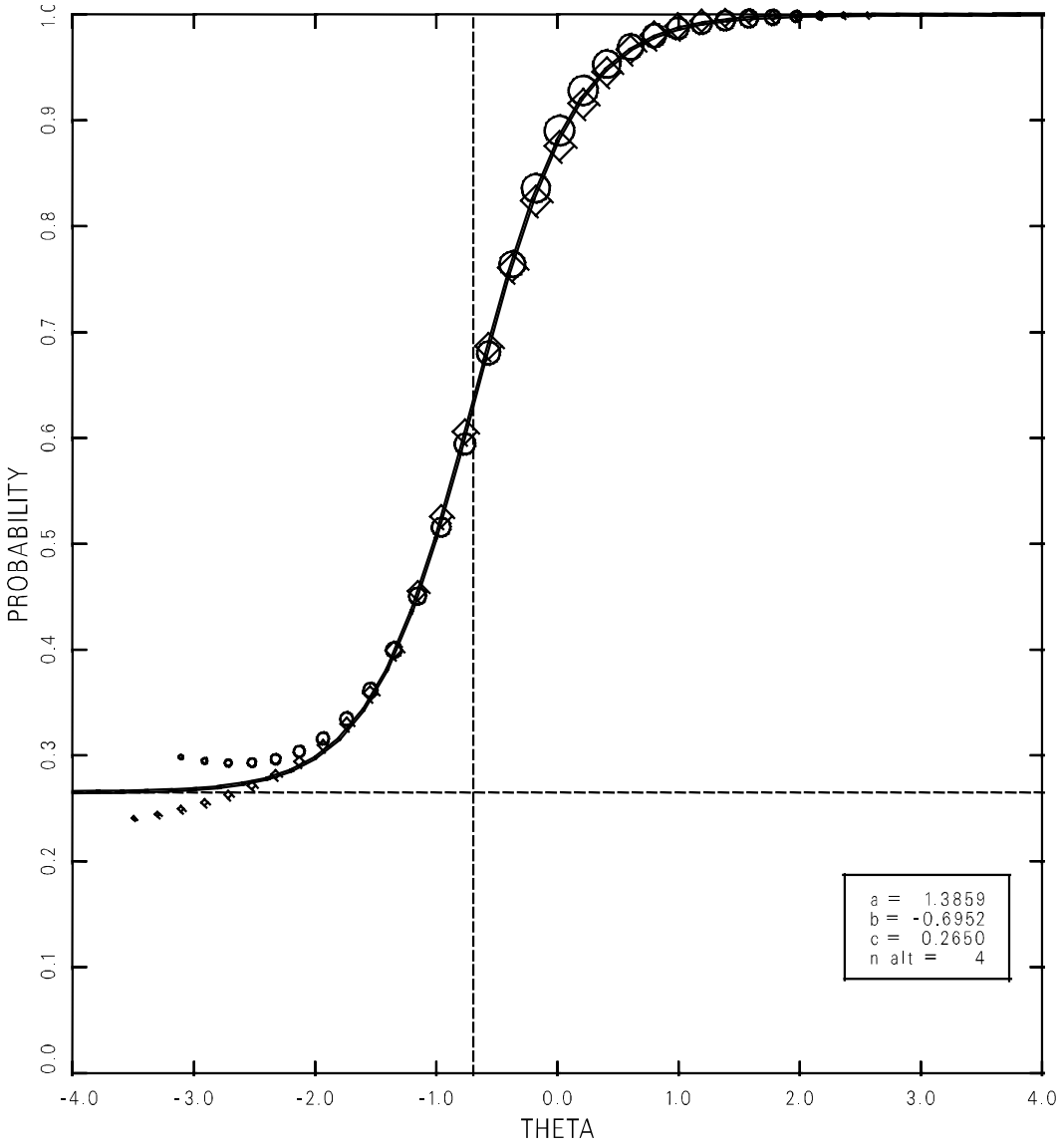


****** Diamonds represent 1997 grade 8 arts assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item category response function (ICRF) using a generalized partial credit model.*

For good fitting items, the empirical and theoretical curves are close together. Therefore, items for which this is not true are examined carefully. Examples of plots for specific items are provided in the subject-area chapters. When the same items are presented in two assessment years, the empirical curves for the two years can be compared. Normally, these curves differ somewhat due to the sampling of students for each of the two years. Figure 12-3 contains a plot for an item from the NAEP 1996 mathematics national assessment with curves of this type. When the empirical curves differ dramatically, one cause might be a change in the meaning of the item due to instructional or societal changes across the years. This type of item is ordinarily treated as two different items—one for each of the assessment years. Figure 12-4 contains the plot for an item that has been treated in this way.

