National Assessment of Educational Progress (NAEP) -- The Sample and Data-Collection Design for Year 15

Morris H. Hansen, Benjamin J. Tepping, Josefina A. Lago and John Burke

1. Introduction

The sample for the Year 15 (1983-84) National Assessment of Educational Progress was a multistage probability sample, with counties or groups of counties serving as first-stage sampling units, elementary and secondary schools serving as second-stage sampling units, the assignment of sessions by type to sampled schools serving as a third stage of sampling, and the selection of students within schools and their assignment to sessions serving as the fourth stage of sampling.

A total of 64 first-stage units was included in the sample, and assessments were conducted at 1,465 schools. Various blocks or packages of exercises were administered in these schools to a total of about 30,000 students in each of the three ages 9, 13 and 17 together with the corresponding modal grades 4, 8, and 11.

To facilitate the transition to a new organization (the Educational Testing Service (ETS) was the new grantee responsible for the NAEP project with Westat as the survey subcontractor) the sample of PSU's and schools was drawn by the Research Triangle Institute (RTI), the earlier survey subcontractor. These samples were drawn following the principles and methods developed by RTI, and similar to those of recent earlier assessments.[1] Procedures more or less similar to those of prior assessments were used for subsequent stages of sampling also, but with some important differences to accommodate new goals adopted by ETS that have an impact on the sampling procedures. The principal new goals (as discussed more fully in the papers by S. Messick and A. Beaton) include the following:

- In prior assessments the students sampled and assessed were those in ages 9, 13, and 17. In this assessment the decision was made to draw samples to assess students of these ages, and also the corresponding modal grades 4, 8, and 11.

- In earlier assessments the test items had been assembled into various packages and the same package of exercises was administered to all students in a session, usually consisting of a sample of about 20 students. In Year 15 ETS developed and specified a new procedure in which exercises were grouped into a larger number of smaller blocks, and assembled into test books in a balanced incomplete block (BIB) design. These books were then assigned to students in a rotating or "spiraled" design so that different books were assigned to each student in a session. In addition, some of the assessments were administered as in the past, to provide comparable procedures for measuring change. In these sessions all students were administered the same "package" of items, and the questions were presented orally from a recorded tape as well as visually, or were paced by a tape recording.

- A questionnaire was obtained for a sample of teachers of sampled students, to permit correlating teacher and student characteristics.

- Earlier assessments had identified and excluded from the assessment students with limited English proficiency or certain handicaps. For Year 15 such students were again excluded, but a questionnaire was obtained for a sample of them to allow additional description and analysis.

Some other changes were made in an effort to reduce costs, or to reduce sampling variances or nonresponse biases, or both. Among these were:

- Assessments were administered in moderately larger session sizes for Year 15 than in earlier assessments.

- Adjustments for nonresponse were made session by session, as in the past, for the comparably administered taped assessments. Somewhat different adjustments for nonresponse were made for the assessments administered by the new spiraled procedures.

- A post-stratification procedure was introduced to replace the earlier "smoothing" procedure.

We provide a brief general description of the Year 15 survey design in what follows.

2. The Sample of First-Stage Units

The first-stage sample was a stratified sample of 64 primary sampling units (PSU's), drawn by RTI to represent the 50 states and the District of Columbia. Each primary sampling unit consisted of a county or a group of

counties. Counties were grouped only as needed to achieve a specified minimum size in terms of numbers of eligible students. The number of PSU's to be selected for the sample and their minimum size were specified by Westat. The specified total of 64 PSU's to be selected was the same as for the Year 13 assessment, and was deemed minimal but sufficient to control the PSU contribution to variance to a reasonable level. Following is a brief description of procedures followed by RTI for defining, stratifying and selecting the sample of PSU's.[2]

- Twenty primary strata of counties were defined, utilizing 1980 Census data, based on four geographic regions by five "Sample Description of Community" (SDOC) classes. The latter separately identified (1) SMSA counties containing at least 10,000 or more population in a big city (a city of 200,000 population or more), (2) remaining counties in "big city" SMSA's, (3) other counties containing any part of a city of 25,000 or more population, (4) all other counties not identified as extreme rural, and (5) counties identified as extreme rural (i.e., with less than 10,000 urban population, non-zero farm employment, and classified as extreme rural on the basis of an occupational index).

