



An Employee-Owned
Research Corporation

1650 Research Boulevard
Rockville, MD 20850-3195
tel: 301-251-1500
fax: 301-294-2040
www.westat.com

Date: March 25, 2013 **Memo:** 2014-3.2PSU/1.2S

To: Bill Ward, NCES David Freund, ETS
Amy Dresher, ETS Nicole Beaulieu, ETS
Ed Kulick, ETS Connie Smith, Pearson
Nancy Caldwell Dianne Walsh
Debby Vivari Lauren Byrne
William Wall Rob Dymowski
Leslie Wallace John Burke
Erin Wilson Joel Wakesberg
Chris Averett Lloyd Hicks
Kavemuii Murangi

From: Dave Hubble

Reviewer: Keith Rust

Subject: Sample Design for 2014 NAEP – Overview

I. Introduction

For 2014, the sample design involves several components, all of which are national assessments of one kind or another.

1. “Operational” assessment in Civics at grades 4, 8, and 12;
2. “Operational” assessment in U.S. History at grades 4, 8, and 12;
3. “Operational” assessment in Geography at grades 4, 8, and 12;
4. “Probe” assessment in Technology and Engineering Literacy (TEL) at grade 8;
5. Pilot tests at grades 4, 8, and 12:
 - a. Science paper & pencil;
 - b. Science Hands-on-Tasks (HOT) lab;
 - c. Science Interactive Computer Tasks (ICT) computer-delivered.
 - i. Small tryout in August 2013
 - ii. Large tryout January 2014

The schools to be included in the small tryout part of Science ICT pilot will be identified by NAEP NSSC staff in a manner sufficient to meet the beta version testing needs and will not be included in the

remainder of this overview. The target sample sizes of assessed students for the remaining components are shown in Table 1 (which also shows an estimate of the required number of participating schools). All of these assessments are to take place in the typical NAEP testing window of late January to early March 2014. Note that about 22% of the assessed students will come from grade 4, about 49% from grade 8, and 29% from grade 12.

Table 1. 2014 NAEP Sample Sizes (Public and Private)

	Session	Public school students	Private school students	Total students
Grade 4				
Civics (O)	0401	5,400	600	6,000
Science (P)		900	100	1,000
U.S. History (O)	0402	5,400	600	6,000
Geography (O)		5,400	600	6,000
Science HOT (P)	0403	450	50	500
Science ICT (P)	0404	2,100		2,100
Total		19,650	1,950	21,600
Schools		380	100	480
Grade 8				
Civics (O)	0801	7,200	800	8,000
Science (P)		900	100	1,000
U.S. History (O)	0802	9,000	1,000	10,000
Geography (O)		7,200	800	8,000
TEL (PR)	0803	18,000	2,000	20,000
Science HOT (P)	0804	450	50	500
Science ICT (P)	0805	2,100		2,100
Total		44,850	4,750	49,600
Schools		1,150	190	1,340
Grade 12				
Civics (O)	1201	7,200	800	8,000
Science (P)		900	100	1,000
U.S. History (O)	1202	9,000	1,000	10,000
Geography (O)		7,200	800	8,000
Science HOT (P)	1203	450	50	500
Science ICT (P)	1204	2,100		2,100
Total		26,850	2,750	29,600
Schools		425	45	470
GRAND TOTAL				
Schools		1,955	335	2,290

(O) = Operational, (PR) = Probe, and (P) = Pilot

II. Assessment Types

From a sampling and operations point of view, many types of assessment sessions can be distinguished. The detailed target counts of assessed students are provided in Table 1. Below is a summary of major points.

1. The U.S. History and Geography spiral (HG) at grades 4, 8, and 12. This spiral must be assessed in a different physical session from the others, but will be in the same schools as the omnibus session type (see immediately below). The session has a target of 12,000 (6,000 U.S. History and 6,000 Geography) assessed students at grade 4 and 18,000 (10,000 U.S. History and 8,000 Geography) each at grades 8 and 12.
2. The omnibus spiral (C) includes Civics and a Science pilot at grades 4, 8, and 12. At grade 4, this session spiral has a total target of 7,000 assessed students. At grades 8 and 12, this session spiral has a total target of 9,000 assessed students.
3. The Technology and Engineering Literacy (TEL) probe assessment for grade 8 will be computer delivered. Because of the different delivery method, this assessment must be in separate sessions. In fact, out of concern for overburdening schools conducting operational assessments, an additional set of PSUs, with minimum overlap with operational PSUs, will be used for conducting TEL with a target of 20,000 assessed students.
4. The Science HOT pilot for grades 4, 8, and 12 must be assessed in different physical sessions from others, but will be in the same schools as the omnibus session type (Civics+) and the History/Geography sessions. The session has a target of 500 assessed students at each grade.
5. The Science ICT pilot (i.e., large tryout) for grade 4, 8, and 12 will be computer delivered and as such these assessments must be in separate sessions. Out of concerns for overburdening schools conducting operational assessments, these assessments will be conducted in the same set of PSUs as the TEL assessment. Additionally, efforts will be made to minimize overlap with TEL sampled schools. The session has a target of 2,100 assessed students at each grade.