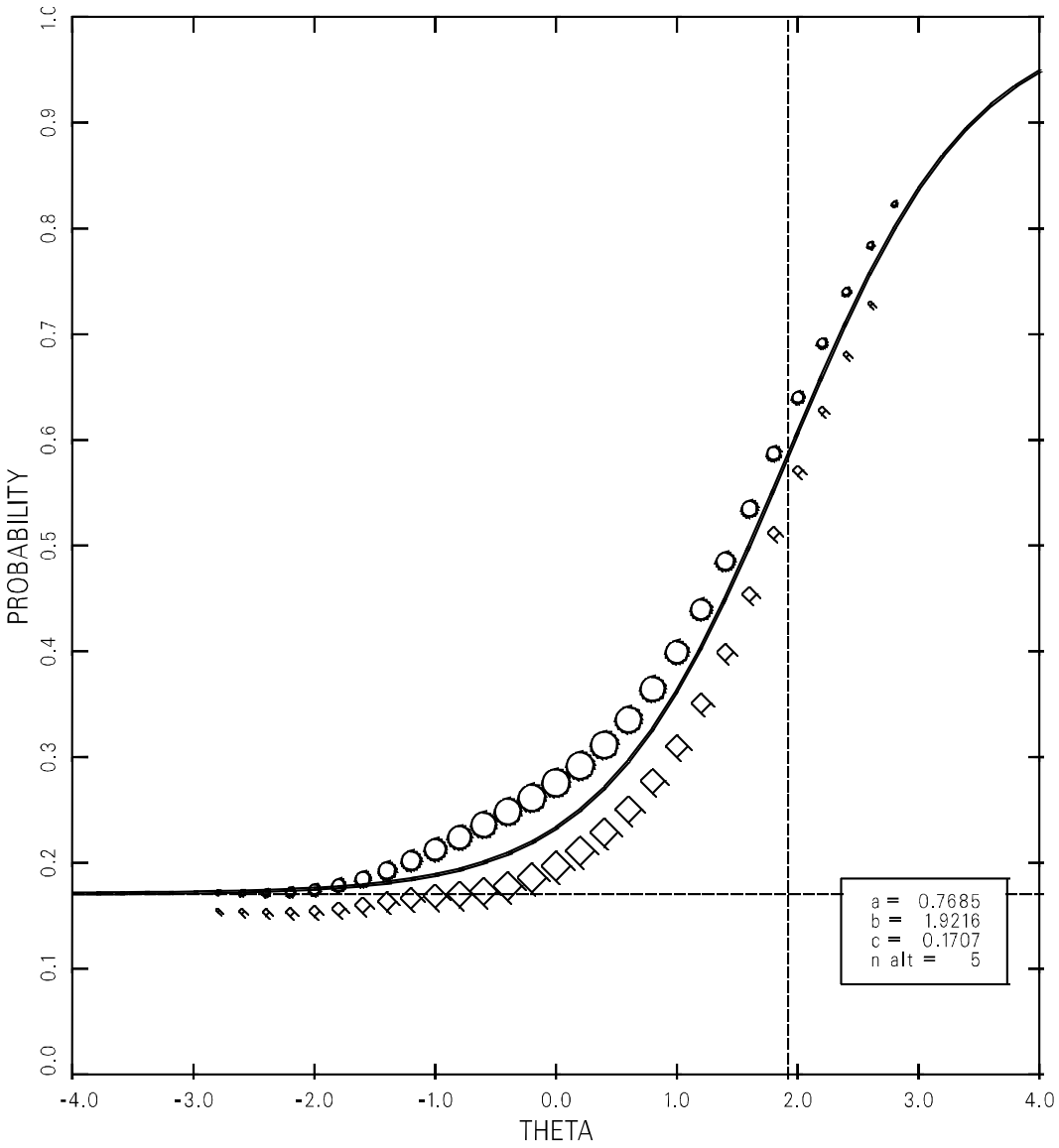
Figure 12-3

*Dichotomous Item (M017901) Exhibiting Good Model Fit Across Assessment Years**



** Circles represent 1996 grade 12 mathematics assessment data; diamonds represent 1992 grade 12 mathematics assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) assuming a logistic form.*

Figure 12-4
*Dichotomous Item (M018901) Exhibiting Different Empirical Item Functions for Different Assessment Years**



** Circles represent 1996 grade 8 mathematics assessment data; diamonds represent 1992 grade 8 mathematics assessment data. They indicate estimated conditional probabilities obtained without assuming a specific model form; the curve indicates the estimated item response function (IRF) using a generalized partial credit model..*

To summarize, using current methodologies in psychometrics, the assumption of conditional independence and the assumption that the data fit the models in Equations 12.1 and 12.3 are examined and controlled in NAEP in several ways. They are examined by considering tests of DIF, item fit statistics, and plots of empirical and theoretical IRFs. They are controlled by treating missing and "not reached" responses in reasonable ways, maintaining the context and administration of items across assessments, collapsing categories of polytomous items when appropriate, combining items into a single item, or making decisions about the inclusion or exclusion of an item in a scale based on data. The identification and amelioration of violations of IRT assumptions is an area of ongoing research in educational measurement. For example, recent studies have investigated local item dependence (Yen,

1993; Habing & Donoghue, in press), assessing the fit of the item response function (Orlando & Thissen, 2000; Donoghue & Hombo, 1999, Hombo & Donoghue, 2000), item parameter drift (Donoghue & Isham, 1998) and detecting and describing multidimensionality (e.g., Roussos, Stout, & Marden; 1998; Zhang & Stout, 1999).

12.3.2 An Overview of Plausible Values Methodology

Item response theory was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 60 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate, $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can then be approximated using an individual's $\hat{\theta}$ values as if they were θ values.

This approach breaks down in the assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a subject area scale. A first problem is that the uncertainty associated with individual θs is too large to ignore, and the features of the $\hat{\theta}$ distribution can be seriously biased as estimates of the θ distribution. (The failure of this approach was verified in early analyses of the 1984 NAEP reading survey; see Wingersky, Kaplan, & Beaton, 1987.) A second problem, occurring even with test lengths of 60, arises when test forms vary across and within assessments as to the numbers, formats, and content of the test items. The measurement error distributions thus differ even if underlying θ distributions do not, causing $\hat{\theta}$ distributions to exhibit spurious changes and resulting in deceptive comparisons in apparent population distributions—easily greater than actual differences over time or across groups. Although this latter problem is avoided in traditional standardized testing by presenting students with parallel test forms, controlled tightly across time and groups, the same constraints cannot be imposed in the design and data-collection phases of the present NAEP. Plausible values were developed as a way to estimate key population features consistently, and approximate others no worse than standard IRT procedures would, even when item booklet composition, format, and content balances change over time. A detailed development of plausible values methodology is given in Mislevy (1991). Along with theoretical justifications, that paper presents comparisons with standard procedures, discussions of biases that arise in some secondary analyses, and numerical examples. The following provides a brief overview of the plausible values approach, focusing on its implementation in NAEP analyses.

Let $\underline{y}$ represent the responses of all sampled examinees to background and attitude questions, along with variables based on the sampling design such as the school where the student is enrolled, and let $\underline{\theta}$ represent the vector of scale score values. If $\underline{\theta}$ were known for all sampled examinees, it would be possible to compute a statistic $t(\underline{\theta},\underline{y})$, such as a scale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient, to estimate a corresponding population quantity T. A function $U(\underline{\theta},\underline{y})$—for example, a jackknife estimate—would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the scaling models are latent variable models, however, $\underline{\theta}$ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by considering $\underline{\theta}$ as "missing data," and approximate $t(\underline{\theta},\underline{y})$ by its expectation given $(\underline{x},\underline{y})$, the data that actually were observed, as follows:

$$\begin{aligned} t^*(\underline{x},\underline{y}) &= E\left[t(\underline{\theta},\underline{y})\middle|\underline{x},\underline{y}\right] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}\middle|\underline{x},\underline{y})\, d\underline{\theta}. \end{aligned} \tag{12.6}$$

It is possible to approximate t^* using random draws from the predictive conditional distribution of the scale proficiencies given the item responses x_i, background variables y_i, and model parameters for sampled student i. These values are referred to as imputations in the sampling literature, and plausible values in NAEP. The value of $\underline{\theta}$ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from the respondent's conditional distribution. Rubin (1987) proposes that this process be carried out several times—multiple imputations—so that the uncertainty associated with imputation can be quantified. The average of the results of, for example, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of Equation (12.6); the variance among them, B, reflects uncertainty due to not observing $\underline{\theta}$, and must be added to the estimated expectation of $U(\underline{\theta},\underline{y})$, which reflects uncertainty due to testing only a sample of students from the population. Section 12.4 explains how plausible values are used in subsequent analyses.

It cannot be emphasized too strongly that **plausible values are *not* test scores for *individuals*** in the usual sense. Plausible values are offered only as intermediary computations for calculating integrals of the form of Equation (12.6), in order to estimate *population* characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in the contrast between plausible values and the more familiar estimates of scale score (e.g., maximum likelihood estimate or Bayes estimate) that are in some sense optimal for each examinee: *Point estimates that are optimal for individual examinees have distributions that can produce decidedly nonoptimal (specifically, inconsistent) estimates of population characteristics* (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects. For further discussion see Mislevy, Beaton, Kaplan, and Sheehan (1992).

