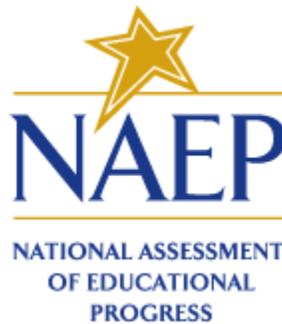


NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values

Slide 1 of 36



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

Slide 2 of 36

Module Objectives

- Describe the key design features of NAEP
- Describe the sampling [weights that must be applied](#) to ensure that estimates are representative of the target population
- Describe appropriate procedures for [calculating standard errors](#)
- Describe 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 3 of 36****Overview of NAEP**

- 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**

- Assesses grades 4, 8, and 12 in public and private schools
- National, state, and selected urban district results
 - Every two years (odd numbered years)
 - Reading
 - Mathematics
 - Every four years (even numbered years)
 - Science
 - Writing
- Other subjects are national-only
 - U.S. History, Geography, Civics, Economics, and Arts

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 5 of 36****NAEP Long-Term Trend (LTT)**

- Designed to provide information on the changes in the performance of the nation's 9-, 13-, and 17- year-old students
 - No state or district level results
- Has measured students' performance in mathematics and reading since 1971
 - 2012 is the most recent; 2020 will be the next administration

Slide 6 of 36**NAEP Reporting**

NAEP reports results for:

- The Nation (Main NAEP National Assessment and/or NAEP Long-Term Trend)
- States (State Assessment)
- Large urban school districts (Trial Urban District Assessment (TUDA))

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 7 of 36****Key Design Features of NAEP**

- 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 procedures ensure 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 Assessment Design**

- 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
 - Takes only one subject area assessment
- 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 Long Term Trend (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

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 9 of 36****A Representative Sample is Required**

[NAEP Authorization Act](#), SEC 303 (b) (2) (A):

- “The Commissioner for Education Statistics... shall...use a random sampling process which is consistent with relevant, widely accepted professional assessment standards and that produces data that are representative on a national and regional basis.”

Slide 10 of 36**Defining the Target Populations**

- NAEP has multiple target populations for which results are reported (the Nation, states, and TUDA districts)
 - 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 the students in the target populations, two conditions must be satisfied:
 - The probability of being sampled must be known
 - Every member of the target populations has to have at least some chance of selection into the sample

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 11 of 36****Included in NAEP's Target Populations**

- The target populations for Main NAEP include students
 - Enrolled in grade 4, 8, or 12
 - Attending public, private, Bureau of Indian Education (BIE), and Department of Defense Education Activity (DoDEA) schools
 - Located in the 50 states and DC
 - Puerto Rico 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**NOT Included in NAEP's Target Populations**

- Students
 - Enrolled in ungraded classrooms
 - Schooled at home, by correspondence, or online
- Students enrolled in schools
 - Serving youth in hospitals or correctional facilities
- Students with disabilities (SDs) and English language learners (ELLs) who cannot be assessed, even with accommodations

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 13 of 36****Sample Designs Used in NAEP**

- Representative samples of students must be drawn from each of these target populations from which valid estimates of what students know and can do will be reported across three reporting components (the Nation, states, and TUDA districts)
- The samples can be thought of in terms of jurisdictions (any government-defined geographic area that is sampled in the NAEP assessment)
 - The United States
 - A state
 - The District of Columbia
 - Puerto Rico
 - An urban district
 - The Department of Defense Education Activity schools (DoDEA)

Slide 14 of 36**Sample Designs Used in NAEP (Continued)**

- 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 state are drawn that are large enough to support state-level estimates
 - In national-only assessment years, about 1,000 Primary Sampling Units (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

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 15 of 36****Sample Designs Used in State Assessment Years**

- 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
 - CCD information is used to stratify the list 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 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**Sample Designs Used in National-Only Assessment Years**

- 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 (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

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 17 of 36****Sampling Schools Across All NAEP Assessment Years**

- 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 a sampled large school are less likely to be selected than students within a sampled small school
 - 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
- Each stratum can have a different sampling rate
 - High-minority schools are oversampled during national-only assessment years
 - Very small schools are undersampled across all NAEP assessment years to reduce data collection cost and burden

