# The New Design for the National Assessment of Educational Progress

Samuel Messick
Educational Testing Service

Established in the 1960s to assess the condition and progress of education in the United States, the National Assessment of Educational Progress (NAEP) collected data for the first time in 1969. Since then, over a million 9-, 13-, and 17-year old students, as well as occasional samples of adults and of 17-year olds who were not in school, have been assessed in a variety of subject-matter areas such as reading, writing, mathematics, science, social studies, literature, music, and art. At the outset, NAEP was profoundly colored by the political climate of the time, especially by the fearfulness expressed by states and school districts that they might be compared and evaluated by means of NAEP, with its potential overtones of national curriculum and national testing.

In light of such concerns, the original NAEP architects developed a sampling plan insuring that accurate results could not be reported at the state or district level. They espoused matrix sampling procedures insuring that no student would take more than a small sample of diverse exercises, so there would be no tests or test scores in the traditional sense and certainly no test scores for any individuals. They capitalized on the strengths of matrix sampling to insure comprehensive coverage of subject matter for population groups, thereby generating sets of objectives and exercises that reflected salient features of most extant curricula. They insisted on analysis and reporting at the exercise level, so that the focus would be not on curriculum units or knowledge and skill domains, but on specific learning outcomes whose nature and importance could presumably be directly judged by laymen and professionals alike. Finally, the assessment was organized in terms of age levels rather than grade levels, which -- while having a number of important points in its favor -- has the consequence of severing NAEP results from the major way in which schools are organized, state and local assessments are reported, and educational policies are formulated.

In addition to the educational disjunction inherent in age sampling as opposed to grade sampling, the usefulness of NAEP results was further attenuated by two critical problems with the original design -- namely, problems of interpretability and comparability of the findings. Interpretations of differences in percent correct on single exercises or in average percent correct on composites of exercises, both in making group comparisons and in establishing trends over time, are complicated or equivocal in the absence of evidence about the coherence and meaning or construct validity of the measures. Worse still, in the original design there was no means of assuring comparability in the meaning of performance across exercises within subject areas and, of prime importance, comparability across different age levels and time periods. Unfortunately, these problems were further compounded by an insidious erosion of funding that led first to the elimination of adult and out-of-school 17-year old samples, then to reduction in the number of subject areas assessed annually, and finally to the forgoing of data collection in some years. As a consequence, the timely scheduling of subject reassessments was disrupted and the timeliness and relevance of the findings reduced. Thus, the challenge for the new design was to improve the timeliness, interpretability, comparability, and policy relevance of NAEP results -- all within limited and fixed budgets.

## Features of the New NAEP Design

Among the many features of the new NAEP design, the following will be briefly discussed: changes in scheduling to improve the timeliness and, through cohort matching, the orderliness of results; the adoption of new data collection methods that permit estimation of correlations, not only among exercises but between exercises and background and program variables, to enhance the interpretability of findings; the introduction of psychometric scaling methods to increase the comparability of performance measures across groups, age levels, and time periods; the expansion of sampling procedures to improve efficiency, representativeness, and school-relatedness of the sample; the reliance on equating samples to link future data collected by the proposed procedures to past data obtained by the original procedures in order to maintain the integrity of trend analyses; and, the elaboration of both the number and variety of background and program variables to facilitate policy-relevant analyses (Messick, Beaton, & Lord, 1983).

### Timeliness and Cohort Matching

The new design retains the previous cyclical scheduling of subject area data collection but changes to a planned schedule of biennial assessment for reasons of cost. At the same time, beginning in school year 1985-86, four subject matter fields are to be included in each assessment wave to insure extensive and timely curriculum coverage. The assessment of reading is introduced into every biennial wave so as to increase the timeliness of information in this basic area as well as to calibrate different cohorts at each age level. One of the major reasons that NAEP has not become a truly useful indicator of educational progress is that assorted subject-matter assessment cycles of three to nine years, which have been characteristic of NAEP in the past, are too infrequent and sporadic either to keep pace with educational change or to keep the public's attention. Worse still, the schedule of subject-matter assessments did not systematically track the student cohorts as they moved through the age levels used in sampling and reporting, so that cohort differences were confounded with educational and social change.