12.3.3 Computing Plausible Values in IRT-Based Scales

Plausible values for each respondent r are drawn from the predictive conditional distribution $p(\underline{\theta}_r|\underline{x}_r,\underline{y}_r,\Gamma,\Sigma)$, where Γ and Σ are regression model parameters defined in this subsection. This subsection describes how, in IRT-based scales, these conditional distributions are characterized, and how the draws are taken. An application of Bayes' theorem with the IRT assumption of conditional independence produces

$$p\left(\underline{\theta}_r\middle|\underline{x}_r,\underline{y}_r,\Gamma,\Sigma\right) \propto P\left(\underline{x}_r\middle|\underline{\theta}_r,\underline{y}_r,\Gamma,\Sigma\right)\times p\left(\underline{\theta}_r\middle|\underline{y}_r,\Gamma,\Sigma\right) = P\left(\underline{x}_r\middle|\underline{\theta}_r\right)\times p\left(\underline{\theta}_r\middle|\underline{y}_r,\Gamma,\Sigma\right) \tag{12.7}$$

where, for vector-valued $\underline{\theta_r}$, $P\left(\underline{x_r}\middle|\underline{\theta_r}\right)$ is the product over scales of the *independent likelihoods* induced by responses to items within each scale, and $p\left(\underline{\theta_r}\middle|\underline{y_r},\Gamma,\Sigma\right)$ is the multivariate—and generally nonindependent—*joint density* of proficiencies for the scales, conditional on the observed value $\underline{y_r}$ of background responses and the parameters Γ and Σ. The provisional scales are determined by the item parameter estimates that constrain the population mean to zero and standard deviation to one. The item parameter estimates are fixed and regarded as population values in the computation described in this subsection.

In the analyses of the data from the national main assessments, a normal (Gaussian) form is assumed for $p\left(\underline{\theta_r}\middle|\underline{y_r},\Gamma,\Sigma\right)$ with a common variance-covariance matrix Σ and with a mean given by a linear model with slope parameters, Γ, based on the first approximately 200 principal components of several hundred selected main-effects and two-way interactions of the complete vector of background variables. The included principal components are referred to as the *conditioning variables*, and are denoted $\underline{y^c}$. (The complete set of original background variables used in the analyses of each subject area are listed in Appendix F.) The following model is fit to the data within each subject area:

$$\underline{\theta} = \Gamma' y^c + \underline{\varepsilon} \tag{12.8}$$

where $\underline{\varepsilon}$ is multivariately normally distributed with mean zero and variance-covariance matrix Σ. The number of principal components of the background variables used for each sample is sufficient to account for 90 percent of the total variance of the full set of background variables (after standardizing each variable). As in regression analysis, Γ is a matrix, each of whose columns contains the *effects* for one scale, and Σ is the matrix *variance-covariance of residuals* between scales.

A model similar to Equation (12.8) is used for the long-term trend assessments, with the difference that $\underline{y^c}$ consists of main effects and interactions from the smaller set of background variables (rather than principal components of those variables) available in the long-term trend assessments.

Maximum likelihood estimates of Γ and Σ, denoted by $\hat{\Gamma}$ and $\hat{\Sigma}$, are obtained with extensions of Sheehan's (1985) MGROUP computer program using the EM algorithm described in Mislevy (1985). The EM algorithm requires the computation of the mean, $\underline{\bar{\theta}_r}$, and variance-covariance matrix, Σ_r^p of the predictive conditional distribution in Equation (12.7) for respondent *r* when there are *p* scales within a subject area. For subject areas with multiple scales, the CGROUP version of the MGROUP program was used to compute the moments using higher order asymptotic corrections to a normal approximation (Thomas, 1993a). For the long-term trend assessments and other assessments with a single scale, the more precise but computationally intensive BGROUP version of MGROUP (Thomas, 1994) was used. BGROUP uses numeric quadrature to evaluate the predictive conditional distribution moments required by the E-step of the EM algorithm for one- and two-dimensional applications (Thomas, 1993a). For estimation of group means on a single scale, CGROUP (Thomas, 1994) and BGROUP results will be nearly identical to those from the original MGROUP program. CGROUP and BGROUP yield better estimates of correlations between scales, and hence better estimates of composite scale means. BGROUP will, theoretically, yield better estimates than CGROUP, but because of the heavy computational demands of the methodology used, its function is limited to bivariate scales. Hence CGROUP is used for assessments involving more than two scales.

After completion of the EM algorithm, the plausible values for all sampled respondents are drawn in the following three-step process. First, a value of Γ is drawn from a normal distribution with

mean being $\hat{\Gamma}$ and variance being the variance of $\hat{\Gamma}$. Second, conditional on the generated value of Γ and the fixed value of $\Sigma = \hat{\Sigma}$, the predictive conditional distribution mean $\underline{\bar{\theta}}_r$ and the predictive conditional distribution variance Σ_r of respondent *r* are computed from Equation 12.7 using the EM algorithm (see Thomas, 1993a). Finally, the $\underline{\theta}_r$ are drawn independently from a multivariate normal distribution with mean $\bar{\theta}_r$ and variance Σ_r approximating the distribution in Equation (12.7). These three steps are repeated five times producing five sets of imputation values for all sampled respondents.

12.4 INFERENCES ABOUT PROFICIENCIES

When survey variables are observed without error from every respondent, usual variance estimators quantify the uncertainty associated with sample statistics from the only source of uncertainty, namely the sampling of respondents. Item-level statistics for NAEP cognitive items meet this requirement, but scale score values do not. The IRT models used in their construction posit an unobservable scale score variable $\underline{\theta}$ to summarize performance on the items in a scale. The fact that $\underline{\theta}$ values are not observed even for the respondents in the sample requires additional statistical analyses to draw inferences about $\underline{\theta}$ distributions and to quantify the uncertainty associated with those inferences. As described above, Rubin's (1987) multiple imputations procedures were adapted to the context of latent variable models to produce the plausible values upon which many analyses of the data from NAEP are based. This section describes how plausible values were employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

12.4.1 Computational Procedures

Even though one does not observe the $\underline{\theta}$ value of respondent *r*, one does observe variables that are related to it: $\underline{x}_r$, the respondent's answers to the cognitive items he or she was administered in the area of interest, and $\underline{y}_r$, the respondent's answers to demographic and background variables. Suppose one wishes to draw inferences about a number $T(\underline{\theta},\underline{Y})$ that could be calculated explicitly if the $\underline{\theta}$ and $\underline{y}$ values of each member of the population were known. Suppose further that if $\underline{\theta}$ values were observable, we would be able to estimate *T* from a sample of *N* pairs of $\underline{\theta}$ and $\underline{y}$ values by the statistic $t(\underline{\theta},\underline{y})$ [where $(\underline{\theta},\underline{y}) \equiv (\theta_1,y_1,\ldots,\theta_N,y_N)$], and that we could estimate the variance in *t* around *T* due to sampling respondents by the function $U(\underline{\theta},\underline{y})$. Given that observations consist of $(\underline{x}_r,\underline{y}_r)$ rather than $(\underline{\theta}_r,\underline{y}_r)$, we can approximate *t* by its expected value conditional on $(\underline{x},\underline{y})$, or

$$t^*\left(\underline{x},\underline{y}\right) = E\left[t\left(\underline{\theta},\underline{y}\right)\middle|\underline{x},\underline{y}\right] = \int t\left(\underline{\theta},\underline{y}\right) p\left(\theta\middle|\underline{x},\underline{y}\right) d\theta. \qquad (12.9)$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\underline{\theta}_i|x_i,y_i)$, which are obtained for all respondents by the method described in Section 12.3.3. Let $\hat{\underline{\theta}}_m$ be the m^{th} such vector of plausible values, consisting of a multidimensional value for the latent variable of each respondent. This vector is a plausible representation of what the true $\underline{\theta}$ vector might have been, had we been able to observe it.