- Preliminary measures of size were computed for each county (frame unit) by separately estimating the enrollment of 9-, 13-, and 17-year-olds in elementary and secondary schools for each county, using Quality Education Data, Inc. (QED)[3] data on school grade-range and total enrollment, and using prediction formulas developed by RTI on the basis of prior experience. The preliminary measure of size was the average enrollment of the three age classes.

- Adjusted measures of size were computed by doubling the preliminary measures of size for counties identified as extreme rural and for low socio-economic status (Low-SES) tracts of "big" cities. (Low-SES Census tracts were identified within the central big cities in the counties included in SDOC class 1, based on an index of SES computed for each Census tract.)

- The number of PSU's to be sampled was allocated to the 20 primary strata, approximately in proportion to the adjusted measures of size.

- PSU's were defined within the 20 primary strata. Single large counties served as PSU's. Other counties were grouped within states (with minor exceptions), each PSU to include a minimum adjusted measure of size of 1,000. The PSU's within each primary stratum were then ordered by state (after states within a region were ordered in a serpentine manner), and by percent minority within state (with reverse ordering in successive states).

- PSU's were selected with probabilities proportionate to the adjusted measures of size without replacement from this ordered list for each of the 20 strata (using a selection procedure developed by Chromy[4]). The two largest PSU's were selected twice.

3. The Initial Sample of Schools

An initial sample of 1,682 schools was selected from the 64 primary sampling units, with the selections carried out independently for the three age classes. A total of 700 schools was selected for age 9 (and grade 4), 588 for age 13 (and grade 8), and 394 for age 17 (and grade 11)[5]. However, some schools contained eligibles for two or more of the age classes and were selected more than once so that a total of 1,587 distinct schools was selected. Enough schools were selected within an age class in each PSU to yield the desired sample size of students, with a reserve to allow for some ineligible schools and for some nonparticipation of schools, based on Year 13 experience.

As an approximation to optimum allocation, the general goal was to draw the successive stages of samples with varying probabilities such that the overall probability of selection of a student to take a particular type of assessment booklet would be the same for each student. This was a goal in the NAEP sample, except for the planned over- and undersampling.

The sample of schools was selected to allow a maximum of about 200 age- or grade-eligibles to be invited to assessment sessions in a school in the 9-year age class and up to about 250 age- or grade-eligibles in the 13- and 17-year age classes. While these specifications allow relatively large samples of students from some individual schools, the average number of students assessed per school was well below the maximum. Moreover, only a small fraction of students assessed in a school is assessed for a given block of exercises. It was recognized that variances would be increased by allowing maximum cluster sizes up to these levels but perhaps not unduly in relation to cost savings.

After some initial study, it was estimated that the number of students in a school that were eligible by either age or modal grade would average roughly 1.3 times the number of age-eligibles. This would vary by age class and from school to school. In sample selection, the number of age-eligibles was used as a preliminary measure of size.

As described below, varying but roughly equal final measures of size were assigned to schools containing estimated age-eligible students ranging from 20 to 160 (for age 9) or to 200 (for ages 13 and 17). Schools with less than 20 estimated age-eligibles were selected with lower probabilities, and schools above the indicated maximum size were selected with probabilities proportional to the estimated numbers of age-eligible students.

With the adoption of these general specifications the sampling of schools by RTI proceeded approximately as follows:

- The estimated number of age-eligibles, E_i, was computed for school i, using QED information for school year 1982-83. The number in each grade was estimated by dividing total enrollment by the number of grades, and the number of age-eligibles was estimated by applying the RTI prediction formulas.[6]

- For the "big-city" PSU's

  - A SES index was assigned to each school (based on employment, unemployment, occupational, and income data from the 1980 Census for each Census tract, and by approximately matching the ZIP codes to the Census tracts).

  - Schools were classified as in low-SES, Stratum 1, and other, Stratum 2. After establishing a cut-off for the SES index to define the two strata, the schools were ordered by size (estimated number of age-eligibles) in ascending order in Stratum 1 and descending order in Stratum 2. For other PSU's the schools were ordered by size.

