DOCUMENT RESUME

ED 279 664 TM 870 051

AUTHOR Bock, R. Darrell
TITLE Designing the National Assessment of Educational Progress to Serve a Wider Community of Users: A Position Paper.
PUB DATE Aug 86
NOTE 32p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Educational Assessment; Educational Research; Educational Testing; Elementary Secondary Education; *Evaluation Utilization; Measurement Techniques; *National Surveys; Policy Formation; *Program Design; Research Design; Scores; Test Construction; *Testing Programs; *Test Results
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

Efforts have been made to increase the dissemination and use of data generated by the National Assessment of Educational Progress (NAEP). Potential users include those concerned with curriculum and methods evaluation, public policymakers, and researchers. NAEP can provide data for curriculum evaluation, including item analysis data which assist in item selection for student testing. Census-like reporting is useful in policy formulation, but this type of information requires estimation of attainment levels in individual pupils and not merely the aggregated responses of groups. There have been problems associated with use of NAEP data in educational research; for example, matrix sampling data were previously difficult to handle. Currently, individual scores are being provided in ways which can be analyzed by conventional statistical methods. It has also been difficult to use school and classroom data as the unit of analysis. A NAEP design which would be suitable for hierarchical regression analysis would expand the possibilities for secondary analysis. Item response theory scaling currently assists in longitudinal analysis of NAEP data. Ways to increase use of the data include the duplex design and rotation sampling of schools, resulting in more frequent sampling. Cooperation with state testing programs can help in providing comparable scores. (GDC)

Designing the National Assessment of Educational Progress To Serve a Wider Community of Users

A Position Paper

R. Darrell Bock

University of Chicago

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

Designing the National Assessment of Educational Progress to Serve a Wider Community of Users

A Position Paper

R. Darrell Bock
The University of Chicago

August, 1986

In the 1960's, Ralph Tyler first advanced the idea that the nation should have an accurate, continuing measure of the productivity of its collective educational effort. He saw in the rapidly developing fields of educational testing, survey sampling theory and computer data processing, the potential for measuring the educational productivity nationwide at no greater cost than was being expended on indices of economic production. Through his efforts, and those of many others, the National Assessment of Educational Progress (NAEP) was created, and the nation wide-survey of educational outcomes that he envisioned became a reality. Now, mandated by federal law, NAEP has completed fifteen years of continuous and comparable measurement of atttainment in subject-matter areas including reading, mathematics, science and writing. Less frequent data are available in areas such as literature, social studies, art, music, and career development.

Although it offers the only dependable national index for monitoring the performance of our schools, NAEP has not, over these years, gained the recognition it deserves, either among professional educators or with the general public. Much less defensible statistics such as state averages of the annual Scholastic Aptitude Test (SAT) scores or those of the American College Testing Program

(ACT), receive more attention from the media and are far more widely known. Surveys of much more limited subject matter, such as the science and mathematics studies of the International Association for the Evaluation of Educational Achievement (IEA), are more prominent in the educational literature and probably have had more influence on current concepts and research in curriculum and instruction than has NAEP.

That the impact of NAEP under the auspices of the Education Commission for the States (ECS) was indeed limited and specialized was documented in a study by Sebring and Boruch (1982). They found that assessment results were used by state education agencies (in 12 states) primarily for the purposes of test development and curriculum design. There was only occasional secondary analysis of NAEP data by educational, psychological or sociological research workers. One of the few studies that made substantial use of NAEP archival data files was an effort by Harnischfeger, Huckins and Wiley (1977) to link state tests results in order to obtain a measure on which to base Title I awards.

* * * * * * * * *

Having been a student of Professor Tyler's, and having served on both the Analysis Advisory Committee of NAEP and the Technical Advisory Committee of the California Assessment Program (CAP), I have followed with interest the development of the accountability and assessment movements in education from their beginnings. I witnessed the frustration the NAEP organization experienced in attempting to impress the Assessment on the national consciousness and to find an audience among professional educators. My involvement with the assessment movement has led me, with the collaboration of colleagues, to examine the discrepancy between actual and potential roles of assessment programs in the conduct of education, and to consider how it could be attenuated. The result of one such study, based on experience with

CAP, appears in Bock, Mislevy and Woodson (1982); another is in a paper recently completed for the NIE Center for Student Testing, Evaluation, and Standards (Bock & Mislevy, 1986). In the present paper, I review the conclusions reached in these and other inquiries into the purposes and design of educational assessment.