The following steps describe how an estimate of a scalar statistic $t(\underline{\theta},\underline{y})$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Five sets of plausible values are used in NAEP analyses.)

1. Using each set of plausible values $\hat{\underline{\theta}}_m$ in turn, evaluate t as if the plausible values were true values of $\underline{\theta}$. Denote the results $\hat{t}_m$, for $m = 1, \ldots, M$.

2. Using the jackknife variance estimator defined in Chapter 10, compute the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

3. The final estimate of t is

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M} \tag{12.10}$$

4. Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M} \tag{12.11}$$

5. Compute the variance among the M estimates $\hat{t}_m$, to approximate the between-imputation variance

$$B = \sum_{m=1}^{M} \frac{\left(\hat{t}_m - t^*\right)^2}{\left(M-1\right)} \tag{12.12}$$

6. The final estimate of the variance of t^* is the sum of two components

$$V = U^* + \left(1 + M^{-1}\right) B \tag{12.13}$$

In this equation, $(1+M^{-1})B$ is the estimate of variance due to the latency of $\underline{\theta}$. Due to the excessive computation that would be required, NAEP analyses do not compute and average jackknife variances over all five sets of plausible values, but uses that computed from the first set. Thus, in NAEP reports, U^* is approximated by U_1.

12.4.2 Statistical Tests

The variance described in Section 12.4.1 is used to make statistical tests comparing NAEP results. This section describes the relationships between these tests and the variance components described above. Chapter 13 contains details of the hypothesis tests used in this assessment.

If $\underline{\theta}$ values were observed for all sampled students, the statistic $(t - T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom, where d is calculated in the usual way. Then the incomplete-data statistic $(t^* - T)/V^{1/2}$ is approximately t-distributed, with degrees of freedom (Johnson & Rust, 1993; Satterthwaite, 1941) given by

$$\nu = \frac{1}{\frac{f^2}{M-1} + \frac{(1-f)^2}{d}} \tag{12.14}$$

where f is the proportion of total variance due to not observing $\underline{\theta}$ values:

$$f = (1 + M^{-1}) B/V \tag{12.15}$$

When B is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case with main NAEP reporting variables. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional $\underline{t}$, such as the k coefficients in a multiple regression analysis, each U_m and U^* is a covariance matrix, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $(T - \underline{t}^*)\, V^{-1}\, (T - \underline{t}^*)$, is approximately F distributed, with degrees of freedom equal to k and with ν defined as above but with a matrix generalization of f:

$$f = (1 + M^{-1})\ Trace\ (BV^{-1})/k\ . \tag{12.16}$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices for multivariate $\underline{t}$.

12.4.3 Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables $\underline{y^c}$ are consistent estimates of the corresponding population values T. This includes interrelationships among scales within a content area that have been treated in the multivariate manner described above in Section 12.3.3. Statistics involving background variables y that were *not* conditioned on, or relationships among scale scores from *different* purposes, content strands or fields, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the variables that were conditioned on and to the scale score of interest. That is, the large sample expectations of certain sample statistics need not equal the true population parameters.

The *direction* of the bias is typically to underestimate the effect of nonconditioned variables. For details and derivations see Beaton and Johnson (1990), Mislevy (1991), and Mislevy and Sheehan (1987, Section 10.3.5). For a given statistic t^* involving one content area and one or more nonconditioned background variables, the *magnitude* of the bias is related to the extent to which observed responses $\underline{x}$ account for the latent variable $\underline{\theta}$, and the degree to which the nonconditioned background variables are explained by conditioning background variables. The first factor—conceptually related to test reliability—acts consistently in that greater measurement precision reduces biases in *all* secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular:

- High shared variance between conditioned and nonconditioned background variables *mitigates* biases in analyses that involve only scale score and nonconditioned variables, such as marginal means or regressions.
- High shared variance *exacerbates* biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

The large number of background variables that have been included in the conditioning vectors for the 1996 assessments allows a large number of secondary analyses to be carried out with little or no bias, and mitigates biases in analyses of the marginal distributions of $\underline{\theta}$ in nonconditioned variables. Analysis of the 1988 NAEP reading data (some results of which are summarized in Mislevy, 1991), which had a similar design and fewer conditioning variables, indicates that the potential bias for nonconditioned variables in multiple regression analyses is below 10 percent, and biases in simple regression of such variables is below 5 percent. Additional research (summarized in Mislevy, 1990) indicates that most of the bias reduction obtainable from conditioning on a large number of variables can be captured by instead conditioning on the first several principal components of the matrix of all original conditioning variables. This procedure was adopted for the 1992, 1994, and 1996 national main assessments by replacing the conditioning effects by the first K principal components, where K was selected so that 90 percent of the total variance of the full set of conditioning variables (after standardization) was captured. Mislevy (1990) shows that this puts an upper bound of 10 percent on the average bias for all analyses involving the original conditioning variables.

12.4.4 A Numerical Example

To illustrate how plausible values are used in subsequent analyses, this subsection gives some of the steps in the calculation of the 1992 grade 4 reading composite mean and its estimation-error variance. This illustration is an example of the calculation of NAEP means and variances and can be used to understand their calculation for any NAEP assessment.

The weighted mean of the first plausible values of the reading composite for the grade 4 students in the sample is 217.79, and the jackknife variance of these values is 0.833. Were these values true θ values, then 217.79 would be the estimate of the mean and 0.833 would be the estimation-error variance. The weighted mean of the second plausible values of the same students, however, is 217.62; the third, fourth, and fifth plausible values give weighted means of 217.74, 218.24, and 218.05. Since all of these figures are based on precisely the same sample of students, the variation among them is due to uncertainty about the students' θs, having observed their item responses and background variables. Consequently, our best estimate of the mean for grade 4 students is the average of the five plausible values: 217.89. Taking the jackknife variance estimate from the first plausible value, 0.833, as our estimate U^* of sampling variance, and the variance among the five weighted means, .063, as our estimate B of uncertainty due to not observing θ, we obtain as the final estimate V of total error variance $0.833 + (1+5^{-1})\ .063 = 0.909$.

It is also possible to partition the estimation error variance of a statistic using these same variance components. The proportion of error variance due to sampling students from the population is U^*/V, and the proportion due to the latent nature of θ is $(1+M^{-1})B/V$. The results are shown in Table 12-1. The value of U^*/V roughly corresponds to reliability in classical test theory and indicates the amount of information about an average individual's θ present in the observed responses of the individual. It should be recalled again that the objective of NAEP is not to estimate and compare values of individual examinees, the accuracy of which is gauged by reliability coefficients. The objective of NAEP, rather, is

to estimate population and subpopulation characteristics, and the marginal estimation methods described above have been designed to do so consistently regardless of the values of reliability coefficients.

