

3.1 INTRODUCTION

Standard statistical procedures should not be applied to the NAEP data without modification because the special properties of the data affect the validity of conventional techniques of statistical inference. There are two reasons for this. First, to ensure accurate results, the relatively small samples of students selected for the NAEP assessments must be truly representative of the entire population and subgroups of this population. Therefore, a complex sampling scheme, rather than simple random sampling was used to collect NAEP data. Second, because scaling models were used to summarize performance in each subject area, measurement error must be taken into account when analyzing scale-score proficiency variables.

In the NAEP sampling scheme, students do not have an equal probability of being selected. Therefore, as in all complex surveys, each student has been assigned a sampling weight. The larger the probability of selection for students within a particular demographic group, the smaller the weights for those students will be. When computing descriptive statistics or conducting inferential procedures, one should weight the data for each student. *Performance of statistical analyses without weights can lead to misleading results.*

Another way in which the complex sample design used by NAEP differs from simple random sampling is that the NAEP sampling scheme involves the selection of clusters of students from the same school, as well as clusters of schools from urbanization, income, and minority strata (in the case of the state assessment) or from the same geographically defined primary sampling unit, or PSU (in the case of the national assessment). As a result, observations are not independent of one another as they are in a simple random sample. Therefore, *use of standard formulas for estimating the standard error of sample statistics such as means, proportions, or regression coefficients will result in values that are generally too small.* The standard error, which is a measure of the variability of a sample statistic, gives an indication of how well that statistic estimates the corresponding population value. It is used to conduct tests of statistical significance. If conventional simple random sampling formulas are used to compute standard errors, the number of statistically significant results will be substantially overestimated in most instances.

Another effect of the NAEP sampling scheme is a reduction of the effective degrees of freedom. In a simple random sample, the degrees of freedom of a variance estimate are based primarily on the number of subjects (although it also depends on the distribution of the variable under consideration). In the NAEP design, the degrees of freedom are a function of the number of clusters of schools (for the state assessment) or clusters of PSUs (for the national assessment), rather than the number of subjects. Therefore, *the standard formulas for obtaining degrees of freedom are not valid with the NAEP data.*

Proficiencies in content areas were summarized through item response theory (IRT) scaling models, but not in the way that these models are used in standard applications in which enough responses are available from each person to estimate his or her proficiency precisely. NAEP administers relatively few items to each respondent in order to track *population* levels of proficiency more efficiently. Because the data are not intended to estimate *individual* levels of proficiency, however, more complicated analyses are required.

The following sections outline the procedures used in NAEP to account for the special properties of the data. Section 3.2 discusses the use of weights to account for the differential sampling rates and certain other adjustments, such as for nonresponse. Section 3.3 discusses jackknife procedures that can be used to estimate sampling variability. Section 3.4 describes the “plausible values” that can be used to estimate population levels of proficiency in the subject areas, and shows how to use them in analyses. Section 3.5 suggests simpler approximations for the procedures described in 3.3 and 3.4, such as using design effects rather than the jackknife to estimate sampling variability. Although this procedure is less precise, it requires substantially less computation. We expect that the resulting degree of accuracy will be acceptable to most users of NAEP data.

3.2 USING WEIGHTS TO ACCOUNT FOR DIFFERENTIAL REPRESENTATION

The 1998 state and national assessments used complex sample designs to obtain the students who were assessed. The goal of the national design was to obtain a series of samples (for the various ages and grades) from which estimates of population and subpopulation characteristics could be obtained with reasonably high precision (low sampling variability) per unit of cost. The goal of the state assessment design was to obtain a sample of students for each jurisdiction from which estimates of population and subpopulation characteristics could be obtained with approximately equal precision for all jurisdictions.

To accomplish these goals, NAEP used multistage cluster sample designs in which the probabilities of selection of the clusters were proportional to measures of their size. To provide improved precision in the estimation of the characteristics of various subpopulations of interest, some schools (corresponding to areas with high concentrations of Black or Hispanic students) were deliberately sampled at approximately twice the normal rate to obtain larger samples of respondents from those subpopulations. Similarly, nonpublic schools were sampled at approximately three times the rate of public schools. The result of these differential probabilities of selection is a series of achieved samples, each containing proportionately more members of certain subgroups than there are in the population. Oversampling procedures were also used to ensure adequate sample sizes of SD/LEP students.

Appropriate estimation of population characteristics must take the sampling design into account. This is accomplished by assigning a weight to each respondent, where the weight properly accounts for the sample design and, in the case of the national assessment, reflects the appropriate proportional representation of the various types of individuals in the population. For instance, census data on the percentage of Hispanic students in the entire student population are used to assign a weight that adjusts the NAEP sample so it is representative of the nation. These weights also include adjustments for nonresponse and, in the case of the national assessment, adjustments (known as poststratification adjustments) designed to make sample estimates of certain subpopulation totals conform to external, more accurate, estimates. For the

present purpose, it is sufficient to note that these weights should be used for all analyses, whether exploratory or confirmatory.