With respect to cohort differences, if a given subject area were assessed in four-year cycles, then the current sample of 17-year olds assessed in mathematics, for example, would be from the same birth cohort as the sample of 13-year olds assessed in math four years earlier and as the sample of 9-year olds assessed in math eight years earlier. By thus matching the assessment intervals to the number of years intervening between the age levels sampled, cohort differences in a given subject area are essentially controlled in the new design and interpretations of trend analyses become both simpler and more powerful. To achieve these benefits, however, the three age levels must be defined in comparable fashion, which unfortunately has not been the case until now. Formerly, ages 9 and 13 were defined by calendar year, but age 17 was defined by birth during the period October 1 through September 30. In the new plan, as of the 1985-86 assessment, all three age levels are defined by the October 1 to September 30 interval, both to attain comparability and to link the birth cohorts more closely to school entrance age requirements. Furthermore, each age level had formerly been assessed at a different time during the school year -- fall for age 13, winter for age 9, and spring for age 17. The new plan calls for all three ages to be assessed in the spring, not only to eliminate gross variation in time of testing but to coordinate the assessment with near completion of the curriculum year.

## BIB Spiralling and the Correlational Basis for Interpretability

Interpretability of findings has been a chronic problem in NAEP as originally implemented because the intended benefits of exercise-level reporting were simply not realized -- namely, the futile hope that the specific learning outcomes embodied in a discrete exercise readily conveyed its own criterion-referenced standard and that a direct link could be easily perceived between the exercise and the educational objective it represented. On the one hand, discrete exercises may often be interpreted to reflect multiple objectives and, on the other hand, it is a rare educational objective of any importance that can be fully captured in a single instance of behavior. This limitation was eventually addressed by also reporting average percent correct on aggregations of exercises presumed to reflect the same educational objective or performance dimension. But these composites were determined on the basis of educators' judgments and may or may not be supported empirically in terms of student performance consistencies. What is needed is not only a means of justifying judgmental exercise composites in terms of student performance consistencies, but of empirically determining the aggregations of exercises that best reflect existing performance consistencies of educational import. The critical requirement for accomplishing this is to be able to estimate the intercorrelations among the exercises as well as between exercises and other variables.

The standard matrix sampling procedure formerly employed by NAEP divided the total assessment battery, which would typically require six to seven hours to complete if it were administered to anyone, into mutually exclusive booklets, each of which was allocated about a 45-minute allotment of exercises. Since no student was administered more than one booklet, this simple matrix design allowed calculation of correlations and cross-tabulations among exercises within the same booklet but not among exercises in different booklets. The new NAEP design remedies this deficiency by using a powerful variant of matrix sampling called Balanced Incomplete Block (BIB) Spiralling. With this procedure, the total assessment battery is divided into blocks of, say, 14 minutes each, and each student is administered a booklet containing three blocks as well as a six-minute block common to all students. Thus, the total assessment time for each student is still about three-quarters of an hour.

The balanced incomplete block part of the method assigns blocks of exercises to booklets in such a way that each block appears in the same number of booklets and each pair of blocks appears in at least one booklet. This generates a large number of different booklets. The spiralling part of the method then cycles the booklets for administration so that typically no two students in any assessment session in a school, and at most only a few students in schools with multiple sessions, receive the same booklet. At each age level, each block of exercises is administered to approximately 2,000 students and each pair of blocks to a smaller number depending upon the particular BIB design.

With BIB spiralling, correlations may be calculated among all exercises, whether in the same booklet or different booklets, on some subset of students, although different correlations will be based on different student subsamples. This permits estimation of the complete matrix of correlations among exercises within a subject area and the subsequent empirical mapping of the structure of achievement in that domain. Since different exercise blocks may derive from different subject-matter areas, BIB spiralling can yield correlations among exercises not only within subject areas but across subject areas as well. This permits examination of cross-area linkages and the tracing of possible facilitating processes from one area to another.