Table 12-1
Estimation Error Variance and Related Coefficients for the 1992 Grade 4 Reading Composite (Based on Five Plausible Values)

			Proportion of Variance Due to...	
			Student Sampling:	Latency of θ:
U*	**$(1+5^{-1})B$**	**V**	***U*/V***	**$(1+5^{-1})B/V$**
0.833	0.076	0.908	0.92	0.08

Chapters 16, 17, 20, 21, and 24 and Appendix H provide values of the proportion of variance due to sampling and due to the latent nature of θ for all 1996 scales and composites for the populations as a whole and, in the appendix, for selected subpopulations. It will be seen that the proportion of variance due to the latency of θ varies somewhat among subject areas, tending to be largest for the long-term trend writing assessment, where there is low correlation between tasks and each student responded to only one or at most two tasks. The proportion of variance due to latency of θ is smallest for the composites of the national main assessment subjects with several scales, where the number of items per student is largest. Essentially, the variance due to the latent nature of θ is largest when there is less information about a student's scale score. (Note the distinction between estimation error variance of a parameter estimate and the estimate of the variance of the θ distribution. The former depends on the accuracy of measurement; the large-sample model-based expected value of the latter does not.) Given fixed assessment time, this decrease in information will occur whenever the amount of information per unit time decreases as can happen when many short constructed-response or multiple-choice items are replaced by a few extended constructed-response items.

12.5 DESCRIBING STUDENT PERFORMANCE

Since its beginning, a goal of NAEP has been to inform the public about what students in United States schools know and can do. While the NAEP scales provide information about the distributions of scale scores for the various subpopulations, they do not directly provide information about the meaning of various points on the scale. Traditionally, meaning has been attached to educational scales by norm-referencing—that is, by comparing students at a particular scale level to other students. In contrast, NAEP achievement levels and scale anchors describe selected points on the scale in terms of the types of skills that are likely to be exhibited by students scoring at that level. In addition, each NAEP item is mapped to a point on its corresponding scale, so that the content of each item provides information about what students at each score level can do in a probabilistic sense. The achievement level process has been applied to the reading, mathematics, science, U.S. history, and geography composites and to the writing and civics unidimensional scales. The achievement levels were set for reading in 1992, mathematics in 1990, science in 1996, U.S. history and geography in 1994, and writing and civics in 1998.

12.5.1 Achievement Levels

NAGB has determined that achievement levels shall be the first and primary way of reporting NAEP results. Setting achievement levels is a method for setting standards on the NAEP assessment that identifies what students should know and be able to do at various points on the composite. For each grade of each subject, three levels were defined—basic, proficient, and advanced. Based on initial policy

definitions of these levels, panelists were asked to determine operational descriptions of the levels appropriate with the content and skills assessed in the assessment. With these descriptions in mind, the panelists were then asked to rate the assessment items in terms of the expected performance of marginally acceptable examinees at each of these three levels. These ratings were then mapped onto the NAEP scale to obtain the achievement level cutpoints for reporting. Further details of the achievement level setting process for subject areas appear in Appendix I for reading and Appendix J for writing and civics.

12.5.2 Item Mapping Procedures

In order to map items (questions) to particular points on each subject area scale, a response probability convention had to be adopted that would divide those who had a higher probability of success from those who had a lower probability. Establishing a response probability convention has an impact on the mapping of assessment items onto the scales. A lower boundary convention maps the items at lower points along the scales, and a higher boundary convention maps the same items at higher points along the scales. The underlying distribution of skills in the population does not change, but the choice of a response probability convention does have an impact on the proportion of the student population that is reported as "able to do" the items on the scales.

There is no obvious choice of a point along the probability scale that is clearly superior to any other point. If the convention were set with a boundary at 50 percent, those above the boundary would be more likely to get an item right than get it wrong, while those below that boundary would be more likely to get the item wrong than right. While this convention has some intuitive appeal, it was rejected on the grounds that having a 50/50 chance of getting the item right shows an insufficient degree of mastery. If the convention were set with a boundary at 80 percent, students above the criterion would have a high probability of success with an item. However, many of the students below this criterion show some level of achievement that would be ignored by such a stringent criterion. In particular, those in the range between 50 and 80 percent correct would be more likely to get the item right than wrong, yet would not be in the group described as "able to do" the item.

In a compromise between the 50 percent and the 80 percent conventions, NAEP has adopted two related response probability conventions: 74 percent for multiple-choice items (to correct for the possibility of answering correctly by guessing), and 65 percent for constructed-response items (where guessing is not a factor). These probability conventions were established, in part, based on an intuitive judgment that they would provide the best picture of students' knowledge and skills.

Some additional support for the dual conventions adopted by NAEP was provided by Huynh (1994, 1998). He examined the IRT information provided by items, according to the IRT model used in scaling NAEP items. Following Bock (1972), Huynh decomposed the item information into that provided by a correct response [$P_{ji}(\theta) \bullet I_j(\theta)$] and that provided by an incorrect response [$(1\text{-}P(\theta)) \bullet I(\theta)$]. Huynh showed that the item information provided by a correct response to a constructed-response item is maximized at the point along the scale at which two-thirds of the students get the item correct (for multiple-choice items with four options, information is maximized at the point at which 75 percent get the item correct). Maximizing the item information, $I(\theta)$, rather than the information provided by a correct response [$P(\theta) \bullet I(\theta)$], would imply an item-mapping criterion closer to 50 percent. Maximizing just the item information, $I(\theta)$, takes into account both responses that are correct and those that are incorrect, however.

For dichotomously scored items the information function as defined by Birnbaum (1968, p. 463) is defined for the j^{th} item as

$$I_j(\theta) = \frac{(1.7a_j)^2 P_{j0}(\theta_k)\left[P_{j1}(\theta_k) - c_j\right]^2}{P_{j1}(\theta_k)(1-c_j)^2}, \tag{12.17}$$

where the notation is the same as that used in Equations (12.1) and (12.2). The item information function was defined by Samejima (1969) in general for polytomously scored items, and has been derived for items scaled by the generalized partial credit model (Donoghue, 1993; Muraki, 1993) as (in a slightly different, but equivalent form)

$$I_j(\theta) = (1.7a_j)^2 \left[\sum_{i=0}^{m_j-1} i^2 P_{ji}(\theta_k) - \left\{ \sum_{i=0}^{m_j-1} i P_{ji}(\theta_k) \right\}^2 \right]. \tag{12.18}$$

12.6 OVERVIEW OF THE 1998 NAEP SCALES

The following IRT scale score analyses were carried out for each grade in the 1998 NAEP assessment:

- Reading: Three IRT scales linked back to the 1992 and 1994 main assessments of reading. These three scales, along with a composite scale, are associated with the 1998 main and state assessments.
- Writing: A single newly developed IRT scale for each grade for the main and state assessments of writing.
- Civics: A single newly developed IRT scale for each grade for the main assessment of civics.

Details are in the following chapters.