1 THE USERS AND USES OF INFORMATION ON EDUCATIONAL ATTAINMENT

The limited response of the public and professional educators to the NAEP program was, no doubt, due in part to poor dissemination efforts and to design features that made effective presentation difficult. An early mistake was the decision to test in the same subject-matter area only once in every four years, rotating through the areas of reading, writing mathematics and science, and including occasional assessemnts of special topics such as music and art. This arrangement made it impossible to conduct discussions of achievement in each area annually at a fixed time the media could anticipate. Moreover, the decision to present results in the form of individual item statistics or average percent correct for subject-matter areas was unfortunate. These indices were not easily comprehended by a general audience, nor were they suitable for policy decisions. Even the so-called "user tapes", intended to make the NAEP archives available for secondary analysis, were unsuccessful. They were too complex to be readily used for secondary analysis by any but advanced practitioners of computer data processing (among the few productive uses were Haertel, Walberg, Jonker & Pascarella, 1981, and Mislevy, Reiser & Zimowski, 1981.)

But a more fundamental reason for the limited currency of NAEP data was that a careful analysis of the possible uses of the results, coupled with an effort to design the assessment so as to serve the widest possible community of users, was never attempted. There was, in effect, nothing resembling a market analysis and survey to identify the constituencies that would ultimately

share in the benefits of information on educational outcomes. To date, this sort of analysis still has not been attempted on a national level.

In the context of state assessment programs, however, Bock and Mislevy (1986) have carried out such an analysis as a prologue to a proposed design for comprehensive educational assessment in the states. Much of their work is also relevant to national assessment: they begin by identifying seven main categories of users of state testing programs, namely:

1) teachers, school counselors, parents, and students.

2) curriculum and instruction specialists.

3) local school system managers, officers, and boards.

4) state departments of education.

5) state legislators and officials.

6) the media and the public.

7) educational research specialists.

Two of these categories—the first and the third—are concerned with local issues of student guidance and school mangagement, and are thus outside the scope of national assessment. In addition, these users require data on all students within their purview and cannot employ the kind of sample survey data on which a national assessment is necessarily based.

For the remaining five categories, however, the NAEP data have a potential relevance. Broadly, the users in these categories fall into three main classes:

1) those concerned with the evaluation of curricula and methods.

2) those dealing with public policy in education.

3) those engaged in substantive research based of secondary analyses on assessment data.

Briefly, we may describe the activities of these three broad classes of users as *evaluation, policy formulation*, and *research.* Let us examine in more detail the themes of these activities, the roles that assessment data can play in them, and the forms of data that best fill these roles.

1.1 EVALUATION

The work of educational evaluation is carried out primarily by professional curriculum specialists, some of whom are members of subject-matter departments in schools of education (e.g., mathematics education, reading, English language and literature, science education, social studies), while others are affiliated with state departments of education, textbook publishers, or educational testing organizations, or are employed in local school systems. These workers are concerned with the many detailed topics, facts, concepts, and skills that make up their respective subject-matter fields. What they want to learn from NAEP data is the suitability of these many curricular units for instruction at specified grade levels. This information can, to a considerable extent, be inferred by inspecting the percent of correct responses to NAEP items at the respective grade levels and in schools classified by demographic features. With these figures, the curriculum specialist can make judgments such as "calculating the area of a triangle is poorly understood at the eighth grade level and, therefore, should be emphasized more in the geometry unit", or "performance of the square root algorithmn is so poor that it should be relegated to the calculator rather than taught".

For NAEP data to be useful in this role the items must be written according to the content and skill specifications that are

currently accepted in these fields. If the items are suitably described, classified, and made available along with item statistics for the national school population and various subpopulations, the specialists in the subject matter fields have an empirical basis for assigning particular topics and units to appropriate grade levels. Sebring and Boruch (1982) found ten different publications of national curriculum organizations (mathematics, writing, English) that used NAEP data in this way.