The NAEP data files include a number of different samples from several populations. Each of these samples has its own set of weights to be used to produce estimates about the characteristics of the population addressed by the sample (the target population). The various samples, their target populations, and their weights are discussed in the following sections.

Table 3-1 provides a guide to the files, weights, and other criteria needed to carry out analyses appropriate to the assessment data.

3.2.1 Student Sample Weights

The target population for each of the national and state samples (one for each grade) consisted of all students who were in the specified grade and were deemed assessable by their school. Sampling weights were used to account for the fact that the probabilities of selection were not identical for all students. All population and subpopulation characteristics based on the assessment data used sampling weights in their estimation.

The overall student weight contained three components – a base weight, an adjustment for school nonparticipation, and an adjustment for student nonparticipation. The overall student weight, ORIGWT, has been scaled so that the sum of weights for appropriate subsamples estimates the total number of assessable students across the nation who are in the applicable grade.

An estimate of the proportion of students in the population who possess some characteristic can be obtained using ORIGWT as the ratio of the sum of the weights for the students with that characteristic, divided by the sum of the weights for all students sampled from that population. The numerator of the proportion is the estimated total number of students with that characteristic and denominator is the estimated population total. Estimated proportions can also be restricted to subpopulations. For example, the estimated proportion of all assessable fourth-grade students who are also in the Northeast region is

$$\frac{W_{TOT}(\text{Grade 4 and Northeast})}{W_{TOT}(\text{Grade 4})}$$

where $W_{TOT}(\text{Grade 4 and Northeast})$ is the sum of the weights (ORIGWT) of all students in the national assessment sample who are both in the fourth grade and the Northeast and where $W_{TOT}(\text{Grade 4})$ is the sum of the weights of all students in the national assessment sample who are in the fourth grade and from any region.

It is also clearly of interest to estimate the relative proportion of a population (say fourth-grade students) who could correctly respond to an assessment exercise. This proportion is estimated by the ratio

$$P = \frac{W_{TOT}(\text{Grade 4, answered item correctly})}{W_{TOT}(\text{Grade 4, presented the item})}$$

where the numerator is the sum of weights (ORIGWT) of all assessed students who are in the fourth grade and who responded to the item correctly and the denominator is the sum of weights

of all students who were in the fourth grade and were presented the item (i.e., reached the item, including those who reached it and left it blank).¹

This total is less than W_{TOT} (*Grade 4*) because not all students are presented every item, either as a result of the spiral design or as a result of not reaching the item. However, the sample of assessed fourth-grade students who had an opportunity to respond to the item (which includes those who did not reach the item) is itself a representative sample of the entire population of assessable fourth-grade students.

3.2.2 Excluded Student Sample Weights

The excluded students from the national and state reading assessments are samples from the population of all grade-eligible students who would not be assessable. For each national and state assessment sample, the excluded student data were combined with the assessed student data, and the weights were adjusted appropriately. The weight to be used for analyses of the excluded students is therefore the same as for the assessed students: the student full-sample weight (ORIGWT). Weighted analyses can be conducted on these data to estimate characteristics of the population of excluded students, analogous to the analyses for the assessed students.

The excluded students for the 1998 samples are included in the data files of assessed students. To select excluded students, use cases where XRPTSAMP=1.

3.2.3 School and Teacher Sample Weights

The 1998 assessment collected questionnaire data from principals and other administrators about aspects of the schools attended by the assessed students. Analyses of these data using the weights described above will produce results that are focused on students (e.g., *What percentage of students attend schools in which mathematics is identified as a special priority?*). This type of analysis requires first merging the school data files with the student data files (see section 5.5 in Part Five). For the school questionnaire data, it is also possible to conduct school-level analyses (e.g., *In what proportion of schools do teachers with bachelor's degrees in mathematics teach mathematics classes?*). The school weights (SRMWTF on the school files) should be used for these purposes.

As described in Part Two, the teacher questionnaire was administered to the reading teachers of fourth-grade and eighth-grade students participating in the assessment. All such teachers were included in the sample, and were asked to complete a questionnaire concerning themselves and their teaching practices, with specific references to each individual class period containing a student included in the assessment.

The purpose of drawing these samples was not to estimate the attributes of the teacher population, but to estimate the number (proportion) of students whose teachers had various attributes and to correlate student characteristics and performance with the characteristics of their teachers. Because the NAEP samples were not selected to contain representative samples of teachers, analogous teacher weights are not provided. Analyses of the teacher questionnaire data should be restricted to student-level analyses using the student weights.

¹ Missing responses after the last observed response are considered "not reached," and are treated as if they had not been presented to the respondent.