Chapter 13

CONVENTIONS USED IN HYPOTHESIS TESTING AND REPORTING NAEP RESULTS[1]

Spencer S. Swinton, David S. Freund, and Nancy L. Allen
Educational Testing Service

13.1 OVERVIEW

Results for the 1998 NAEP assessments were disseminated in several different reports: the *NAEP 1998 Reading Report Card for the Nation and the States* (Donahue, Voelkl, Campbell, & Mazzeo, 1999), the *NAEP 1998 Writing Report Card for the Nation and the States* (Greenwald, Persky, Campbell, & Mazzeo, 1999), the *NAEP 1998 Civics Report Card for the Nation* (Lutkus, Weiss, Campbell, Mazzeo, and Lazer, 1999), and, published only on the web, summary data tables for each report. These reports are published on the NCES/NAEP web site *http://nces.ed.gov/nationsreportcard*. Several other reports based on 1998 NAEP data will be forthcoming.

The *NAEP 1998 Reading Report Card for the Nation and the States*, the *NAEP 1998 Writing Report Card for the Nation and the States*, and the *NAEP 1998 Civics Report Card for the Nation* highlight key assessment results for the nation and summarize results across the jurisdictions participating in the assessments. These reports contain composite scale score results (e.g., scale score means) for the nation, for each of the four regions of the country, and for public-school students within each jurisdiction participating in the state assessments of reading and writing, both overall and by primary reporting variables. The seven key reporting variables (referred to here as primary reporting variables) are gender, race/ethnicity, level of parents' education, Title I participation, eligibility for free or reduced cost school lunch, type of location, and type of school (public, Catholic schools, other religious schools, and other private schools). For public-school students, scale score means were reported for a variety of other subpopulations defined by responses to items from the student, teacher, and school questionnaires and by school and location demographic variables provided by Westat[2]. Upcoming reports will include estimates of scale score means and selected percentiles for specific subgroups of students of interest in each report.

The second type of summary report is an electronically delivered collection of summary data tables (available on the NCES/NAEP web site) that contain detailed breakdowns of the scale score data for each sample according to the responses to the student, teacher, and school questionnaires for the public-school, nonpublic-school, and combined populations as a whole and for important subgroups of the public-school population, as defined by the primary reporting variables. There are six sections in each collection of summary data tables:

[1] Spencer S. Swinton played a role in making decisions about hypothesis-testing methods and procedures and worked with David S. Freund, who implemented many of the methods and procedures in computer programs. Nancy L. Allen contributed to the current version of this chapter.

[2] Some of these variables were used by Westat, in developing the sampling frame for the assessment and in drawing the sample of participating schools.

Student Summary Data Tables break down the composite scale score data according to the students' responses to questions in the three student questionnaires (common core, subject-specific background, and motivational section) included in the assessment booklets.

Teacher Summary Data Tables break down the composite scale score data according to the teachers' responses to questions in teacher questionnaires, where they are available.

School Summary Data Tables break down the composite scale score data according to the principals' (or other administrators') responses to questions in the school characteristics and policies questionnaire.

Question Summary Data Tables provide the response data (percent of students choosing each option) for each cognitive item in the assessment.

Achievement-Level Summary Data Tables provide estimates of the percentage of students at or above each achievement level as well as the percentage of students below the *Basic* level.

Percentile Summary Data Tables provide selected composite-scale and subscale percentiles for the public-school, nonpublic-school, and total populations and for the major demographic subgroups of the national school population.

The production of the *Report Cards* and the summary data tables required many decisions about a variety of data analysis and statistical issues. For example, certain categories of the reporting variables contained limited numbers of examinees. A decision was needed as to what constituted a sufficient sample size to permit the reliable reporting of subgroup results, and which, if any, estimates were sufficiently unreliable to need to be "flagged" as a caution to readers. As a second example, the performance for subgroups of students were compared. A number of inferential rules, based on logical and statistical considerations, had to be developed to ensure that conclusions are adequately supported by the data from the assessment. Practical comparison procedures were required to control for Type I errors without paying too large a penalty with respect to the statistical power for detecting real and substantively interesting differences. Prior to 1998, the Bonferroni procedure (Hochberg, 1988) was the principal method used by NAEP to protect against Type I error. Currently, a new multiple comparison criterion, false discovery rate or FDR (Benjamini & Hochberg, 1994), is used. FDR controls the *rate* of false rejections (e.g., 5 false rejections per 100 rejections), rather than controlling the probability of one such error (familywise error rate, or FWE), as the Bonferroni procedure does. To implement the use of the FDR, the 1994 procedure of Benjamini and Hochberg was selected.

The purpose of this chapter is to document the major conventions and statistical procedures used in generating the *Report Cards* and the summary data tables. Additional details about procedures relevant to the *Report Cards* can be found in the text and technical appendices of those reports. Information is available on the Internet, describing procedures used in creating the summary data tables.

13.2 MINIMUM SCHOOL AND STUDENT SAMPLE SIZES FOR REPORTING SUBGROUP RESULTS

In all of the reports, estimates of quantities such as composite and scale score means and percentages of students indicating particular levels of background variables (as measured in the student, teacher, and school questionnaires) are reported for the population of students in each grade. These estimates are also reported for certain key subgroups of interest as defined by primary NAEP reporting

variables. Where possible, NAEP reports results for gender, for five racial/ethnic subgroups (White, Black, Hispanic, Asian American/Pacific Islander, and American Indian/Alaskan Native), three types of locations (central cities, urban fringes/large towns, rural/small town areas), four levels of parents' education (did not finish high school, high school graduate, some college, college graduate), Title 1 participation, eligibility for the free or reduced-cost school lunch component of the National School Lunch Program, and type of school. However, for some regions of the country and sometimes for the nation as a whole, school and/or student sample sizes were too small for one or more of the categories of these variables to permit accurate reporting.

A consideration in deciding whether to report an estimated quantity is whether the sampling error is too large to permit effective use of the estimates. A second, and equally important, consideration is whether the standard error estimate that accompanies a statistic is itself sufficiently accurate to inform potential readers about the reliability of the statistic. The precision of a sample estimate (be it sample mean or standard error estimate) for a population subgroup from a three-stage sample design (the one used to select samples for the national assessments) is a function of the sample size of the subgroup and of the distribution of that sample across first-stage sampling units (i.e., PSUs in the case of the national assessments). Hence, both of these factors were used in establishing minimum sample sizes for reporting.

Here a decision was reached to report subgroup results only if the student sample size exceeded 61.[3] A design effect of two was assumed for this decision, implying a sample design-based variance twice that of simple random sampling. This assumption is consistent with previous NAEP experience (Johnson & Rust, 1992). In carrying out the statistical power calculations when comparing a subgroup to the total group, it was assumed that the total population sample size is large enough to contribute negligibly to standard errors. Furthermore, it was required that the students within a subgroup be adequately distributed across PSUs to allow for reasonably accurate estimation of standard errors. In consultation with Westat, a decision was reached to publish only those statistics that had standard error estimates based on five or more degrees of freedom. The same minimum student and PSU sample size restrictions were applied to proportions and to comparisons of percentages or proportions as well as average scale scores and comparisons of average scale scores.