- A measure s_i' was assigned to each school, based on the estimated number of age-eligibles E_i, illustrated as follows for age 9, for which $\bar{n} = 20$ is the planned full-session size:

  - If school i had six or less estimated age-eligibles, $s_i' = .25$;

  - If school had seven to 19 estimated age-eligibles, $s_i' = E_i/20$;

  - If school had 20 or more age-eligibles but less than 160,

    $$s_i' = \frac{E_i}{20k_i}$$

    where k_i is the number of sessions of 20 that can be accommodated by E_i; and

  - If school had 160 or more age-eligibles

    $$s_i' = \frac{E_i}{160} .$$

- A final measure of size, s_i, was computed for each school by using $s_i = 2s_i'$ for those schools in "big-city" PSU's that had been assigned to the low-SES stratum, and by using $s_i = s_i'$ for all other schools. (We note that the extreme rural PSU's were already oversampled by a factor of 2, which had the effect of doubling the school sample in these.)

- The number of schools to be selected in an age class was computed separately for each PSU to yield approximately the desired number of students to be tested, after making approximate allowance for school and student nonresponse and for ineligible schools. The number of schools to be selected, t, is

  $$t = \frac{\bar{n}m}{\bar{k}}$$

  where

  $\bar{n}$ is the number of students per full age session (e.g., 20 for age 9),

  m is the number of full age-eligible sessions assigned to the PSU, and

  $$\bar{k} = \frac{\Sigma s_i' k_i}{\Sigma s_i'}$$

  i.e., the weighted average of the k_i (the number of age-eligible sessions available in school i, as used in computing the measures of size), and

  s_i' is defined above.

- The t schools were then selected in the PSU for the age class by sampling with probabilities proportionate to the final measures of size, s_i. It was recognized that a school might be selected twice for the same age class by this procedure, and thus (in order to avoid administering more than

10 sessions in a school) it might be necessary to transfer sessions to another sampled school. (Actually, only three schools were selected twice, and these were for age 17.)

A detailed description of the initial selection of the sample schools is given in the RTI Final Report cited earlier.

4. Updating the School Sample

ETS made the initial contacts with sampled school districts to obtain participation. The participating districts were then requested by Westat to identify schools that were new since the time of the QED list, or schools with changes in grade range or major changes in enrollment. These were given appropriate chances to be in the sample using probability-sampling procedures. Also, the sample was supplemented in a few PSU's where losses due to closed schools or other changes left too few schools in the sample. A Principal's Questionnaire showing updated grade and enrollment figures and certain other school characteristics was requested from each of the cooperating schools.

Some substitutions were made, as needed and to the extent feasible, for noncooperating schools. Generally, substitutions were made for schools refusing to participate in the assessments if their omissions would result in an unacceptable balance in school type among the schools assessed, according to the size of the school and the socioeconomic status of the community or would result in a substantial reduction in the number of students tested. In general, substitution of schools was made within the same PSU, but in a few cases losses in one PSU were compensated for by additional assessments in the sampled schools in another PSU. In three cases substitute schools were obtained from a neighboring and similar county (not a member of the primary sample of PSU's).

Table 1 summarizes the selection and participation of schools. The cooperation rates obtained were approximately the same as obtained for the Year 13 NAEP (an overall rate of 88.1 for Year 15 and of 88.0 for Year 13).

5. The Assignment of Sessions to Schools, by Type

The assignment of sessions to schools was done separately by the two types of sessions, designated "spiral" and "tape."

As discussed in the papers by Messick and Beaton, the balanced incomplete block (BIB) design together with spiraling (or interspersing) the assessment booklets was introduced into NAEP for the first time in Year 15. This made it possible to correlate results for all pairs of exercises in the BIB design. The exercises were divided into blocks of items, each block also containing some background questions. The blocks were assembled into 63 test booklets, most containing three blocks as well as a set of background questions common to all the booklets, so that each block occurred in the same number of booklets and also each pair of blocks occurred in the same number of booklets. As a result, it was expected that each block of items would be administered to about 2,600 students in each age class and each pair of blocks would be administered to about 280 students in each age class. The booklets were assembled systematically into packages, so arranged that the starting book was varied from session to session.