III. Primary Sampling Units Selection and Overlap Control

As the U.S. History, Geography, Civics, and pilot samples assessments are national, with a total sample size of assessed students of about 85,000, for reasons of operational efficiency in conducting the assessments a sample of Primary Sampling Units (PSUs) was selected, and all sampled schools will be drawn from within the sampled PSUs. With a smaller sample size of about 26,000 assessed students for computer delivered TEL probe in grade 8 and Science ICT in grades 4, 8, and 12, a separate sample of PSUs was selected with the largest PSUs being in common to both PSU samples.

The PSUs were created from aggregates of counties. Data on counties were obtained from the 2010 Census, and the definitions of Metropolitan Statistical Areas (MeSAs) used were the December 2009 Office of Management and Budget (OMB) definitions. Each Metropolitan Statistical Area (MeSA) constitutes a PSU, except that MeSAs that cross Census region boundaries were split into their individual regional components.

Non-metropolitan PSUs were formed by aggregating counties into geographic units of sufficient minimum size to provide enough schools to constitute a workload of about 1% of the total sample. These

PSUs were made of contiguous counties where possible, and almost contiguous counties (separated by MeSA counties) otherwise. Each PSU falls within a single state.

This process generated a frame of approximately 1,000 PSUs. The PSUs were stratified, using characteristics aggregated from county-level characteristics, found by analysis to be related to NAEP achievement in past assessments. A sample of 67 PSUs was selected for the Operational sample. The 29 largest MeSAs were selected with certainty, and the remaining sample was a stratified probability proportional to size (PPS) sample, where the size measure was a function of the number of children as given in the most recent population estimates prepared by the U.S. Census Bureau. For the Operational sample, 76 such strata were formed and paired and a single PSU was selected from one stratum in each of the 38 pairs for a total of 67 PSUs. For the computer-delivered TEL grade 8 and Science ICT grades 4, 8, and 12 samples, the same certainty PSUs were selected. However, in order to ensure no overlap in sampled schools a single PSU was selected from each of the strata in the 38 pairs not selected for the Operational sample for a total of 67 PSUs. Also note that the PSUs for the TEL and Science ICT samples were selected in such a way as to minimize overlap with the NAEP 2013 sample PSUs. This was done to reduce the chance that a school was selected for the 2013 TEL pilot and the 2014 operational TEL assessment.

IV. Stratification and Oversampling

As in the recent past, the plan is to draw separate public and private school samples. This approach has proven to be useful, in that, selecting the samples separately has three advantages: 1) it permits the timing of sample selection to vary between public and private schools, should this prove necessary; 2) it allows us to readily assume different response and eligibility rates for public schools and private schools; and 3) it makes it easier to use different sort variables for public schools and private schools. It also allows for the possibility of a late change of mind concerning the sample sizes that differ between public and private schools. Note that the Science ICT computer-delivered design will not include a private school component as the assessment goals can be better met through other means in this case.

Explicit stratification will take place at the PSU level. For schools within PSUs, stratification gains are achieved by sorting the school file prior to systematic selection. As in past national samples, the expectation is that, within the set of certainty MeSA PSUs within a census region, PSU will not necessarily be the highest level sort variable. Thus, type of location will be used as the primary sort variable. Consider for example the large MeSAs in the Midwest region. The design is aimed primarily at getting the correct balance of city, suburban, town, and rural schools, as a priority over getting exactly a proportional representation from each MeSA (Chicago, Detroit, Minneapolis), although of course it should be possible to get a high degree of control over both of these characteristics. The sort of the schools will use other variables beyond the type of location variable, such as, a race/ethnicity percentage variable. The exact set of variables used in sorting the schools prior to sampling will be specified in the particular sampling specification memos.