13.3 IDENTIFYING ESTIMATES OF STANDARD ERRORS WITH LARGE MEAN SQUARED ERRORS

As noted above, standard errors of average scale scores, proportions, and percentiles play an important role in interpreting subgroup results and in comparing the performances of two or more subgroups. The jackknife standard errors reported by NAEP are statistics whose quality depends on certain features of the sample from which the estimate is obtained. In certain cases, the mean squared error[4] associated with the estimated standard errors may be quite large. This result typically occurred when the number of students upon which the standard error is based is small or when this group of students comes from a small number of participating PSUs. The minimum PSU and student sample sizes that were imposed in most instances suppressed statistics where such problems existed. However, the possibility remained that some statistics based on sample sizes that exceed the minimum requirements had standard errors that were not well estimated. Therefore, in the reports, estimated standard errors for published statistics that are themselves subject to large mean squared errors are followed by the symbol "!".

[3] This number was obtained by determining the sample size necessary to detect an effect size of 0.5 with a probability of 0.8 or greater.

[4] The mean squared error of the estimated standard error is defined as $\mathscr{E}[\hat{s} - \sigma]^2$, where $\hat{s}$ is the estimated standard error, σ is the "true" standard error, and $\mathscr{E}$ is the expectation, or expected value operator.

The magnitude of the mean squared error associated with an estimated standard error for the mean or proportion of a group depends on the coefficient of variation (*CV*) of the estimated size of the population group, denoted as $\hat{N}$ (Cochran, 1977, Section 6.3). The coefficient of variation is estimated by:

$$CV(\hat{N}) = \frac{SE(\hat{N})}{\hat{N}}$$

where $\hat{N}$ is a point estimate of *N* and $SE(\hat{N})$ is the jackknife standard error (described in Chapter 10 of this report) of $\hat{N}$.

Experience with previous NAEP assessments suggests that when this coefficient exceeds 0.2, the mean squared error of the estimated standard errors of means and proportions based on samples of this size may be quite large. (Further discussion of this issue can be found in Johnson & Rust, 1992.) Therefore, the standard errors of means and proportions for all subgroups for which the coefficient of variation of the population size exceeds 0.2 are marked as described above. In the *Report Cards* and the summary data tables, statistical tests involving one or more quantities that have standard errors, confidence intervals, or significance tests so flagged should be interpreted with caution.

13.4 TREATMENT OF MISSING DATA FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

As previously described, responses to the student, teacher, and school questionnaires played a prominent role in all reports. Although the return rate on all three types of questionnaire was high,[5] there were missing data for each type of questionnaire.

The reported estimated percentages of students in the various categories of background variables, and the estimates of the average scale score of such groups, were based on only those students for whom data on the background variable were available. In the terminology of Little and Rubin (1987), the analyses pertaining to a particular background variable presented in the reports are contingent on the assumption that the data are missing completely at random.[6]

The estimates of proportions and proficiencies based on “missing completely at random” assumptions are subject to potential nonresponse bias if, as may be the case, the assumptions are not correct. The amount of missing data was small (usually, less than 2%) for most of the variables obtained from the student, school, and teacher questionnaires. For analyses based on these variables, reported results are subject to little, if any, nonresponse bias. However, for particular background items in these questionnaires, the level of nonresponse was somewhat higher, and so the potential for nonresponse bias is also somewhat greater. Results for background questions for which more than 10 percent of the responses were missing should be interpreted with caution.

To analyze the relationships among teachers’ questionnaire responses and their students’ achievement, each teacher’s questionnaire had to be matched to the students who were taught by that teacher. If a student could not be matched to a teacher, all teacher questionnaire responses are missing for that student. Lower percentages of students with teacher questionnaire data indicate that there is less

[5] Information about survey participation rates (both school and student), as well as proportions of students excluded by each jurisdiction from the assessment, is given in Appendix A. Sampling adjustments intended to account for school and student nonresponse are described in Chapters 10 and 11.

[6] The term "missing completely at random" means that the mechanism generating the missing data is independent of the response to the particular background items and the scale score.

certainty about results for variables from the teacher questionnaire. Note that these match rates do not reflect the additional missing data due to item-level nonresponse. The amount of additional item-level nonresponse in the returned teacher questionnaires can be found in the summary data tables.

13.5 HYPOTHESIS-TESTING CONVENTIONS

13.5.1 Comparing Means and Proportions for Different Groups of Students

Many of the group comparisons explicitly commented on in the reports involved mutually exclusive sets of students. Examples include comparisons of the average scale score for male and female students, White and Hispanic students, students attending schools in central city and urban fringe or large-town locations, students who reported watching six or more hours of television each night, and students who report watching less than one hour of television each night.

The text in the reports indicate that means or proportions from two groups were different only when the difference in the point estimates for the groups being compared was statistically significant at an approximate simultaneous α level of .05. An approximate procedure was used for determining statistical significance NAEP staff judged to be statistically defensible, as well as being computationally tractable. Although all pairs of levels within a variable were tested and reported in the summary data tables, some text within the reports was developed for only a subset of these comparisons, although the family size was maintained at that of the original tests. For example, text was included in the reports to compare the majority ethnic group and each minority group, but text for all possible comparisons of groups may not have been included. The procedure used to make statistical tests is described in the following paragraphs.

Let A_i be the statistic in question (e.g., a mean for group *i*) and let S_{A_i} be the jackknife standard error of the statistic. The text in the reports identified the means or proportions for groups *i* and *j* as being different if:

$$\frac{|A_i - A_j|}{\sqrt{S^2_{A_i(A_i)} + S^2_{A_j(A_j)}}} \geq T_{\frac{.05}{2c}}$$

where T_α is the (1 - α) percentile of the *t* distribution with degrees of freedom, *df*, as estimated below, and *c* is the number of related comparisons being tested. See the following section (Section 13.5.2) for a more specific description of multiple comparisons. In cases where group comparisons were treated as individual units, the value of *c* was taken as 1, and the test statistic was equivalent to a standard two-tailed *t*-test for independent samples. When *c* is greater than 1, this test is based on the Benjamini and Hochberg (1995) procedure of controlling the FDR, described below.

The procedures in this section assume that the data being compared are from independent samples. Because of the sampling design in which PSUs, schools, and students within school are randomly sampled, the data from mutually exclusive sets of students may not be strictly independent. Therefore, the significance tests employed are, in many cases, only approximate. Another procedure, one that does not assume independence, could have been conducted. However, that procedure is computationally burdensome. A comparison of the standard errors using the independence assumption and the correlated group assumption was made using NAEP data. The estimated standard error of the difference based on independence assumptions was approximately 10 percent larger than the more complicated estimate based on correlated groups. In almost every case, the correlation of NAEP data across groups was positive. Because, in NAEP, significance tests based on assumptions of independent

samples are only somewhat conservative, the approximate (assuming independence) procedure was used for most comparisons.

