

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Slide 1 of 36

Title Slide: NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Slide 2 of 36

As discussed in the common modules, when analyzing data from complex sample surveys, certain procedures must be used to ensure that estimates made from the data are representative of the population and that hypothesis tests are accurate. This module describes the key design features of NAEP. It describes the sampling weights that must be applied when using NAEP data to ensure that estimates are representative of the target population. This module also describes the appropriate procedures for calculating standard errors using NAEP data. Finally, this module describes the implications of the NAEP assessment design, IRT models, and plausible values on analysis of NAEP data. For information about weighting and calculating appropriate standard errors more generally, please view the relevant common modules, if you have not already done so.

Throughout this module, underlined blue screen text indicates a link to additional resources.

Slide 3 of 36

NAEP is the largest continuing and nationally representative assessment of what our nation's students know and can do in various subjects. NAEP provides a common measure of how students across the nation are performing. NAEP is a survey designed to produce national, state, and large urban school district results. NAEP is not a testing program for individual students or schools.

Slide 4 of 36

Main NAEP is administered to samples of students in grades 4, 8, and 12 in public and private schools. The Main NAEP assessment component yields national, state, and selected urban district results in reading and mathematics every two years, in odd numbered years; and in science and writing every four years, in even numbered years.

Other NAEP subjects are assessed on a voluntary basis at regularly scheduled intervals to provide national-only results in U.S. history, geography, civics, economics, and the arts.

Slide 5 of 36

NAEP Long-Term Trend, or LTT, assessments are designed to provide information on the changes in the performance of the nation's 9-, 13-, and 17-year-old students. Since the LTT assessments are administered at the national level only, there are no state or district level results.

Because the long-term trend program uses substantially the same assessments decade after decade, it has measured students' performance in mathematics and reading since

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

1971. The most recent LTT assessments were administered in 2012. The next administration will take place in 2020.

Slide 6 of 36

NAEP reports results for the Nation (through the Main NAEP National Assessment and/or the NAEP Long-Term Trend assessment), States (through the state assessment), and for large urban school districts (through the Trial Urban District Assessment, or TUDA, assessment).

Slide 7 of 36

The assessment design of NAEP works together with the sampling design to ensure representative and valid measures of student achievement across various subjects for each of the reporting components (the Nation, states, and TUDA districts). The NAEP assessment design ensures that valid estimates of what students know and can do across various subjects are produced while minimizing burden on students, schools, and costs associated with assessment administration. The NAEP sampling design ensures that the students and schools selected for participation in NAEP are representative of the groups for which results are reported.

This module presents general information about common design features across NAEP administrations. It is important to note that the assessment needs of any given administration typically vary. Documentation should be reviewed carefully to understand the specific details regarding assessment design and sampling procedures relevant to analyses of interest.

Slide 8 of 36

NAEP uses an assessment design that ensures the response burden on students and schools is minimized while producing valid estimates of achievement. NAEP assessment booklets are constructed such that no student responds to all of the items in the item pool. A participating student, takes only a portion of the complete set of cognitive items developed for any assessment; and takes only one subject area assessment.

It is important to note that there are not separate subject area assessments for the Main NAEP National, State, and TUDA assessments. There are separate reading and mathematics assessments used as part of LTT NAEP that are different from the Main NAEP reading and mathematics assessments.

NAEP administers the assessment to samples of students that are representative of the nation, states, and TUDA districts from which results are reported.

Slide 9 of 36

The NAEP Authorization Act, SEC 303 (b) (2) (A) requires “a random sampling process” and “data that are representative on a national and regional basis.” To meet these requirements, sampling methods are utilized that balance the need for valid inferences with burden on schools and students as well as data collection costs.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 10 of 36**

NAEP has multiple target populations for which results are reported (the Nation, states, and TUDA districts). Accordingly, the NAEP sampling design consists of multiple samples that either stand alone, or can be combined, to provide estimates of what students know and can do for each reporting component. To make valid inferences about students in the target populations, two conditions must be satisfied. First, the probability of selection must be known. Second, every member of the target populations has to have at least some chance of selection into the sample.