The tape design used an administration procedure like that of earlier rounds of NAEP, so as to provide direct comparison with the results of earlier rounds and to calibrate the results of the spiral design. The administration of each booklet utilized a tape recording, as in earlier rounds. The specified sample size was such that each tape-administered booklet was expected to be administered to about 1,250 students.

A preliminary allocation of sessions was made to the sampled schools based on the QED 1982-83 information on enrollment and grade range for use in making initial arrangements with the schools. These were revised later on the basis of the Principal's Questionnaire which provided enrollment by grade and information on SES status and minority enrollment for the school.

For the final allocation of sessions to schools, small schools were clustered with others in the sample so that there was an estimated minimum of eight (usually more) age-eligible students in each school cluster. The allocation of tape sessions was made first, by ordering the school clusters by an index of socio- economic status (based on the information provided in the Principal's Questionnaire) and by size and then selecting a systematic sample of four school clusters with probability proportional to the estimated number of available sessions for age-eligibles in the school cluster. The next step was to assign one spiral session to each school cluster not selected for a tape session and to allocate the balance of the spiral sessions specified for the PSU to school clusters proportionate to the estimated number of remaining sessions available (for students eligible by age or grade[7]).

6. The Samples of Students

A total of about 29,300 students was to be tested for each age class, including students for the corresponding modal grade. This means an average of about 460 completed assessments per PSU for each age class. On the basis of the experience in Year 13, conservative estimates were made of the proportion of students that would be excluded from testing because of language or other disability and of the proportion of students invited for assessment that would actually complete the assigned test. These estimates led to the determination of the sampling rate to be applied in each sample school. Since the estimates were conservative, the number of students assessed was expected to exceed the target. For age 9 and grade 4, about 31,700 students were assessed; for age 13 and grade 8, about 33,900 students were assessed; for age 17 and grade 11, about 35,200 students were assessed.

A Student Listing Form (SLF) was filled out for each participating school; all enrolled students of the specified age (9, 13, or 17) and all others in the corresponding modal grade (4, 8, or 11) were to be entered on the SLF in any order convenient for the school. In a few instances for very large schools, only a sample of students was listed on the SLF. The SLF was ordinarily prepared by the school, but Westat staff assisted or prepared the form as found desirable or necessary.

After the SLF was completed the selection of sample students was carried out briefly as follows:

- A computer generated listing of sample SLF line numbers was prepared in advance by Westat to identify the students to be included in the sample. When the number of students listed on the SLF was widely different from the anticipated number, communication was handled by telephone and a new set of sample line numbers was supplied.

- The sample line numbers also identified the particular session to which a sampled student was assigned, that is, whether spiral or a particular tape session.

- The names of students selected for the sample were reviewed by appropriate school personnel to identify sampled students who for language reasons or certain types of handicaps would be unable to take the test and thus should be excluded.

Make-up sessions were scheduled in schools in which the students assessed constituted less than 75 percent of the selected sample in the case of spiral sessions, less than 50 percent in the case of tape sessions for 9-year-olds and 13-year-olds, and less than 75 percent in the case of 17-year-olds. Very few make-up sessions were necessary for 9- and 13-year-olds. For the 17-year-olds, make-up sessions were conducted in about 20 percent of the sample schools.

7. Assignment of Weights for Estimation

7.1 Base Weights

The base weight assigned to a student is the reciprocal of the probability that the student is invited to a particular type of assessment session, i.e., a spiral session or a particular one of the four tape sessions. That probability is the product of:

1. The probability that the PSU is selected;

2. The conditional probability, given the PSU, that the school is a member of the sample selected by RTI or any supplementary sample selected by Westat;

3. The conditional probability, given the sample of schools in a PSU, that the school is allocated the specified type of session; and

4. The conditional probability within a school that the student is invited to the specified type of session.

The probabilities (1) and (2) were provided by RTI for each PSU and for each school originally selected by them. There were occasions, described in Section 3, where the school sample was modified and this affected the weights. When supplemental schools were selected in updating the school sample or to compensate for losses due to closed or ineligible schools, the weights of the originally selected schools were modified to reflect the modified sample size. When substitutions were made for refusals, the weights assigned to the substitute schools are what they would have been, had the schools been original selections.