Table 3-1
Summary of 1998 NAEP Reading Samples and Their Use in Analyses

Samples	Filenames*	Overall Weight	Replicate Weights	Selection Criteria	Use
National Reading Student & Teacher Samples	RNT1STUD.DAT RNT2STUD.DAT RNT3STUD.DAT	ORIGWT	SRWT01-SRWT62	RPTSAMP=1	1) National student-level analyses of reading proficiency by scale or composite 2) At grades 4 and 8, national student-level analyses of reading teacher data
				RPTSAMP=1 and SCHTYPE=1	1) National public-school student-level analyses of reading proficiency by scale or composite (use for comparing nation to states) 2) At grades 4 and 8, national public-school student-level analyses of reading teacher data
				XRPTSAMP=1	Estimates of numbers and characteristics of students excluded from the national reading assessment
National Reading School Samples	RNT1SCHL.DAT RNT2SCHL.DAT RNT3SCHL.DAT	SRMWTF	SRMWT01-SRMWT62	None	National school-level estimates of school questionnaire data
				SSCHTY7=1	National public-school school-level estimates of school questionnaire data
State Reading Student & Teacher Samples	RST1ST__.DAT RST2ST__.DAT	ORIGWT	SRWT01-SRWT62	RPTSAMP=1	1) Student-level analyses of reading proficiency by scale or composite within a state 2) Student-level analyses of reading teacher data within state
				RPTSAMP=1 and SCHTYPE=1	1) Public-school student-level analyses of reading proficiency by scale or composite within a state, across states, or comparing state to nation 2) Public-school student-level analyses of reading teacher data within a state, across states, or comparing state to nation
				XRPTSAMP=1	Estimates of numbers and characteristics of students excluded from the state reading assessment
State Reading School Samples	RST1SC__.DAT RST2SC__.DAT	SRMWTF	SRMWT01-SRMWT62	None	School-level estimates of school questionnaire data within a state.
				SSCHTY7=1	Public-school school-level estimates of school questionnaire data within a state

*File naming convention: RNT=Reading NaTional data, followed by a single-digit number indicating grade assessed (1=grade 4, 2=grade 8, 3= grade 12), the next four characters are either "STUD" for student or "SCHL" for school (e.g., RNT1STUD.DAT contains reading national grade 4 student data). RST=Reading STate data, followed by a single-digit number indicating grade assessed, the next four characters are either "ST__" for student, or "SC__" for school, followed by the appropriate state abbreviation (e.g., RST1STAL.DAT contains the reading state data for grade 4 in Alabama).

Data collected from the teacher questionnaires are appended to the appropriate student records in the secondary-use student data files. Adding the teacher data to the appropriate student file allows correct and efficient analysis of the teacher/student data without requiring users to match data from separate files.

3.2.4 Replicate Weights

In addition to overall estimation weights, a set of replicate weights is provided for each student, excluded student, and school. These replicate weights are used in calculating the sampling errors of estimates obtained from the data, using the jackknife repeated replication method. The methods of deriving these weights were aimed at reflecting the features of the sample design appropriately, so that when the jackknife variance estimation procedure is implemented as intended, approximately unbiased estimates of sampling variance result.

Replication estimates the variance of the full sample. This process involves repeatedly selecting portions of the sample to calculate the statistic of interest. The estimates that result are called replicate estimates. The variability among these calculated quantities is used to obtain the full sample variance. The process of forming these replicate estimates involves first dividing the sample elements among a set of replicate groups, then using the pattern of replicate groups in a systematic fashion to apply replicate weights to the file.

The student replicate weights SRWT01-62 correspond to the overall weight ORIGWT. They are used for estimating the sampling errors of estimates derived using the full sample weights. These weights are designed to reflect the method of sampling schools, and account for the type of stratification used and whether or not the student's school was included in the sample with certainty. The method of sampling students within schools is also reflected, implicitly in the case of noncertainty schools and explicitly for schools included with certainty. These replicate weights also reflect the impact on sampling errors of the school- and student-level nonresponse adjustments applied to the overall estimation weights.

At the school level, the replicate weights on the school data files should be used to estimate the variance for population estimates obtained using the school weight. The replicate weights for schools in the various assessment samples are shown in Table 3-1.

The appropriate replicate weights for excluded students are also shown in Table 3-1.

3.3 PROCEDURES USED BY NAEP TO ESTIMATE SAMPLING VARIABILITY (JACKKNIFING)

This section describes how the sampling variability of statistics based on the NAEP data can be estimated. The jackknife variance estimator described below gives fairly precise estimates of the total sampling error for population estimates derived from NAEP student and school data, and for conducting multivariate analyses. To aid secondary users who have fewer resources than those available for the NAEP reports, section 3.5 provides a less expensive approximation for estimating sampling variances.