Slide 11 of 36

The target populations for the Main NAEP include students enrolled in grades 4, 8, and 12 attending public and private elementary and secondary schools, Bureau of Indian Education (BIE), and Department of Defense Education Activity (DoDEA) schools physically located in the 50 states and the District of Columbia. Students physically located in Puerto Rico are included in NAEP's target population for mathematics only.

The target population for the 2012 NAEP Long-Term Trend assessments consisted of 9-, 13-, and 17-year-old students enrolled in public and private schools nationwide.

Slide 12 of 36

The target populations for NAEP's national assessments do not include students enrolled in ungraded classrooms, students schooled at home, by correspondence, or online. Students enrolled in schools serving youth in hospitals or correctional facilities, and students with disabilities (SDs) and English language learners (ELLs) who cannot be assessed, even with accommodations are also not included in NAEP's target population.

Slide 13 of 36

To meet the requirement that, 'data are representative on a national and regional basis,' representative samples of students must be drawn from which valid estimates of what students know and can do will be reported across three reporting components (the Nation, states, and TUDA districts).

Accordingly, multiple samples are designed and used in NAEP. These samples can be thought of in terms of jurisdictions. A jurisdiction is any government-defined geographic area that is sampled in the NAEP assessment (for example, the United States, a state, the District of Columbia, Puerto Rico, an urban district, or the Department of Defense Education Activity schools (DoDEA)).

Slide 14 of 36

Each assessment cycle, a sample of students in designated grades (or ages in LTT) within both public and private schools throughout the United States (and sometimes specified territories and possessions) is selected for assessment. The multi-stage complex sampling design varies based on the NAEP assessment year. In state assessment years, samples of public and private schools and their students in each

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

state are drawn that are large enough to support state-level estimates. In national-only assessment years, about 1,000 Primary Sampling Units, or PSUs, are created, which consist of counties or groups of counties that are constructed based on geographic factors.

In all cases, the selection process utilizes a probability sample design in which every school and student has a chance to be selected, and standard errors can be calculated for the derived estimates.

Slide 15 of 36

For years with state assessments (e.g., mathematics 2013), the Common Core of Data (CCD) contains a comprehensive list of public schools, which is used as the sampling frame. The CCD also contains additional information about each school, which is then used for stratification (i.e., to create strata) based on school characteristics. A sample of schools is then drawn within each jurisdiction (the 50 individual states, the District of Columbia (DC), Puerto Rico, BIE, DoDEA, and 21 TUDA districts) and then a sample of students is drawn from within each school using a procedure that we will discuss later in this module.

A separate sample of Private schools is drawn from a list that has been stratified by school characteristics prepared from the Private School Universe Survey (PSS).

Slide 16 of 36

In years when the NAEP samples are intended only to provide representation at the national level and not for each individual state, the public and private school selection process is somewhat different. For years with national-only assessments, about 1,000 Primary Sampling Units, or PSUs, are created, which consist of counties or groups of counties that are constructed based on geographic factors. Of these, 50 to 100 are sampled. Once these PSUs have been selected, lists of schools are generated using the CCD and PSS. Schools are sampled within PSUs using a procedure that we will discuss later in this module.

Slide 17 of 36

Schools are sampled with probability proportional to size of enrollment so that all students have a similar chance of selection overall, regardless of the size of school that they attend. Larger schools are more likely to be selected than smaller schools. Students within larger schools are less likely to be selected than in smaller schools. This assures that there are a sufficient number of students sampled from schools of varying sizes to allow valid estimates and comparisons to be made. It is important to note that NAEP does not sample schools at the same rate within each stratum, meaning each stratum can have a different sampling rate. For example, high-minority schools are oversampled during national-only assessment years; and very small schools are undersampled across all NAEP assessment years to reduce data collection cost and burden.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Slide 18 of 36

Across all NAEP assessment efforts, in a year when state NAEP assessments are administered, approximately one million students are assessed in around 19,000 schools.