The probability (3) was computed by determining all possible outcomes of the algorithm used by Westat to allocate the tape and spiral sessions to the schools selected for the sample in each PSU. Probability (4) is a function of the sampling intervals used to select individuals for testing from the list of eligible students provided by each cooperating school.

7.2 Adjustments for Nonresponse

School Nonresponse

Within each PSU, school nonresponse adjustment classes were defined on the basis of affiliation (i.e., public, parochial or private) and size, with no class having fewer than five schools. Consolidations of classes were often necessary.

For each class, the school nonresponse factor for spiraled assessments is

$$f_1 = \frac{\sum_{i \in A} W_i G_i}{\sum_{i \in B} W_i G_i}$$

where in the numerator the summation is over all schools in the original sample within the adjustment class (that is, including refusing and supplemental schools but excluding substitute schools), and the summation in the denominator is over the cooperating schools (including schools that were substituted for noncooperating schools). W_i denotes the school base weight (the reciprocal of the probability of selection of the school, conditional on the PSU). G_i denotes the number of grade- or age-eligible students estimated from QED data. This factor was used to adjust the school base weights, W_i, for school nonresponse. No school nonresponse adjustment was made for tape assessments since substitutions were made as necessary for noncooperating schools in order to achieve four taped administrations in each PSU.

Student Nonresponse

Factors for the adjustment of the base weights for student nonresponse were computed separately for students assigned to spiral sessions and for each type of tape session.

For spiral sessions the student nonresponse adjustment was made separately for students in or above the modal grade for his/her age, and for those below the modal grade for his/her age, for each PSU. The factor for an adjustment class was

$$f_{2s} = \frac{\Sigma u_i n_i}{\Sigma u_i n_{Ri}}.$$

Here, the summations are over the schools with students in an adjustment class and n_i and n_{Ri} denote respectively the number of students invited and the number responding, i.e., completing the assessment, in the adjustment class in school i. The weight, u_i, is the reciprocal of the probability of assignment of a student in school i to a spiral session, conditional on the PSU.

For each tape session, t, there is only one adjustment class per PSU. The adjustment factor is

$$f_{2t} = \frac{n_t}{n_{Rt}}$$

where n_t is the number of students that were invited to the particular tape session and n_{Rt} is the number who completed the assessment.

The student response rates separately for urban and rural type PSU's are given in Table 2. The overall student response rates were 92 percent for age 9 and grade 4, 90 percent for age 13 and grade 8, and 82 percent for age 17 and grade 11. These were higher than anticipated based on previous experience.

7.3 Variation in Weights

As mentioned earlier, the general goal was a design with uniform overall sampling fractions except for oversampling in certain types of areas or schools to improve estimates for certain subgroups. However, additional variation in weights arises from a number of factors, including especially the undersampling by a factor of four of schools with less than seven expected age-eligibles. Variation arises also from the use of the same PSU's for each age class, with selection probabilities proportionate to average measures of size, and with subsampling of an approximately constant number of students in an age class from each PSU. Also, the noncooperation of some schools for which substitutions were not made resulted in additional allocation of assessments to other schools, and the necessity to allocate separately tape and spiral sessions to schools introduced some added variation in probabilities of selection and weights. In addition, adjustment for nonresponse at the school and student levels, and certain other factors, added to variations in weights.

Such variability in weights contributes to the variance of overall estimates from the survey, approximately by a factor $F = 1 + V^2$, where V^2 denotes the relvariance of the student weights. For Age Class 13 in Year 15, for the spiral sample (before trimming and post-stratification, see below), F = 1.21. For the four tape assessments, the factors are respectively 1.25, 1.25, 1.31 and 1.21. These may be compared with the factors obtained in Year 13, which ranged from 1.13 to 1.96 for individual packages and averaged 1.30, 1.31 and 1.30 for Age Class 9, 13, and 17, respectively.