A major source of uncertainty in the estimation of the population value for 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 likely 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 proficiencies. For the 1998 assessment, proficiencies 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 proficiency. In NAEP, each individual responds to relatively few items so that individual proficiency values are not well determined. Consequently, the variance of any statistic based on proficiency values has a component due to the imprecision in the measurement of the proficiencies of the sampled individuals in addition to a component measuring sampling variability.

The NAEP samples are obtained via a stratified multistage probability sampling design that includes provisions for sampling certain subpopulations at higher rates. Additional characteristics of the sample include adjustments for both nonresponse and poststratification. The resulting samples have different statistical characteristics than those of a simple random sample. In particular, because of the effects of cluster selection (students within schools, schools within PSUs) and nonresponse and other weighting adjustments, observations made on different students cannot be assumed to be independent of each other. Furthermore, to account for the differential probabilities of selection and the various sample weighting adjustments, each student has an associated sampling weight that must be used in the computation of any statistic and is itself subject to sampling variability.

Treatment of the data as a simple random sample, with disregard for the special characteristics of the NAEP sample design, will produce underestimates of the true sampling variability. A procedure known as jackknifing is suitable for estimating sampling errors from such a complex design. This procedure has a number of properties that make it particularly suited to the analysis of NAEP data:

- ◆ It provides unbiased estimates of the sampling error arising from the complex sample selection procedure for linear estimates such as simple totals and means, and does so approximately for more complex estimates.
- ◆ It reflects the component of sampling error introduced by the use of weighting factors, such as nonresponse adjustments, that are dependent on the sample data actually obtained.
- ◆ It can be adapted readily to the estimation of sampling errors for parameters estimated using statistical modeling procedures, as well as for tabulation estimates such as totals and means.
- ◆ Once appropriate weights are derived and attached to each record, jackknifing can be used to estimate sampling errors. A single set of replicate weights is required for all tabulations and model parameter estimates that may be needed.

Here the method of applying the jackknife procedure involves first defining pairs (or occasionally triples) of replicate groups. A replicate group consists of a single PSU, a pair of PSUs, or (for the large certainty PSUs) schools within a PSU. The replicate groups were paired in accordance with the sample design. The pairing was done independent of performance information obtained from the sample. For the 1998 assessment, Westat defined 62 such pairs.

Components of the sampling variability of an estimate are each estimated as the squared difference between the value of the statistic for the complete sample and a pseudoreplicate formed by recomputing the statistic on a specially constructed pseudodataset. This pseudodataset is created from the original dataset by eliminating one member of a pair and replacing it with a copy of the remaining unit or units in the pair. For computational purposes, a pseudoreplicate associated with a given pair is the original dataset with a different set of weights (referred to as the student replicate weights SRWT01 through SRWT62 on the data files, where SRWT i is for the i^{th} pair). This set of weights allows measurement of the total effect of replacing one member of the pair with a copy of the other(s), including adjustments for nonresponse and poststratification. The i^{th} pseudoreplicate for a given statistic is obtained by recalculating the statistic using the weights SRWT i instead of the original sampling weights.

As a specific example of the use of the student replicate weights, let $t(\underline{y}, \underline{w})$ be any statistic that is a function of the sample responses y and the weights w that estimates population value T . For example, t could be a weighted mean, a weighted percent-correct point, or a weighted regression coefficient. The $t(\underline{y}, \underline{w})$, computed with the sampling weights (ORIGWT on the data files) is the appropriate sample estimate of T . To estimate $\hat{V}ar(t)$, the sampling variance for this statistic, proceed in the following manner:

- 1) For each of the 62 pairs of first-stage units, compute the associated pseudoreplicate for the statistic. For the i^{th} pair, this is

$$t_i = t(\underline{y}, \underline{SRWT}_i),$$

which is the statistic t recalculated by using SRWT i instead of the original sampling weights.

- 2) The estimated sample variance of t is

$$\hat{V}ar(t) = \sum_{i=1}^{62} (t_i - t)^2.$$

We refer to this estimation technique as the multiweight jackknife approach. Part Five provides SPSS and SAS code for carrying out the above in the special case of a weighted mean.

A similar procedure is followed to estimate the sampling variability for school-level statistics based on the school data and using the school weights.

Replicate weights are provided on the data files for each of the 1998 assessment samples.

Table 3-1 provides a summary of the samples, sampling weights, and replicate weights and their use in analyses of 1998 NAEP data.

As a very simple example of how the jackknife variance estimate is computed, consider the following table (Table 3-2) designed to demonstrate the steps. Although the full set of NAEP data consists of thousands of observations and 62 student replicate weights, for the example we will consider a dataset with eight observations and two student replicate weights. Furthermore, the weights have been simplified for clarity.