Approximately 600 schools and 12,000 students per subject and grade are included in the NAEP national assessments. It is important to note that the number of students assessed may vary between 6,000 and 20,000 depending on the grade and subject assessed.

Approximately 100 schools and 3,000 students per state, subject, and grade are included in the NAEP state assessments.

Approximately 50 schools and 1,600 students per district, subject and grade were included in the 2013 NAEP Mathematics and Reading TUDA assessments. It is important to note that the five largest TUDA districts have samples that are 50% larger than this.

The slides that follow will show the basic procedures for selecting schools and students in the state assessment years and the national-only assessment years. It is important to note, that these procedures differ.

Slide 19 of 36

Here is a table that shows the sampling frame and sample selected for public schools in one of the 50 states by strata. Urbanicity (or **Location category**) in this state consists of **Large or Mid-Sized Central City, Urban Fringe or Large or Mid-Size City, Large or Small Town, Rural, and New schools**. Within each of these urbanicity strata, schools are further grouped by minority density (or **Minority proportion**) groups of **Low-minority, Mid-minority, and High-minority** for two of the urbanicity strata. Within the **Rural** urbanicity strata, **NonMSA** and **MSA** are used to denote **non-metropolitan statistical area** and **metropolitan statistical area**.

As seen here, all of the 395 schools within this state are then assigned according to these strata, as seen in the **Frame** columns. Sampling frame data are obtained using the Common Core of Data (CCD) and Census data during state assessment years. The list obtained from CCD is updated to include any new schools that have been opened since the list was obtained, from a sample of districts in each state. A sample of new schools is selected from among those identified, at the same overall sampling rate as the initial school sample. As you can see, one new school was selected for the sample for this state.

Schools are then selected from each strata. About thirty students per subject are selected per school. This target of about thirty students per subject per school is a function of the psychometric properties of the NAEP assessment design.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 20 of 36**

It is important to remember that because the NAEP sampling design consists of multiple samples that either stand alone, or can be combined, to provide estimates of what student know and can do at for each reporting component (the Nation, states, and TUDA districts), the procedure for selecting schools and students just described in the previous slide is conducted for the remaining 49 states, the District of Columbia (DC), Puerto Rico (mathematics only), BIE, DoDEA, and 21 TUDA districts.

Each of the resulting tables, like the table previously shown, are then combined to form the NAEP sample from which results are reported – a sample representative of students enrolled in grades 4, 8, and 12 attending public and private elementary and secondary schools, Bureau of Indian Education (BIE), and Department of Defense Education Activity (DoDEA) schools physically located in the 50 states and the District of Columbia. Students physically located in Puerto Rico are included in NAEP's target population for mathematics only. Accordingly results from Puerto Rico are not included in the U.S. national results.

Slide 21 of 36

The national public results are obtained by combining the data obtained from students sampled as part of the 50 state samples, the DC sample, the BIE sample, the DoDEA sample, and the 21 TUDA samples. Each of these samples can be thought of in terms of the strata table shown in a previous slide.

The state results are obtained by combining the data obtained from students sampled as part of the individual state sample and any TUDA districts within the state. For example, the California sample consists of the California state sample, and three TUDA district samples (Fresno, San Diego, and Los Angeles).

It is important to note that if special equipment or procedures must be utilized to ensure effective administration, additional information from PSU-based samples may also be included to obtain national results. For example, the 2011 writing assessments were administered to PSU-based samples of students as field staff had to carry laptops into the schools and set them up for the assessment. The 2015 grade 12 assessment, although national-only, was not administered to PSU-based samples of students because the same field staff could administer these assessments along with the state/TUDA assessments.

Documentation should be reviewed carefully to understand the specific details regarding assessment design and sampling procedures relevant to analyses of interest.

Slide 22 of 36

In years when the NAEP samples are intended only to provide representation at the national level and not for each individual state or TUDA, a list of approximately 3,000 counties with known demographic characteristics is used. These characteristics are derived from Census data. Some of them are then merged together into PSUs. Each Metropolitan Area becomes a single PSU. Next, smaller, contiguous counties are

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

merged together to obtain a minimum population of around 50,000 people within the PSU. This yields a sampling frame of about 1,000 PSUs from which the sample can be drawn.