The use of post-stratification also adds, by design, a small amount to the variation in weights, but presumably reduces the variance of overall estimates because it reduces the variability in the sizes of subclasses that respond differently. For the spiral assessment for 13-year-olds, the factor F was increased from 1.21 to approximately 1.26 as a result of the post-stratification.

7.4 Trimming the Weights for Outliers

As in previous assessments, the weights for students with extremely large weights were reduced, i.e., trimmed, in order to reduce the effect of potentially extreme contributions of a few schools to any particular estimate. The trimming algorithm was similar, but not identical, to that used in earlier assessments and had the effect, approximately, of trimming the weight of any school that contributed more than a specified proportion, θ, to the variance of the estimated number of students in the spiral assessment, and in each of the four tape assessments. θ was set equal to 10 divided by the number of schools involved. For Age Class 13 this resulted in trimming the weight for three schools in the spiral assessment and for one school for a tape assessment.

7.5 Post-Stratification

The weights determined in the manner described above were adjusted by post-stratification in order to reduce the variance of estimates relating to student populations that spanned several subgroups. For this purpose, 39 subgroups were defined for each age class as the logical products of 13 subgroups (defined in terms of race, ethnicity, region and community size (SDOC)) and 3 subgroups (defined in terms of age and grade) as shown in Table 3.

For each of the 39 cells so defined, independent estimates of the number of students were made by Westat on the basis of data provided by the Bureau of the Census in the form of special tabulations of the education supplement of the Current Population Survey (CPS) for 1981 and 1982, together with the projections made by the Bureau of the Census of the population by single years of age for each year from 1981 to 1983. These estimates were then combined with the estimates yielded by NAEP, in a composite estimator in which the weights are inversely proportional to the approximate variances of the estimates from NAEP and the estimates made by Westat from the CPS and Census data. This was done separately for each of the three age classes.

The final weight for any student is then the product of the weight previously adjusted for nonresponse and a factor which is the ratio of the composite estimate of the number of students in the cell to which the student belongs to the NAEP estimate for the same cell.

8. Variance and Variance-Component Estimation

Variance estimates will be made using a "jackknife" procedure. Approximate estimates of total variances will be made by grouping the 64 primary sampling units into 32 pairs. A jackknife replicate will be formed by successively dropping one PSU from a pair, at random, and doubling the weight of the other PSU in that pair, and including all other PSU's. Thus 32 replicates will be identified. The post-stratification estimation procedure will be carried through separately for each replicate. The replicates also appropriately reflect the effect of the nonresponse adjustments since these were made within PSU's.

The variance of any estimate, $\bar{x}$, is then estimated by computing

$$s^2_{\bar{x}} = \sum_{h}^{32} (\bar{x}_h - \bar{x})^2$$

where $\bar{x}_h$ is the same statistic as $\bar{x}$ but computed for replicate h.

Components of variance will be estimated approximately by a similar procedure but by redefining the jackknife replicates. Thus, the schools in each PSU will be divided into two random half-samples. Sixty-four jackknife replicates will be identified from these 64 pairs, following the procedure described earlier for PSU's. The estimated variance of the statistic $\bar{x}$ computed from these 64 replicates, $s^2_{\bar{x}2}$, will reflect the variance contribution from sampling schools and from all subsequent stages of sampling, but not from sampling PSU's. Then $s^2_{\bar{x}} - s^2_{\bar{x}2}$ is an estimate of the PSU contribution to the variance of $\bar{x}$. In a similar manner contributions to variance of sampling from the subsequent stages of sampling can be estimated.

The estimates of variance will guide in improving future sample design and can, among other things, evaluate the effect of BIB spiral sampling as compared with assigning the same package to every student in a session.

We note that the BIB spiraling generally spreads a given question or set of questions across more sessions and more schools than results from the assignment of the same package of

exercises to each student in a session. As a consequence the sampling errors will be equal to or less than the sampling errors in an assessment in which all students in a session take the same package. Gains will depend on the definition of the statistic, $\bar{x}$. We will be able to estimate these gains, and the impact of alternative design decisions for future surveys, from the variance-component analyses. Our advance speculations made in designing the Year 15 sample were that the relative variance reductions from BIB spiraling might be of the order of 20 to 25 percent.