Because of clustering and differential weighting in the sample, the degrees of freedom are less than for a simple random sample of the same size. The degrees of freedom of this *t*-test is defined by a Satterthwaite (Johnson & Rust, 1992) approximation as follows:

$$df = \frac{(\sum_{k=1}^{N} S_{A_k}^2)^2}{\sum_{k=1}^{N} \frac{S_{A_k}^4}{df_{A_k}}}$$

where N is the number of subgroups involved, and df_{A_k} is as follows:

$$df_{A_k} = \left(3.16 - \frac{2.77}{\sqrt{m}}\right)\left[\frac{\left(\sum_{j=1}^{m}\left(t_{jk} - t_k\right)^2\right)^2}{\sum_{j=1}^{m}\left(t_{jk} - t_k\right)^4}\right]$$

where m is the number of jackknife replicates (usually 62 in NAEP), t_j is the j^{th} replicated estimate for the mean of a subgroup, and t_k is the estimate of the subgroup mean using the overall weights and the first plausible value.

The number of degrees of freedom for the variance equals the number of independent pieces of information used to generate the variance. In the case of data from NAEP, the 62 pieces of information are the squared differences $(t_{jk} - t_k)^2$, each supplying at most one degree of freedom (regardless of how many individuals were sampled within PSUs). If some of the squared differences $(t_{jk} - t_k)^2$ are much larger than others, the variance estimate of m_k is predominantly estimating the sum of these larger components, which dominate the remaining terms. The effective degrees of freedom of S_{A_k} in this case will be nearer to the number of dominant terms. The estimate df_{A_k} reflects these relationships.

The two formulae above show us that when df_{A_k} is small, the degrees of freedom for the *t*-test, *df*, will also be small. This will tend to be the case when only a few PSU pairs have information about subgroup differences relevant to a *t*-test. It will also be the case when a few PSU pairs have subgroup differences much larger than other PSU pairs.

The procedures described above were used for testing differences of both means *and* nonextreme percentages. The approximation for the test for percentages works best when sample sizes are large, and the percentages being tested have magnitude relatively close to 50 percent. Statements about group differences should be interpreted with caution if at least one of the groups being compared is small in size or if "extreme" percentages are being compared.

Differences in percentages were treated as involving "extreme" percentages if for either percentage, *P*:

$$P < P_{lim} = \frac{200}{N_{EFF} + 2},$$

where the effective sample size is

$$N_{EFF} = \frac{P(100 - P)}{(SE_{JK})^2}, \text{ and } SE_{JK}$$

is the jackknife standard error of *P*. Similarly, at the other end of the 0 – 100 scale, a percentage is deemed extreme if $100 - P < P_{lim}$. In either extreme case, the normal approximation to the distribution is a poor approximation, and the value of *P* was reported, but no standard error was estimated and hence no significance tests were conducted.

13.5.2 Multiple Comparison Procedures

Frequently, groups (or families) of comparisons were made and were presented as a single set. The appropriate text, usually a set of sentences or a paragraph, was selected for inclusion in a report based on the results for the entire set of comparisons. For example, some reports contain a section that compared average scale scores for a predetermined group, generally the majority group (in the case of race/ethnicity, for example, White students) to those obtained by other minority groups. The entire set of tests was presented in the summary data tables. The procedures described above and the certainty ascribed to intervals (e.g., a 95 % confidence interval) are based on statistical theory that assumes that only one confidence interval or test of statistical significance is being performed. However, in some sections of a report, many different groups are compared (i.e., multiple sets of confidence intervals are being analyzed). In sets of confidence intervals, statistical theory indicates that certainty associated with the entire set of intervals is less than that attributable to each individual comparison from the set. To hold the significance level for the set of comparisons at a particular level (e.g., .05), adjustments—called "multiple comparison procedures"—must be made to the methods described in the previous section. One such procedure, the false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995) was used to control the certainty level.

Unlike the other multiple comparison procedures (e.g., the Bonferroni procedure) that control the familywise error rate (i.e., the probability of making even one false rejection in the set of comparisons), the FDR procedure controls the expected proportion of falsely rejected hypotheses. Furthermore, familywise procedures are considered conservative for large families of comparisons (Williams, Jones, & Tukey, 1999). Therefore, the FDR procedure is more suitable for multiple comparisons in NAEP than other procedures.

The 1998 assessment is the first time NAEP has used the Benjamini-Hochberg procedure to maintain FDR for all multiple comparisons. Prior to the 1996 assessment, the Bonferroni procedure was used for multiple comparisons. In 1996, either the Bonferroni or Benjamini-Hochberg FDR procedure was used, depending on the testing situation. The Benjamini-Hochberg FDR procedure was used for large numbers of comparisons (i.e., any comparisons involving all of the states): (a) all pairwise comparisons of the states; (b) all comparisons of individual states to the national average; and (c) the trend for each state, which compared the current mean for the state to the state's mean in the previous

assessment. All other multiple comparisons for the 1996 assessment used the Bonferroni procedure. The 1994 NAEP reading assessments used the Bonferroni procedure exclusively for multiple comparisons.

The Benjamini and Hochberg application of the false discovery rate (FDR) criterion can be described as follows. Let q be the number of significance tests made and let $P(1) \leq P(2) \leq \ldots \leq P(q)$ be the ordered significance levels of the q tests, from lowest to highest probability. Let α be the combined significance level desired, usually .05 for one-tailed tests (or .025 for two-tailed tests). The procedure compares $P(q)$ with α, $P(q\text{-}1)$ with $\alpha\,(q\text{-}1)/q$, . . ., $P(j)$ with α_j/q, stopping the comparisons with the first j such that $P(j) \leq \alpha_j/q$. All tests associated with $P(1)$, . . ., $P(j)$ are declared significant; all tests associated with $P(j+1)$, . . . , P_q are declared nonsignificant.

13.5.3 Comparing Proportions Within a Group

Certain analyses involved the comparison of proportions. One example was the comparison of the proportion of students who reported that a parent graduated from college to the proportion of students who indicated that their parents did not finish high school to determine which proportion was larger. There are other such proportions of interest in this example, such as the proportion of students with at least one parent graduating from high school but neither parent graduating from college. For these types of analyses, NAEP staff determined that the dependencies in the data could not be ignored.

Unlike the case for analyses of the type described in Section 13.5.1, the correlation between the proportion of students reporting a parent graduated from college and the proportion reporting that their parents did not finish high school is likely to be negative and large. For a particular sample of students, it is likely that the higher the proportion of students reporting “at least one parent graduated from college” is, the lower the proportion of students reporting “neither parent graduated from high school” will be. A negative dependence will result in underestimates of the standard error if the estimation is based on independence assumptions (as is the case for the procedures described in Section 13.5.1). Such underestimation can result in an unacceptably large number of “nonsignificant” differences being identified as significant.

The procedures of Section 13.5.1 were modified for analyses that involved comparisons of proportions within a group. The modification involved using a jackknife method for obtaining the standard error of the difference in dependent proportions. The standard error of the difference in proportions was obtained by first obtaining a separate estimate of the difference in question for each jackknife replicate (using the first plausible value only) then taking the standard deviation of the set of replicate estimates as the estimate. The procedures used for proportions within a group differed from the procedures of Section 13.5.1 only with respect to estimating the standard error of the difference; all other aspects of the procedures were identical.