In addition, we will implement oversampling of high-minority schools within the public schools. That is, as used in past national assessments, a public school with over 15 percent Black and Hispanic combined enrollment will be given twice the chance of selection of a public school of the same size with a lower percentage of these two groups. This approach is effective in increasing substantially the sample sizes of Black and Hispanic students, without inducing undesirably large design effects on the sample, either overall or for particular subgroups. This oversampling will be performed for the Operational, TEL, and Science ICT samples. Beyond this, we will also explore the oversampling of Black and Hispanic students at the student level in schools not being oversampled at the schools level, that is, schools with less than 15 percent Black and Hispanic students.

The updated preliminary 2011/12 CCD and the updated 2011/12 PSS school files have been approved for use by NCES. They will serve as the public and private school frames for the 2014 NAEP.

V. New Schools

To compensate for the fact that the CCD file used to create the NAEP public school sampling frames is out of date at the time of frame construction, we will supplement the samples for the Operational and TEL assessments with a sample of new public schools for each grade. New school samples will not be developed for the private school samples.

The new school samples will be drawn using a two-stage design. At the first stage, a national sample of school districts will be selected from the Operational and TEL sample PSUs. The sampled districts will be asked to review lists of their respective schools and identify new schools. Frames of new schools will be constructed from these updates, and new schools will be drawn with probability proportional to size using the same sample rates as their corresponding original school samples.

Note that the student and school sample sizes in Table 1 do not reflect these new school samples. However, some schools from the original sample will prove to be closed or otherwise ineligible, and the new school procedure essentially compensates for the sample losses from these sources, as well as ensuring full coverage of the population.

VI. Within PSU Overlap Control with Other Samples

In keeping with the efforts at the PSU level to reduce potential overlap between the Operational paper-based and computer delivered assessment samples, methods will be employed to reduce overlap during sample school selection within the PSUs that contain more than one sample. We anticipate that it should be possible to avoid any school overlap among the different school samples at a given grade. Schools may be selected to participate at more than one grade.

The Keyfitz method will be used to compute conditional probabilities to reduce the overlap between the samples within grade. That is, in the 29 certainty PSUs that overlapped between Operational paper-based and TEL, the conditional probabilities of selection for the TEL schools will be based on the Operational schools sampling outcomes. Specifically, this will be done to reduce overlap between Operational grade 8 sample schools and the TEL grade 8 sample schools.

In addition, overlap control will also be performed to reduce the number of Science ICT sample schools that have already been selected for the Operational and the TEL assessments in the 29 certainty PSUs and the TEL assessment in the 38 TEL/Science ICT non-certainty PSUs.

The Keyfitz method will be employed in a fashion similar to that described above for areas that overlap between the Science ICT and the TEL sampling outcome. In areas where all 3 samples are located, the conditional probabilities used for the Science ICT pilot will be based on the Operational and TEL sampling outcomes.

VII. Substitute Samples

Substitute samples will be selected for each of the 2014 samples in the following order: Operational, TEL, then Science ICT. Within these samples, the order for selecting substitute schools will be from “oldest” to “youngest”. That is, grade 12, 8, and then 4 for the Operational samples, grade 8 for TEL sample, and grade 12, 8, and then 4 for the Science ICT samples. This ordering of samples and

grades is necessary since no school can be selected as a substitute more than once. So, it is more critical for operational and probe samples to precede the pilot samples and higher grades to precede lower grades due to having fewer schools available to serve as substitutes at the higher grades. This will be done separately for both public and private schools. The general steps for selecting substitutes are to put the school frames in their original sampling sort order, and take the 'nearest neighbor' of each original sampled school, excluding schools selected for any of the NAEP 2014 samples, schools already selected to serve as a substitute school, and schools which cross PSU or state boundaries, as potential substitutes.

The nearest neighbor is the school adjacent (immediately preceding or succeeding) the original school in the sorted frame with the closer estimated grade enrollment value. If estimated grade enrollment of both potential substitute schools differs from the original school by the exact same amount, the selection procedure will randomly choose one of the schools. If neither the preceding or succeeding school is eligible to be a substitute, then the sampled school is not assigned a substitute.

In addition, the few sampled private schools whose school affiliation is unknown will not get substitutes nor can such private schools not in sample serve as substitute schools. Also, new schools will not get substitute schools nor serve as substitutes.