This list of about 1,000 PSUs is then stratified by region and urbanicity. About 95 to 100 PSUs are sampled, with probability of selection proportional to size of the PSU to minimize the difference in the overall probability of selection of students across the nation. About two dozen Metropolitan Areas are selected with certainty because they are so large.

Slide 23 of 36

This map shows a sample of PSUs selected for a NAEP, national-only sample. You can see that big cities like Los Angeles, New York, and Chicago were selected. You can also see that some are selected in each of the regions (Northeast, South, Midwest, and West). You may also notice that states with large areas and lower populations like Wyoming and North Dakota had no sampled PSUs in the national-only sample. National assessment field operations are mounted only in the sampled PSUs to minimize costs.

Once the PSUs are selected, schools within the PSUs are sampled using a national stratification based on region, Metropolitan Statistical Area, or MSA/nonMSA, and various characteristics derived from American Community Survey, or ACS, data found to be related to achievement at the PSU level. These include such factors as the percent of adults who graduated high school, and the percent of children living in poverty.

Sampling frame data are obtained using the Common Core of Data (CCD), Census data, and the Private School Universe Survey (PSS) during national-only assessment years. The list obtained from CCD is updated to include any new schools that have been opened since the list was obtained.

About thirty students per subject are selected per school. This target of about thirty students per subject per school is a function of the psychometric properties of the NAEP assessment design.

It is important to note that this procedure for selecting schools and students is used for both the Main NAEP (during national-only assessment years) and the LTT NAEP.

Slide 24 of 36

As was discussed in the common module titled, **Statistical Analysis of NCES Datasets Employing a Complex Design**, weights must be used in order to make estimates representative of the target population from the sample data.

As NAEP samples are not simple random samples, weights account for differential selection probabilities.

Sampling weights must be used to counterbalance the unequal probabilities of selection and ensure unbiased estimates for the population of interest; which within NAEP could be the Nation, states, or TUDA districts.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Additionally, as some members of the sample will not participate in NAEP assessments, sampling weights also account for differential response patterns to the NAEP assessments, to ensure that the data are representative of the population.

Slide 25 of 36

In the common module titled, **Statistical Analysis of NCES Datasets Employing a Complex Sample Design**, two standard error calculation procedures were discussed: Replication Techniques and Taylor Series linearization. Replication is a method that calculates appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples, or replicates. The Taylor Series linearization method uses PSU and strata identifiers to compute the appropriate standard errors.

Slide 26 of 36

Replication is a method that calculates appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples, or replicates. Within NAEP, the replication method should be used anytime you are interested in analyzing data regarding student characteristics, or contextual variables.

If this is the method you will use to calculate standard errors, you need to select the 62 replicate weights (**SRWT01** through **SRWT62**) that are associated with the student weight (**ORIGWT**).

NAEP replication weights use the paired jackknife (JK2) replication method so that is the method that should be specified within the statistical software used for analysis.

Appropriate software, such as AM, must be utilized to appropriately calculate standard errors for estimates generated with NAEP's plausible values, or PVs. AM is described within the module titled, **Considerations for Analysis of NAEP Data**. Plausible values will be discussed in the next section of this module.

Slide 27 of 36

The second method for calculating standard errors is the Taylor Series linearization method. This method uses PSU and strata identifiers to compute the appropriate standard errors. Within NAEP, Taylor Series Approximation should be used anytime you are interested in analyzing average scale scores or regressing achievement scores onto student characteristics or contextual variables.

If this is the method you will use to compute standard errors, you will select the cluster ID and strata identifiers, **JKUNIT** and **REPGRP1**, associated with the sampling weight named **ORIGWT**, and specify Taylor Series linearization as the method within the statistical software used for analysis.