Furthermore, since two minutes of each block are currently allocated to background and attitude items, as is the common block taken by all students, BIB spiralling yields correlations between educational performance and a host of background, attitudinal, and program variables. BIB spiralling is also statistically more efficient than ordinary matrix sampling for some estimates. By administering more different exercises within a particular school and by administering a particular exercise in more different schools, the school clustering effect is reduced and the BIB sampling design is consequently more efficient. However, although balanced incomplete block designs are typically implemented within subject-matter areas, feasible designs cutting across several subject areas are often partially balanced.

## Item-Response Theory and the Quest for Comparability

Comparability of findings has also been a chronic problem in NAEP ever since its inception. Since many factors affect percent success on a given exercise, the measurement of change in terms of either single exercises or composites of exercises is inherently difficult to interpret. A key problem is that the relationships between percentages and quantitative variables such as those descriptive of background or program characteristics are typically nonlinear, so interpretations of the meaning or sources of percentage change are often either misleading or abstruse. This difficulty may be overcome by employing a statistical scaling model such as Item-Response Theory (IRT) that transforms percent correct (P) to a logit scale ($\log\frac{P}{1-P}$) to define latent continua (i.e., ability or performance dimensions) which are typically linearly related to other quantitative variables (Lord, 1980).

Item-response theory defines the probability of answering an exercise correctly as a mathematical function of ability level or skill. The particular mathematical function most widely used, the logistic function, has one parameter for each individual -- namely, ability or proficiency level -- and from one to three parameters characterizing each exercise (Lord, 1980; Lord & Novick, 1968). The item parameters reflect difficulty level, discriminating power, and likelihood of guessing. The model involving three item parameters is used in NAEP because the one- and two-parameter versions do not adequately cope with the realities of exercise variation. Item-response models postulate that the probability of success depends on the difference between the respondent's proficiency level and the item's difficulty level -- as weighted by item discriminating power and adjusted for guessing -- and on nothing else. Thus, item-response models apply only to unidimensional subject areas or subareas assessed by sets of exercises that all reflect a single underlying ability or skill, in the sense that only one dimension of response consistency contributes to systematic variations in item difficulty. With item intercorrelations available by virtue of BIB spiralling, this important model requirement of unidimensionality can be empirically evaluated for NAEP data via factor analysis and other techniques of multivariate analysis. This is especially important if a subject area such as science proves to be multidimensional overall but is comprised of some unidimensional subareas that could be identified for separate scaling.

One of the critical properties of item-response scaling is that item parameters are invariant across groups of examinees, while at the same time estimates of examinee proficiency levels are invariant across sets of items measuring the same ability or skill. Thus, IRT analyses yield a common scale on which group performance may be estimated and meaningfully compared for any group or subgroup, even though all respondents did not take all of the NAEP exercises in a subject area. Furthermore, since many of the same exercises are administered to the different age levels and in different assessment years, a common scale may be established, if the model fits, across age levels as well as across time (Lord, 1980).

By virtue of the invariance both of item parameters across respondent groups and of respondents' skill levels across calibrated exercises, not only may each student's skill level be estimated from any subset of calibrated exercises but exercises may be added or retired from the assessment at any time without affecting comparability of results. Moreover, since the skill scales are unbounded, they are not warped by floor and ceiling effects in the way that percentages and total scores are, so they tend to be more linearly related to other quantitative variables.

## Sampling to Improve Coverage and Relevance

The new NAEP design retains the previous deeply stratified three-stage sampling plan but as modified to meet some new purposes in addition to the old. The first stage of sampling entails classifying the primary sampling units or PSUs into strata defined by geographic region and communify type. The PSUs are typically counties, but small counties are aggregated so that no PSU has fewer than an estimated 1,000 students at each assessment age. For each age level, the second stage entails enumerating, stratifying, and selecting schools, both public and private, within each PSU selected at the first stage. The third stage involves randomly selecting students within a school for participation in NAEP, with the proviso that students judged to be untestable by current NAEP procedures were excluded from the sample -- those excluded being primarily functionally disabled and limited-English proficient pupils. Originally, samples of adults 26 to 35 years of age, as well as of 17-year olds who were not in school, were also located via household surveys and administered one or more NAEP booklets. However, limited funding has led to the elimination of this important feature.