Table 3-2
Example Dataset to Demonstrate the Jackknife

First-Stage Unit	Pair	Pair Member	Y	ORIGWT	SRWT01	SRWT02
1	1	1	5	10	20	10
	1	1	4	9	18	9
2	1	2	6	12	0	12
	1	2	3	8	0	8
3	2	1	8	4	4	8
	2	1	9	6	6	12
4	2	2	7	5	5	0
	2	2	10	4	4	0

In the example dataset there are four first-stage units, 1 through 4, each consisting of two of the eight observations. The first-stage units are divided into two pairs, as identified by the column labeled "Pair." Within each of those pairs, one first-stage unit is designated as the first member of the pair (1) while the other is designated as the second (2) (labeled "Pair Member" in the table). A detailed discussion of the pairing procedure can be found in the forthcoming *1998 Technical Report* (Allen, et al., in press).

The statistic of interest is the weighted average of the responses Y , using the weights ORIGWT, and is equal to

$$t = \text{NUM}_1 / \text{DEN}_1 = 5.914$$

where

$$\text{NUM}_1 = 10 \times 5 + 9 \times 4 + 12 \times 6 + 8 \times 3 + 4 \times 8 + 6 \times 9 + 5 \times 7 + 4 \times 10 = 343$$

is the weighted sum of the responses, and

$$\text{DEN}_1 = 10 + 9 + 12 + 8 + 4 + 6 + 5 + 4 = 58$$

is the sum of the weights ORIGWT.

The first pseudoreplicate of the statistic t is the weighted mean recomputed using the SRWT01 as the weights and is

$$t_1 = \text{NUM}_1 / \text{DEN}_1 = 5.842$$

where

$$\text{NUM}_1 = 20 \times 5 + 18 \times 4 + 0 \times 6 + 0 \times 3 + 4 \times 8 + 6 \times 9 + 5 \times 7 + 4 \times 10 = 333$$

and

$$\text{DEN}_1 = 20 + 18 + 0 + 0 + 4 + 6 + 5 + 4 = 57.$$

Similarly, $t_2 = 354/59 = 6$ is the weighted mean computed using SRWT02 as the weights. The jackknife variance estimate is then

$$\begin{aligned} \hat{V}ar(t) &= \sum_{i=1}^2 (t_i - t)^2 = (t_1 - t)^2 + (t_2 - t)^2 \\ &= (-0.072)^2 + (0.086)^2 = 0.1258 \end{aligned}$$

and the jackknife standard error of t is .112, the square root of the variance.

3.3.1 Degrees of Freedom of the Jackknife Variance Estimate

The effective number of degrees of freedom of the variance estimate $\hat{V}ar(t)$ will be at most equal to the number of pairs used in forming the pseudoreplicates. The number of degrees of freedom in sampling from normally distributed variates with uniform variances is sufficient information to indicate the variability of the variance estimate, and is equal to the number of independent pieces of information used to generate the variance. For each assessment sample, the pieces of information are the 62 squared differences $(t_i - t)^2$, each supplying at most one degree of freedom, regardless of how many individuals were sampled within any replicate groups.

The effective number of degrees of freedom of the sample variance estimator can be less than the number of pairs (62) if the differences are not normally distributed or if some of the squared differences $(t_i - t)^2$ are markedly different in magnitude than others. An extreme case of the latter is when one or more of the t_i are identical to t , so that $(t_i - t)^2 = 0$. This may happen, for example, when the statistic t is a mean for a subgroup, such as a type of location, and no members of that subgroup come from the pair i . Such a pair contributes zero to the effective number of degrees of freedom of the variance estimate.

An estimate of the effective number of degrees of freedom for $\hat{V}ar(t)$ comes from an approximation due to Satterthwaite (1941). (See Cochran, 1977, p. 96, for a discussion.)

If the t_i are normally distributed, the effective number of degrees of freedom using this approximation is

$$df_{eff} = \frac{\left[\sum_{i=1}^K (t_i - t)^2 \right]^2}{\sum_{i=1}^K (t_i - t)^4}$$

where K is the number of pairs used.

3.4 PROCEDURES USED BY NAEP TO HANDLE IMPRECISION OF INDIVIDUAL MEASUREMENT

Jackknifing provides a reasonable estimate of uncertainty due to the sampling of respondents when the variable of interest is observed without error from every respondent. Population percents correct for cognitive items meet this requirement, but scale-score proficiency values do not. The item response theory (IRT) models used to summarize performance in a subject area or subarea posit an unobservable proficiency variable θ to summarize performance on the items in that area. The fact that θ values are not observed even for the respondents in the sample requires additional statistical machinery to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. To this end, we have adapted Rubin's (1987) "multiple imputations" procedures for missing data to the context of latent variable models to produce the "plausible values" that appear in the NAEP secondary-use data files.