Slide 18 of 36**Providing Data that are Representative on a National and Regional Basis**

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

- National assessments
 - Approximately 600 schools and 12,000 students per subject and grade
 - May vary between 6,000 and 20,000 students depending on the grade and subject assessed
- State assessments
 - Approximately 100 schools and 3,000 students per state, subject, and grade
- TUDA assessments
 - Approximately 50 schools and 1,600 students per district, subject and grade were included in the 2013 NAEP Mathematics and Reading TUDA assessments
 - The five largest TUDA districts have samples that are 50% larger than this

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

Slide 19 of 36

Procedure for Selecting Schools and Students in State Assessment Years

Strata		Frame			Sample		
		Schools	Students		Schools selected	Students	
			Enrollment	Population		Selected students	Estimate of population
Location category	Minority proportion						
Large or Mid-Sized Central City	Low-minority	12	2,549	5.1%	6	433	4.6%
	Mid-minority	14	2,428	4.8%	5	465	4.8%
	High-minority	15	2,581	5.2%	6	558	5.8%
Urban Fringe of Large or Mid-Size City	Low-minority	26	4,368	8.7%	9	777	8.1%
	Mid-minority	25	3,995	8.0%	10	807	8.5%
	High-minority	22	4,268	8.5%	9	747	7.8%
Large or Small Town		83	10,822	21.6%	25	2,039	21.3%
Rural	NonMSA	154	12,659	25.3%	38	2,359	24.7%
	MSA	44	6,461	12.9%	15	1,395	14.5%
New schools					1	21	
Total		395	50,131	100.0%	124	9,601	100.0%

Slide 20 of 36

Combining the Samples in State Assessment Years

- 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 students know and can do for each reporting component (the Nation, states, and TUDA districts), the procedure for selecting schools and students just described for Kentucky 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 shown in the previous slide, 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
 - Results from Puerto Rico are not included in the U.S. national results

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 21 of 36****Combining the Samples in State Assessment Years (Continued)**

- 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 the example
- 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
 - The California sample consists of the California state sample, and three TUDA district samples (Fresno, San Diego, and Los Angeles)
- 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
 - 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**Sampling Schools in National-Only Assessment Years**

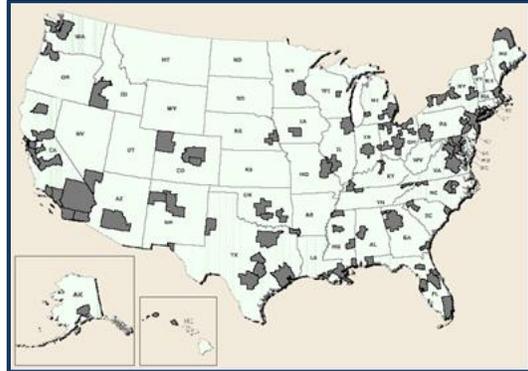
- Begins with a list of approximately 3,000 counties with known demographic characteristics derived from Census data
- Some of them are then merged together into PSUs
 - Each Metropolitan Area becomes a single PSU
 - Smaller, contiguous counties are merged to obtain a minimum population of around 50,000 people within the PSU
- This yields a sampling frame of about 1,000 PSUs from which the sample can be drawn
- The PSUs are stratified by region and urbanicity
- About 95-100 of the PSUs are sampled
 - The probability of selection is proportional to size of the PSU
 - About two dozen Metropolitan Areas are selected with certainty because they are so large

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

Slide 23 of 36

Sampling Schools and Students in National-Only Assessment Years

- Schools within the PSUs are sampled using a national stratification based on
 - Region
 - Metropolitan Statistical Area (MSA)/ nonMSA
 - Characteristics derived from American Community Survey (ACS) data related to achievement at the PSU level
 - Percent of adults who graduated high school
 - 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
- About 30 students per subject are selected per school. This target of about 30 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 is used for both the Main NAEP (during national-only assessment years) and the LTT NAEP



Slide 24 of 36

Use of Weights: Review

- Used to make estimates representative of target population from the sample data
- Account for differential selection probabilities
- Must be used to counterbalance the unequal probabilities of selection and ensure unbiased estimates for the population of interest (the Nation, states, or TUDA districts)
- Account for [differential response patterns](#) to the NAEP assessments