9. The Sample of Excluded Students

After preparing the list containing the names of the age- or grade-eligible students enrolled in a school, each participating school was asked to review the list and to decide who, in the school's judgment, was to be excluded. Students who were non-English-speaking, educable mentally retarded, or functionally disabled were to be excluded from the assessment. Students were not to be excluded merely because of poor academic performance or normal discipline problems. Excluded students were sampled at the same rates as any other eligible student but were excluded from testing. Instead, a special excluded student questionnaire focusing on the nature of the student's problem and the school's approach to handling it was to be completed by the school. About 4 percent of the eligible students in age class 13 were excluded. In Year 13 about 5 percent of the 9 and 13 year-olds and about 3-1/2 percent of the 17-year-olds were excluded.

For age 13 a total of approximately 1,500 students were excluded. The distribution of reasons for exclusion is given below:

	Percent
Physical or mental handicap	67
Behavior disorder	7
Handicap and limited English proficiency	6
Limited English proficiency	20
	100

10. The Associated Student-Teacher Sample

In each sample school, one or more of the students selected for spiral assessment was subsampled and a teacher questionnaire was obtained for the principal English or language arts teacher of each of these students. For this purpose one student was to be selected at random from each spiral session. The principal English or language arts teacher was also recorded for each student who participated in the spiral assessment, and the characteristics of the surveyed teacher are associated with each spiral-assessed student of that teacher. Thus, the set of students for whom teacher characteristics were obtained is a probability subsample of the student sample selected for spiral assessment.

The conditional probability that a spiral assessment selected student (of teacher k) had his teacher in the survey is given by

$$P_k = 1 - \frac{\binom{n-n_k}{t}}{\binom{n}{t}}$$

where the symbol $\binom{a}{b}$ denotes the number of combinations of a things taken b at a time, and

- n = the total number of students invited to spiral assessments;
- n_k = the number of students invited to spiral assessments whose teacher is the k^{th} teacher of the school, k=1, 2, ..., K, and
- t = the number of spiral-invited students subsampled for the teacher survey.

Since the principal teacher was recorded only for assessed students, P_k was approximated by replacing n_k and n by the numbers of assessed rather than invited students. Students whose teachers were surveyed have their weights multiplied by the reciprocal of P_k in any analyses that involve relating teacher characteristics to student characteristics. The weights were further adjusted for nonresponse, within PSU's, to account for the fact that not all assessed students indicated their principal language arts teacher and not all sampled teachers returned a completed questionnaire. They were also adjusted within PSU's by a first-stage post-stratification procedure so that the sum of the weights for students in the teacher sample were equal to the sum of the weights for all students in the spiral sample. A final post-stratification adjustment was then applied as described in Section 8 above for the full student samples.

11. Control Activities

Quality control field visits were conducted by Westat and ETS at a sample of 64 schools: 32 schools for 13-year-olds, 12 for 9-year-olds and 20 for 17-year-olds. Both purposive samples and probability samples were selected for this purpose. Purposive samples were drawn from schools in the first set of PSU's worked by each supervisor to

identify and correct, as soon as possible, problems in within-school sampling operations and other assessment activities. Probability samples were drawn to represent the remaining schools.

In addition, during the within-school sampling of students Westat's sampling staff monitored student sample yield and supported the field staff on resolution of sampling problems.

Weighting procedures and overall results were evaluated for each age class and type of session by comparing individual school and PSU estimates to expected figures. Also, estimates of total numbers of students from spiral session, overall and by PSU and school, were compared to estimates from tape sessions, and to expected sample sizes, and variability in weights and the magnitudes of school and student nonresponse adjustments were examined. A few problems were identified and corrective action was taken.

12. Planning for Future Assessments

It has always been true that planning for a future assessment must begin before the current assessment is completed. Consideration is being given to assessment in the Spring for all ages, instead of Fall for age 9, Winter for age 13, and Spring for age 17. Also, there are potentials for participation in the assessment by a number of states, and there are discussions of the desirability and feasibility of oversampling for blacks and other minorities. These and other possible changes may influence the sample design, including the definition and selection of PSU's and schools. Additional work on evaluation of unit costs and variance components may also contribute to design modifications. Consideration is being given to such issues in the limited time available before decisions must be made to proceed on the sample design for Year 17 assessments.