The essential idea of plausible values methodology is that even though we do not observe the θ value of respondent i , we do observe other kinds of variables that are related to it: x_i , the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i , the respondent's answers to demographic and background variables. Suppose we would like to draw inferences about a number $T(\theta, Y)$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\theta, y)$, where $(\theta, y) \equiv (\theta_1, y_1, \dots, \theta_N, y_N)$, and that we could estimate the variance in t around T due to sampling respondents by the function $U(\theta, y)$. Given that observations consist of (x_i, y_i) rather than (θ_i, y_i) , we can approximate t by its expected value conditional on (x, y) , or

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

It is possible to approximate t^* with random draws from the conditional distributions $p(\theta | x, y)$. Let $\hat{\theta}_m$ be the m^{th} such vector of "plausible values." It is a plausible representation of what the true θ might have been, had we been able to observe it. The following steps describe how an estimate of a scalar statistic $t(\theta, y)$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Note: five sets are provided on the data files for each subject area or subarea analyzed by these procedures.)

- 1) Using each set of plausible values $\hat{\theta}_m$ in turn, evaluate t as if the plausible values were true values of θ . Denote the results \hat{t}_m , for $m=1, \dots, M$.
- 2) Using the multiple weight jackknife approach, compute the estimated sampling variance of \hat{t}_m , denoting the result as U_m .

- 3) The final estimate of t is

$$t^* = \sum_{m=1}^M \hat{t}_m / M .$$

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

$$U^* = \sum_{m=1}^M U_m / M .$$

- 5) Compute the variance among the M estimates \hat{t}_m , to approximate uncertainty due to not observing θ values from respondents (sampling error):

$$B = \sum_{m=1}^M (\hat{t}_m - t^*)^2 / (M-1) .$$

- 6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1}) B .$$

Note: NAEP reports use a single jackknife estimate U_m in place of the average of five, as would be required for U^ ; see section 3.5.*

Suppose that the statistic $[t(\theta, y) - T]/U^2$ would follow a t -distribution with d degrees of freedom. Then the distribution of $(t^* - T)/V^2$ is also approximately t -distributed, with degrees of freedom given by

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

where f is the proportion of total variance due to not observing θ values:

$$f = (1 + M^{-1}) B / V .$$

When B is small relative to U^* , and d is large, a normal approximation suffices. This is the case with main NAEP reporting variables, and the normal approximation is routinely applied to flag “significant” results.

For k -dimensional t , such as the k coefficients in a multiple regression analysis, each U_m and U^* are covariance matrices, and B is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity

$$(T - t^*) V^{-1} (T - t^*)$$

is approximately F -distributed, with degrees of freedom equal to k and v , with v defined as above but with a matrix generalization of f :

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

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

Computation of statistics t^* involving the plausible values and categories of variables included in the conditioning variables y yields consistent estimates of the corresponding population values T . Statistics involving background variables y that were not conditioned on are *subject to biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variable(s) to the variables that were conditioned on. The direction of the bias is typically to underestimate the effect of nonconditioned variables.*²

For a given statistic t^* involving one or more nonconditioned background variables, the magnitude of the bias is related to the extent to which observed responses account for the latent variable θ , and the degree to which the nonconditional background variables are explained by conditioning background variables.

3.5 APPROXIMATIONS

A jackknife estimate of the variability of a statistic based on one or more observed NAEP variables in the 1998 samples requires computing the statistic 63 times. Estimating the variability for a statistic involving a scale-score could require computing the statistic as many as 315 times, including 63 runs to obtain a variance estimate for each of five sets of plausible values. Because the cost of the full procedure may well prove prohibitive in many studies, approximate procedures that produce reasonable estimates at lower costs are provided below. Section 3.5.1 gives approximations for sampling variation; 3.5.2 gives approximations for variation due to measurement error associated with scale-scores; 3.5.3 discusses strategies for combining the suggestions in 3.5.1 and 3.5.2.

3.5.1 Approximations for Sampling Variability

The major computational load in calculating uncertainty measures for any statistic exists in the computation of the uncertainty due to sampling variability. As noted in the last section, a jackknife estimate of the variability of a statistic based on one or more observed NAEP variables in the 1998

² For details, see section 10.3.5 of Implementing the New Design: The NAEP 1983-84 Technical Report (Beaton, 1987) section 8.4.3 of Expanding the New Design: The NAEP 1985-86 Technical Report (Beaton, 1988), and Mislevy, 1991.

assessment samples requires computing the statistic 63 times. This section describes a less computationally intensive approximation to sampling variability of any statistic.

As indicated in section 3.3, it is inappropriate to estimate the sampling variability of any statistic based on the NAEP database by using simple random sampling formulas. These formulas, which are the ones used by most standard statistical software such as SPSS and SAS, will produce variance estimates that are generally much smaller than is warranted by the sample design.