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 25 of 36****Calculating Standard Errors Using NAEP Data**

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 (replicates)
- The Taylor Series linearization method uses PSU and strata identifiers to compute the appropriate standard errors

Slide 26 of 36**Standard Error Calculation in NAEP: Replication Techniques**

- This method calculates appropriate SEs based on differences between estimates from the full sample and a series of created subsamples (replicates)
- This method should be used anytime you are interested in analyzing data regarding student characteristics, or contextual variables
- Select replicate weights that are associated with the student weight (ORIGWT) and the associated 62 replicate weights (SRWT01 through SRWT62)
- NAEP replication weights use the paired jackknife (JK2) replication method
- Appropriate software, such as AM, must be utilized to appropriately calculate standard errors for estimates generated with NAEP's plausible values (PVs)

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 27 of 36****Standard Error Calculation in NAEP: Taylor Series linearization**

- This method uses primary sampling unit (PSU) and strata identifiers to calculate appropriate SEs
- This method should be used when you are interested in analyzing average scale scores or regressing achievement scores onto student characteristics or contextual variables
- Select the cluster and stratum identifiers (JKUNIT and REPGRP1) associated with the sampling weight variable (ORIGWT)
- Taylor Series linearization must be specified as the method to be used for analysis within the statistical software
- Appropriate software, such as [AM](#), must be utilized to appropriately calculate standard errors for estimates generated with NAEP's plausible values (PVs)

Slide 28 of 36**Sample Design Implications for NAEP Analysts**

- Sampling 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, states, 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
 - 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

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 29 of 36****NAEP Assessment Design: Review**

- NAEP assessments are designed such that no student responds to all of the items in the item pool
 - Multiple matrix sampling is used with Item Response Theory (IRT) models to minimize testing burden while ensuring broad subject-matter coverage
 - 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 student test scores, but rather produces estimates of scale score [distributions for groups of students](#)
 - 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**NAEP IRT Models**

- Item Response Theory (IRT) models are used to describe the relationships between the item responses provided by students and the underlying score scales
 - Provides a common scale on which 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
 - Constructed-response items that have more than two categories use a generalization of that IRT model
- Items of all types contribute to NAEP score scales

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 31 of 36****Plausible Values**

- 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
- 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**Assessment Design Implications for NAEP Analysts**

All available plausible values should be used when calculating summary statistics for groups of students

- Appropriate point estimates of individual scale scores cannot be calculated by averaging the available plausible values attached to a student's file

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

Slide 33 of 36

Summary of Implications for NAEP Analysts

- Remember that the NAEP 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)
- These two key design features have implications for analysts of NAEP data
 - The complex sample design must be accounted for (**weights must be used**)
 - The assessment design must be accounted for (**plausible values (PVs) must be used**)
- **Weights and PVs must be used simultaneously** when analyzing NAEP data

Slide 34 of 36

Summary of Implications for NAEP Analysts (Continued)

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 properly and simultaneously accounts for these key design factors
 - NAEP Data Explorer (NDE)
 - NAEPEX
 - AM Software
 - Other software that can handle both complex survey designs (via jackknife replicate weights or Taylor Series linearization, AND plausible values)
- Analysts should always review the documentation carefully to understand the specific details regarding assessment design and sampling procedures relevant to analyses of interest

NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values**Slide 35 of 36****Module Summary**

- Described the key design features of NAEP
- Described the sampling weights that must be applied to ensure that estimates are representative of the target population
- Described appropriate procedures for calculating standard errors
- Described the implications of the NAEP assessment design, IRT models, and plausible values on analysis of NAEP data

Slide 36 of 36**Module Resources**

- [Weights that must be applied](#)
- [Calculating standard errors](#)
- [Overview of the NAEP Assessment Design](#)
- [NAEP Authorization Act](#)
- [NAEP multi-stage complex sampling design](#)
- [Treatment of Missing Responses in NAEP](#)
- [AM Software](#)
- [NAEP Analysis and Scoring](#)
- [Item Scaling Models](#)
- [Plausible Values Versus Individual Scale Scores](#)