VIII. Student Sampling

Student sample sizes within each school are determined as the combined result of several factors:

1. We wish to take all students in relatively small schools.
2. We wish to avoid the situation where all but a few students (e.g., more than 90%, but fewer than 100%) are tested.
3. We do not wish to have a sample that is too clustered for any one assessment subject.
4. We do not wish to have many physical sessions that contain only a very small number of students, as this is inefficient.
5. We wish to minimize the number of unique combinations of session types in the schools and to avoid three session types in a given sample school.
6. We do not wish to overburden the schools with unduly large student samples.
7. At grade 4, if we sample fewer than all students, in a school with fewer than 120 students, the school will probably elect to have us sample all students.
8. We wish to be cognizant of the fact that the Administration Schedule template has 34 lines on it.
9. We wish to consider what will be the bundle sizes of booklets. For our purposes at this time, this was based on historic bundle sizes at grades 4, 8, and 12.

The plans below reflect the design that results from considering each of these factors and balancing them.

Operational: Grade 4 Schools

We will select all students, up to 120. In schools with more than 120 students, we will select 100.

There are three session types, the Civics/Science Pilot (C/P) assessments, the U.S. History/Geography (H/G) assessments, and the Science HOT (HOT) pilot assessment. The proportion of students assigned is 24/39 for the H/G session, 14/39 for the Civics/Science Pilot session, and 1/39 for the Science HOT session for fourth grade. Minimum session size is 12 within schools with 12 or more students. In order to achieve the minimum session size, every school will not be assigned each session. In fact, since H/G comprises over half the sample, efforts were made to reduce the number of unique session combinations within sampled schools. This assignment of the C/P, H/G, and HOT sessions, based on the number of students in the school, is detailed in Table 2.

Table 2. 4th grade school session allocations and proportions

Enrollment size	Grade 4			
	1 to 23	24 to 47	48 to 94	95 and higher
Probability of being assigned H/G and HOT only	0	2/39	1/10	1/8
Proportion of sample students assigned to HOT (in schools with H/G and HOT session types)	NA	1/2	10/39	8/39
Probability of being assigned C/P and H/G only	0	28/39	9/10	7/8
Proportion of sample students assigned to H/G (in schools with C/P and H/G session types)	NA	1/2	211/351	437/741
Probability of being assigned C only	14/39	0	0	0
Probability of being assigned H/G only	24/39	9/39	0	0
Probability of being assigned HOT only	1/39	0	0	0

Operational: Grade 8 and Grade 12 Schools

We will select all students, up to 110. In schools with more than 110 students, we will select 100.

There are three session types, the Civics/Science Pilot (C/P) assessments, the U.S. History/Geography (H/G) assessments, and the Science HOT (HOT) assessments. The proportion of students assigned is 36/55 for the H/G session, 18/55 for the C/P session, and 1/55 for the HOT session for eighth and twelfth grade. Minimum session size is 12 within schools with 12 or more students. In order to achieve the minimum session size, every school will not be assigned each session. In fact, since H/G comprises over half the sample, efforts were made to reduce the number of unique session combinations within sampled schools. This assignment of the C/P, H/G, and HOT sessions, based on the number of students in the school, is detailed in Table 3.

Table 3. 8th and 12th grade school session allocations and proportions

Enrollment size	Grade 8 and 12			
	1 to 23	24 to 47	48 to 94	95 and higher
Probability of being assigned H/G and H/G only	0	2/55	1/14	1/11
Proportion of sample students assigned to H/G (in schools with H/G and H/G session types)	NA	1/2	14/55	11/55
Probability of being assigned C and H/G only	0	36/55	13/14	10/11
Proportion of sample students assigned to H/G (in schools with C and H/G session types)	NA	1/2	463/715	16/25
Probability of being assigned C only	18/55	0	0	0
Probability of being assigned H/G only	36/55	17/55	0	0
Probability of being assigned H/G only	1/55	0	0	0

TEL Probe: Grade 8 Schools

We will select all students, up to 30. In schools with more than 30 students we will select 30. All students will be assigned to the TEL Probe.

Science ICT Pilot: Grade 4, 8, and 12 Schools

We will select all students, up to 30. In schools with more than 30 students we will select 30. All students will be assigned to the Science ICT Pilot sessions for each grade.

IX. Weighting Requirements

Operational Samples

These samples will have a single set of weights for each subject (Civics, U.S. History, and Geography at grades 4, 8, and 12) applied to reflect probabilities of selection, school and student nonresponse, any trimming, and the random assignment to the particular subject. There will be a separate replication scheme by grade and public/private.

TEL Probe Sample

As with the Operational samples, the TEL Probe sample at grade 8 will be fully weighted.

Pilot Test Samples

We will not weight the students in the Science Pilot, Science H/G pilot, or Science ICT pilot test studies at grades 4, 8, and 12. However, preliminary weights will be available for these pilot test samples, if needed.