It may be possible to account approximately for the effects of the sample design by using an inflation factor, the design effect, 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 simple random sampling variance estimate based on the same number of elements. The design effect may be used to adjust error estimates based on simple random sampling assumptions to account approximately for the effect of the design. In practice, this is often accomplished by dividing the total sample size by the design effect and using this effective sample size in the computation of errors. Note that 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 clustering effects occurring among sampled elements and the effects of any variable weights resulting from variable overall sampling fractions.

On the basis of empirical results and theoretic considerations, Kish and Frankel (1974) have developed several conjectures about design effects:

- 1) Generally, the design effects for complex statistics from complex samples are greater than 1, causing variances based on simple random sampling assumptions to tend to be underestimated.
- 2) The design effects for regression coefficients tend to be smaller than the corresponding design effects for means of the same variables. Hence, these latter estimates, which are more easily computed, tend to overestimate the design effects of complex statistics. For correlation coefficients and partial correlation coefficients, the design effect for the mean should be used (Skinner, Holt, & Smith, 1989, p. 70).
- 3) The size of the design effects of complex statistics tends to parallel those of means; variables with a high design effect of the mean also tend to have high design effects for complex statistics involving those variables.

To incorporate the design effect idea in a statistical analysis, proceed in the following manner:

- 1) For a given class of statistics (e.g., means, proportions, regression coefficients), compute the jackknife variance described in section 3.3.1 for a number of cases. The cases should cover the range of situations for which the approximation is to be used. If various subpopulations are to be considered, it is important to have information on the relative variability within each subgroup. This is especially important if certain subgroups are more highly clustered in the sample.

- 2) For the identical cases, compute the simple random sampling variance given the elements in the sample. To account properly for the difference between the number of individuals being sampled and the total of the sampling weights, the weights should be scaled so that their sum equals the sample size.
- 3) For each case, compute the design effect where the design effect for case j is

$$deff_j = Var_{JK}(t_j) / Var_{CON}(t_j) ,$$

the ratio of the jackknife variance estimate of the statistic to its simple random sampling variance estimate.

- 4) If the design effects for the various cases are tolerably similar, choose an overall composite design effect. If the design effects for certain subgroups appear to cluster around a markedly different value from the remaining cases, treat those subgroups separately.
- 5) In the case that a consistent overall design effect has been found:
 - a) Rescale the weight of each individual so that the sum of the scaled weights is equal to the effective sample size

$$N_{eff} = \frac{\text{sample size}}{\text{design effect}}$$

(that is, multiply each weight by N_{eff}/W_{TOT} , where W_{TOT} is the sum of the original weights).

- b) Conduct a traditional weighted analysis using these scaled weights.

The number of degrees of freedom for any variance estimates obtained by using this approach is still, at best 62, the number of pairs, as it was for the jackknife. Accordingly, tests of significance produced by standard programs (which will use the effective sample size minus the number of parameters for error degrees of freedom) should be interpreted with extreme caution because they are likely to be too liberal. Significance and inferential procedures will be improved if based on the smaller error degrees of freedom that were indicated for the degrees of freedom for the jackknife variance estimate (section 3.3.1), although this estimate of effective degrees of freedom will also be an overstatement, since the paired variables will not be normally distributed.

3.5.2 Approximations for Measurement Error Variability

A second method of reducing costs applies to statistics that involve scale-score proficiency values: using fewer runs on plausible value sets. A statistic t^* based on a single set of plausible values has the same expectation as the average of five, but cannot take into account the uncertainty caused by the fact that θ is unobserved. Compared to using all five sets of plausible values, using at least two but fewer than five sets to evaluate a statistic allows one to account for this component of uncertainty and reduce costs at the same time. One merely applies the formulas given in section 3.4

with $M=2, 3, \text{ or } 4$, as appropriate. It may be seen that the resulting decrease in computation is accompanied by an increase in total variance associated with t^* , but one that may be worth the price.

Note: It is not recommended to compute the average of the five plausible values associated with each respondent, then analyze these averages. This procedure does not generally give the correct value of a statistic.

3.5.3 Approximating Both Sampling and Measurement Variability

Full implementation of the procedures for estimating the variability of a statistic involving a scale-score variable requires 315 runs. Combining the approximations suggested above in various ways allows the researcher to trade off precision and cost in a manner that suits his or her needs. Some options are discussed below.