Appropriate software, such as AM, must be utilized to appropriately calculate standard errors for estimates generated with NAEP's plausible values, or PVs. AM is described within the module titled, **Considerations for Analysis of NAEP Data**. Plausible values will be discussed in the next section of this module.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Slide 28 of 36

In summary, when analyzing NAEP data, weights must be used to counterbalance the unequal probabilities of selection, account for differential response patterns, and ensure unbiased estimates for the population of interest (the Nation, State, or TUDA districts). If using AM to conduct analyses, the replication method should be used anytime you are interested in analyzing data regarding student characteristics, or contextual variables; while the Taylor Series linearization method should be used when you are interested in analyzing average scale scores or regressing achievement scores onto student characteristics or contextual variables.

Slide 29 of 36

As discussed in the module titled, **Data Collected Through NAEP**, the NAEP assessments are designed such that no student responds to all of the items in the item pool. This method, known as multiple matrix sampling, is used with Item Response Theory (IRT) models to minimize testing burden while ensuring broad subject-matter coverage. In this approach, a participating student takes only a portion of the complete set of cognitive items developed for any assessment.

NAEP is not designed to report individual test scores, but rather produces estimates of scale score distributions for groups of students. The resulting scale score distributions describing student performance are transformed to a NAEP scale, and summary statistics of the scale scores are estimated.

Slide 30 of 36

Item Response Theory (IRT) models are used to describe the relationships between the item responses provided by students and the underlying score scales. IRT provides a common scale on which the performance of students receiving different blocks of items can be placed. For each item, item parameters that are used in the models are estimated from student response data. Different IRT models with different types of item parameters are used to describe multiple-choice items, constructed-response items that are scored right-wrong, and constructed-response items that have more than two categories. Multiple-choice items and constructed-response items that are scored right-wrong use the same IRT model; while constructed-response items that have more than two categories use a generalization of that IRT model. It is important to note that items of all types contribute to NAEP score scales.

Slide 31 of 36

Plausible values are proficiency estimates for an individual NAEP respondent, drawn at random from a conditional distribution of potential scale scores for all students in the sample who have similar characteristics and identical patterns of item responses. NAEP assigns plausible values to each respondent. The plausible values are not test scores for individuals in the usual sense; they are offered only as intermediary computations for calculating summary statistics for groups of students.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Plausible values are calculated and attached to the record of each student and are used to calculate summary statistics for NAEP reports and are available for the use of NAEP data users in secondary analyses of NAEP data.

Slide 32 of 36

In summary, all available plausible values should be used when calculating summary statistics for groups of students. There are five plausible values that should be used when analyzing NAEP data released before 2013. There are 20 plausible values that should be used when analyzing data from the 2013 Mathematics and Reading assessments forward. It is important to note that appropriate point estimates of individual scale scores cannot be calculated by averaging the available plausible values attached to a student's file.

Slide 33 of 36

To summarize, it is important to remember that the NAEP assessment design and sample design work together to ensure representative and valid measures of student achievement across various subjects for each of the reporting components (national, state, and TUDA). These two key design features have implications for analysts of NAEP data. The complex sample design must be accounted for when analyzing NAEP data. This means that weights must always be used that are appropriate to the analysis. The assessment design must also be accounted for and plausible values (PVs) must be used.

Weights and PVs must be used simultaneously when analyzing NAEP data.

Slide 34 of 36

As weights and PVs must be used simultaneously when analyzing NAEP data, to avoid unintentional errors, analysts of NAEP data should always use an analytic tool that simultaneously accounts for the complex sample design and properly handles the plausible values. These tools include the NAEP Data Explorer (NDE), NAEPEX, and AM Software, which will be discussed in the module titled, **Considerations for Analysis of NAEP Data**.

Analysts should always review the documentation carefully to understand the specific details regarding assessment design and sampling procedures relevant to analyses of interest.

Slide 35 of 36

This module described the key design features of NAEP. It described the sampling weights that must be applied when using NAEP data to ensure that estimates are representative of the target population. This module also described the appropriate procedures for calculating standard errors using NAEP data. Finally, this module described the implications of the NAEP assessment design, IRT models, and plausible values on analysis of NAEP data.

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 36 of 36**

Important resources provided throughout the module are summarized here for your reference.

You may now proceed to the next module in the series, or click the **Exit** button to return to the landing page.