The new NAEP plan reintroduces an adult sample; improves the representativeness of the sample of Hispanics in terms of the major cultural subgroups of Puerto Rican, Cuban, and Mexican Americans; documents the extent and nature of sample exclusions; includes a sample of teachers of sampled students to permit correlating teacher and student characteristics; and, undertakes sampling by grade as well as by age. The adult sample consists of 21- to 25-year olds assessed by means of a household survey in a special study of adult literacy, which it is hoped will become a recurrent feature of NAEP; any out-of-school 17-year olds identified in the household survey will also be assessed. The representativeness of the sample of Hispanics will be improved probably by increasing the number of PSUs overall as well as by oversampling. Functionally-disabled and limited-English proficient students who were often excluded from past samples are now included in the sampling frame and excluded only if selected for the sample, at which point an extensive questionnaire is completed by school personnel on the excluded students' characteristics and programs. This yields a significant body of data on a national probability sample of students deemed by their schools to be untestable by current NAEP procedures.

The teacher sample consists of a random sample of teachers of the assessed students, one teacher of English or language arts being selected for each session in the 1983-84 assessment of reading and writing and probably up to four teachers being selected for each session in the 1985-86 assessment of reading, mathematics, science, and computer competence. The selected teachers are administered a questionnaire covering background, education, and training; chacteristics of the instructional program; and, teacher perceptions of the school and its curricula.

Lastly and most importantly, the new NAEP plan entails sampling by grade level as well as by age level. Specifically, national samples are drawn of the modal grades in which most 9-, 13-, and 17-year old students fall; formerly these were grades 4, 8, and 11 but with the revised age definition, the modal grades are 3, 7, and 11. Even though the meaning of grade level varies in different parts of the country depending on the age at which children are admitted to school and on the advancement and retention policies of local school systems, this important step is taken to link NAEP results more directly to school practices, state and local assessments, and educational policies, which are all typically tied to grade levels.

But it should be noted that grade sampling is not undertaken at the expense of eliminating age sampling because there are also important reasons for sampling by age, not the least of which are that age has a common meaning across geographic regions and school practices and that age sampling retains comparability with past assessment data. Another critical reason for not relying on grade sampling alone is that many disadvantaged students are overage for their grade placement, which would seriously distort the meaning of average grade-level performance and seriously compromise the interpretation of grade trends as indicators of educational "progress" for key subgroups. The price paid for this two-way sampling is that approximately 30 percent more students are required for national samples of both modal grades and ages, which brings the sample size per block to approximately 2,600 students at each age-grade combination.

### Trade Offs in Statistical Bridges to Past Data

The power of BIB spiralling -- which brings with it sampling efficiencies, the computation of intercorrelations among exercises, and the application of IRT scaling -- is bought at the expense of an important data collection procedure. In past assessments, exercises were aurally presented and paced using a tape recorder. This is not feasible with BIB spiralling because students in each session are assessed on many different booklets. The forgoing of aural presentation is potentially important because poor readers, whether from disadvantaged minority groups or not, tend to perform somewhat better with aural as well as printed presentation, while good readers appear not to be unduly distracted on the average. On the other hand, state and local assessments, to our knowledge, have rarely if ever adopted aural presentation procedures and it is doubtful that many will. This renders previous NAEP procedures, and hence NAEP results, noncomparable to the mainstream of educational measurement practice, at least for poor readers.

In any event, since the same exercise presented by printed page alone will probably exhibit different response properties than when presented aurally as well, past NAEP results cannot be expected to be strictly comparable to those obtained in the redesigned NAEP with tape presentation eliminated. For this reason, the new NAEP design incorporates equating samples for each subject area so that statistical bridges may be established to the past data in each area, thereby maintaining the capability for continued trend analyses. This equating strategy requires that data be collected on some student samples by the past method and on other random samples by the new method during the same assessment wave in each affected subject area. Similar bridge samples are also employed when other substantial changes in data collection methodology are introduced, as when time of testing is shifted uniformly to the spring and age definitions are revised. The assignment of bridge sessions and primary assessment sessions to sampled schools interposes an additional stage in the sampling plan between the selection of schools and the selection of students within schools.