- 1) *Full implementation (315 runs)*. This option is the most costly, but the most precise. Each estimate of a statistic and each jackknife estimate of its sampling variance is calculated on all five sets of plausible values.
- 2) *Estimate based on five sets of plausible values, jackknife based on one set of plausible values (67 runs)*. This option involves computing the statistic t^* exactly as described in section 3.4, but basing its variance estimate on the sum of B (the variance of five \hat{t}_m estimates) and one U_m value (rather than the average of five). This is the option routinely used by NAEP in its own reports. It gives the same point estimate of T as the full implementation, but the variance estimate, while still consistent, is less precise. Using the jackknife as opposed to a design effect accounts for 62 of the runs, but allows for differential impacts of the respondent-sampling design upon the variability of different statistics.
- 3) *Estimate based on five sets of plausible values, design effect for sampling variance (5 runs, assuming a design effect has already been estimated)*. This option gives the same point estimate of T as options 1) and 2), but a less precise estimate of its variability. It is obtained by computing t^* and $V(t^*)$ just as described in section 3.4, but with each U_m value obtained by boosting the SRS sampling variance estimates in accordance with a design effect as described in section 3.5.1. Note that additional initial runs will be needed to estimate the design effect.
- 4) *Estimate based on M sets of plausible values, where $1 < M < 5$; design effect for sampling variance (M runs)*. The point estimates provided by this option now differ from those in previous options. They have the same expectations as those described above, but now the point estimates themselves, rather than just the estimates of their variability, are less precise. By using at least two sets of plausible values, however, the researcher ensures that both the sampling and the measurement components of variability are taken into account. This option is attractive for researchers who have very limited resources.
- 5) *Estimate based on one set of plausible values, design effect for sampling variance (1 run)*. The point estimate obtained here has the same expected value as those described above, but is again less precise. Measurement variability cannot be estimated with only one set of plausible values, and statements of variability or significance tests based on sampling

variability only are incorrect because they underestimate variability. The degree of underestimation depends on the statistic being computed. For population or subpopulation averages of proficiency on background variables included in all booklets, the degree of underestimation of variability is roughly 20 percent (Rubin, 1987, Table 4.1). For statistics that are more complex or involve background variables that appear on only a subset of booklets, the underestimation can easily exceed 50 percent. This option is not recommended for such statistics.

Again, a strategy that *should not* be considered deserves repeated emphasis: *Computing the average of the five plausible values associated with each respondent, then analyzing these averages, does not generally give the correct value of a statistic.*

3.6 ADDITIONAL SOURCES OF ERROR

In addition to errors due to sampling and imprecision of individual measurement, NAEP results are also subject to other kinds of errors, including the effects of necessarily imperfect adjustment for student and school nonresponse and other largely unknowable effects associated with the particular instrumentation and data collection methods used. Nonsampling errors can be attributed to a number of sources: inability to obtain complete information about all selected students in all selected schools in the sample (some students or schools refused to participate, or students participated but answered only certain items); ambiguous definitions; differences in interpreting questions; inability or unwillingness to give correct information; mistakes in recording, coding, or scoring data; and other errors of collecting, processing, sampling, and estimating missing data. The extent of nonsampling errors is difficult to estimate. By their nature, the impacts of such errors cannot be reflected in the data-based estimates of uncertainty.

Users of NAEP data should also be aware that there are additional components of variance, due to the statistical nature of the scaling and linking process, that are not included in the various estimation procedures discussed in sections 3.3 and 3.4. In NAEP, as in other applications of IRT, item parameters are unknown; estimates must be used. Research is underway on how uncertainty associated with item parameter estimates affects the estimation of proficiency distributions (see, e.g., Tsutakawa & Johnson, 1990). The estimation error associated with scale linking represents another source of uncertainty. Some preliminary investigations into estimating the uncertainty associated with scale linking have been carried out by Sheehan and Mislevy (1988) and Johnson, Mislevy, and Zwick (1990). At present, standard errors for NAEP results reflect only the estimation due to sampling of students and due to imprecision of individual measurement. Research is underway to determine mechanisms for including other sources of uncertainty into the variance estimation procedures.

3.7 A NOTE CONCERNING MULTIPLE COMPARISONS

If many statistical tests are conducted at one time, it is likely that significance tests will overstate the degree of statistical significance of the results. In the preceding sections, we noted that because of the design of the NAEP sample, conventional significance tests will overstate significance because they fail to consider the effects of clustering. In contrast, the problem of multiple comparisons noted here is independent of sample design; it arises even if one uses the appropriate statistical tests described

previously. The problem arises because the more statistical tests are calculated, the more likely it becomes that one will find a “significant” finding because of chance variation. In other words, the chance of a type I error – a spurious “significant” finding – rises with the number of tests conducted.

In sets of confidence intervals, statistical theory indicates that the 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., 0.05), adjustments called “multiple comparison procedures” (Miller, 1966) must be made to the methods described in the previous sections. 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, et al., 1994). Therefore, the FDR procedure is more suitable for multiple comparisons in NAEP than other procedures. A detailed description of the FDR procedure appears in the forthcoming *NAEP 1998 Technical Report* (Allen, et al., in press).

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 that state’s mean in the previous assessment. All other multiple comparisons for the 1996 assessment used the Bonferroni procedure. The 1994 reading assessment used the Bonferroni procedure exclusively for multiple comparisons.