Footnotes

1. See Final Report on National Assessment of Educational Progress: Sampling, Weighting, and Quality Check Activities for Assessment Year 13. June 1983 (RTI/1967/00-02F).

2. For a detailed description of the selection of PSU's, see the RTI Final Report (RTI/2589/03-00F) entitled, "Primary Sample for Years 15-19 of the National Assessment of Educational Progress."

3. Quality Education Data, Inc. (QED) maintains and updates annually lists of schools showing, for each school, the grade span, total enrollment, school district, principal's name, and other information. The initial data provided by QED were evaluated against Census school-enrollment data by RTI, which led to some corrections of the QED file, made before the data were used in computing measures of size for sampling.

4. Chromy, James R., "Sequential Sample Selection Methods," Proceedings of the Section on Survey Research Methods, 1979, American Statistical Association, pp. 401-406.

5. Three schools were selected twice for age 17 and grade 11.

6. See Section 3.1.4 of School Sampling Procedure for Year 15 of the National Assessment of Educational Progress, September 1983 (RTI/2589/02-00F).

7. The final report on sampling will give the details of the allocation.

Table 1. Summary of NAEP Year 15 school participation experience

	Age 9/ grade 4	Age 13/ grade 8	Age 17/ grade 11	Total sample
Initially selected schools	700	588	394	1,682
Supplemental selections	17	2	1	20
New schools added	2	1		3
Total original sample	719	591	395	1,705
Out-of-range or closed (A)	15	12	17	44
No eligibles enrolled (B)	17	64	17	98
District refused (C)	61	42	40	143
School refused (D)	19	14	21	54
Cooperating - No student sample (F)	0	4	1	5
Cooperating - Assessment conducted (E)	607	455	299	1,361
Cooperation rate = $\frac{B+E+F}{B+C+D+E+F}$	88.6	90.3	83.9	88.1
(Year 13)	(88.0)	(89.2)	(86.5)	(88.0)
Replacement for refusals	67	28	34	129
Out-of-range or closed	3	0	0	3
No eligibles enrolled	5	3	1	9
Refusals	5	2	6	13
Assessment conducted	54	23	27	104
Total contacted schools	786	619	429	1,834
Total assessments conducted	661	478	326	1,465

Table 2. Preliminary student response rates

Age -- grade	Session type	Urban PSU's*			Rural PSU's**			Total assessed
		Number invited	Number assessed	Response rate	Number invited	Number assessed	Response rate	
9 -- 4	spiral	19,946	18,334	91.9%	8,324	7,829	94.1%	26,163
	tape	4,304	3,943	91.6%	1,680	1,580	94.0%	5,523
								31,686
13 -- 8	spiral	21,959	19,722	89.8%	9,660	8,899	92.1%	28,621
	tape	4,163	3,653	87.7%	1,745	1,612	92.4%	5,265
								33,886
17 -- 11	spiral	26,096	21,159	81.1%	8,888	7,768	87.4%	28,927
	tape	5,661	4,423	78.1%	2,105	1,805	85.7%	6,228
								35,155
Total								100,727

*SDOC's 1, 2, and 3 as defined in Section 2

**SDOC's 4 and 5 as defined in Section 2

Table 3. Definition of subgroups used in post-stratification

Subgroup	Race	Ethnicity	Region	SDOC
1	White	Non-Hispanic	1	1, 2
2	White	Non-Hispanic	1	3, 4, 5
3	White	Non-Hispanic	2, 3	1, 2
4	White	Non-Hispanic	2, 3	3
5	White	Non-Hispanic	2, 3	4, 5
6	White	Non-Hispanic	4	1, 2
7	White	Non-Hispanic	4	3, 4, 5
8	Any	Hispanic	1, 2, 3	Any
9	Any	Hispanic	4	Any
10	Black	Non-Hispanic	1	Any
11	Black	Non-Hispanic	2	Any
12	Black	Non-Hispanic	3, 4	Any
13	Other	Non-Hispanic	1, 2, 3, 4	Any

Subgroup	Eligibility status of student
1	Age and grade
2	Age only
3	Grade only