### Expanding the Measurement Base and Enhancing Policy Relevance

An important consequence of BIB spiralling is that NAEP exercises and composites may be correlated with any of the background and attitude items that are spiralled into the student booklets (or are taken in common by all students) as well as with teacher, school, and program variables that are tied to the students via the teacher and school questionnaires, school records, or other means. These background and program variables may also be used to generate group comparisons, such as students in public versus private schools or language minority students in bilingual programs versus those who are not. Given the availability of other background variables characterizing the groups in question, such group comparisons may also be conducted controlling for a variety of demographic, home, and school factors by means of analysis of covariance techniques.

The only limitation on the number and nature of educational and policy questions that can be addressed in this fashion is whether or not relevant background and program variables are included in the student, teacher, and school questionnaires or are derivable from other sources. The opportunity to elicit student information bearing on policy issues is greatly amplified by means of BIB spiralling -- as an instance, 351 background and attitude items were administered to the 13-year olds in the 1983-84 assessment. These student questions covered demographic characteristics and home environment; educational background and current practices; exposure to courses and computers; use of time both in school and out; and, orientation toward school, studying, and subject matters.

In addition to the teacher questionnaire previously described, extensive contextual data also derive from a school questionnaire covering characteristics of the principal, staff, and student body; of standards, programs, and computers; and, of school climate, finances, and resources. Thus, the new NAEP affords ample opportunity to examine the background and program correlates of student educational performance in assorted educational contexts in relation to a variety of policy issues.

### Enriched Data Analysis Capabilities

The introduction of BIB spiralling into data collection has profound implications for data analysis. To begin with, the availability of covariances among exercises provides a number of immediate benefits. First, it contributes to construct validation (Cronbach, 1971; Messick, 1975, 1980) in that the coherence of exercises designed to measure the same educational objectives can be empirically evaluated, as can the degree to which an exercise relates to other objectives for which it was not intended. A second benefit of covariances is that by identifying exercises that assess the same objective or performance dimension regardless of exercise format or content, the generalizability of process interpretations receives some empirical grounding. A third benefit is economy of measurement. By empirically grouping sets of exercises that reliably and sensibly assess a common dimension or objective, composite scores can be used which entail smaller sampling errors. In short, covariances provide an empirically-grounded conceptual basis for establishing meaningful and efficient scales.

Moreover, the entire matrix of intercorrelations, or selected submatrices, can be analyzed by such multivariate techniques as metric and nonmetric factor analysis and multidimensional scaling to ascertain the structure of educational achievement in a subject area, taking adequate account of the potential problems created when different covariances in the matrix are based on different random samples of individuals. One may also inquire whether performance dimensions have the same meaning and are measured with the same precision in different population groups such as males and females or blacks and whites. This may be accomplished through the application of confirmatory factor analysis of covariance structures in different groups of the same age to see if the same number of dimensions emerge in each group and if they are interrelated in the same way (Jöreskog & Sörbom, 1979). This is an important point because the interpretation of group differences in mean level of performance depends upon common covariance structures. Similarity or difference of covariance structures in different age groups may also be analyzed in the same manner. Of particular concern in age-group comparisons is the possibility of developmental trends not only in mean level of performance but in the degree of differentiation and integration of the skill dimensions at different age levels -- that is, age differences in the factor-score variance-covariance matrices. We may also inquire whether age-related differences in factor intercorrelations occur in the same way, for example, in all sex and age groups. Again, any obtained age-group differences in the number and nature of underlying dimensions have critical implications for the interpretation of mean differences between the age groups, because this would imply that the same dimensions are not being measured or not being measured in the same way at different ages. Finally, with the variety of background, attitude, home, school, and program variables available in the new NAEP, powerful structural equation or path models of educational attainment may be formulated and tested (Jöreskog & Sörbom, 1979; Bentler, 1980).

## References

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. Annual Review of Psychology, 31, 419-456.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike, Educational measurement(2nd ed.). Washington, DC: American Council on Education.

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V, estimation of linear structural equation systems by maximum likelihood methods: A program. Chicago, IL: National Educational Resources.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21, 215-237.

Messick, S., Beaton, A., & Lord, F. (1983). National Assessment of Educational Progress reconsidered: A new design for a new era (NAEP Report 83-1). Princeton, NJ: National Assessment of Educational Progress.