Even more directly, the educational test developers can use NAEP item statistics to choose items for attainment tests at specified grade levels. From its inception, NAEP has had a policy of releasing items and item statistics for this purpose. Test specialists can obtain released items along with percent-correct estimates for the national population and major subpopulations. Either the actual items, or new items constructed according to the same specification, can be incorporated in the tests under development. The use of released items in this way by state assessment programs has been one of NAEP's more successful dissemination programs (Sebring and Boruch, 1982).

Matrix sampling. To serve the need for curriculum evaluation in detailed content of major subject-matter areas, NAEP pioneered the use of matrix-sampling techniques for gathering information on student attainment. Because any one student cannot possibly be administered enough items to measure attainment reliably in all the units within a subject-matter area, assessment instruments consist of many distinct forms or booklets, typically about 30, containing items drawn from the item content classes. Any given student takes only one such form, but because different students take different forms, and because the sample of students is very large, there is sufficient information to estimate accurately the difficulty of all the items in the complete set. Moreover, responses to subsets of items representing main content categories or skills (i.e. problem solving in physical science), can be aggregated for use as indices of attainment when comparing groups of students in

different educational programs. The matrix-sampling technique, highly efficient in these applications, is the mainstay of evaluation oriented educational assessment.

That NAEP has been relatively successful in the evaluation area is no doubt a reflection on Ralph Tyler's long involvement with this field. Many of the familiar concepts of curriculum development, especially the use of cross-classifications of content topics and behavioral outcomes as a rubric for specifying curricular objectives, are his contributions. He understood clearly that the aims of measuring student performance for purposes of evaluation are quite different from those of the study of individual difference in attainment. Although the raw data of both are student responses to cognitive tasks, it is the curriculum, program or material that is being examined in evaluation, and not the student. This evaluation orientation proves to be a source of difficulty for other users of assessment results, however, because most of them need case-by-case measurement of individual student attainment. These users are concerned primarily with the relationship of background characteristics of students with levels of attainment; investigation of these relationships by standard statistical methods requires scores for the students, and not just group-level statistics. A principle objective of the assessment design proposed in the present paper is to provide detailed evaluation of schools and programs jointly with accurate case-by-case measurement of student achievement, simultaneously with one assessment instrument.

1.2 POLICY FORMULATION

The term is applied here to the activities of all those persons who are concerned with the formulating, proposing, and influencing of educational policy, both at the state and the national level. The media, public interest groups, teachers organizations, state legislators, chief state school officers, professional educators, state governors, and the leaders of federal programs in support of edu-

cation all belong in this class. The interests and backgrounds of these parties to public education are varied and specialized. Not all are directly involved with the school system. Agencies responsible for promoting industrial development in the states, for example, need measures of educational quality and outcomes as indicators of the quality of the available work force and as an inducement to workers who would be moving into the state.

Members of this constituency are alike, however, in attending only to broad indices of student attainment in main subject matter areas; the detailed information on curricular objectives that is required by evaluators has little relevance to them. They may get some benefit from aggregate assessment indices, such as average percent correct of items in such areas as Reading or Science, especially if the index is broken down by relevant categories within the population of students (e.g., by income level of communities). They might also find time trends in such statistics meaningful. But they are most comfortable with enumeration data that describe the dispositions of persons in the political and administrative units relevant to them.

Census-like reporting. The problem with item average-percent-correct statistics is that they are nonlinear measures of effects and support only gross ordinal or directional judgments. Policy decisions are much better served by setting standards of attainment and reporting the percent of students who surpass these standards at various stages in their educational careers. We find educational standards used in this way in reference to literacy, as in the statement that "13 percent of the U. S. adult population are functionally illiterate". Apart from the question of the behavioral criterion for "functional" literacy, or of the age distribution of such literacy, it must be admitted that this kind of figure presents the condition of reading attainment in terms that are directly relevant to policy decisions: either the proportion of persons who are illiterate is too large to tolerate and some corrective action must be taken, or the proportion is small and not demanding of immediate attention. To

be most useful in policy formulation, measures of educational performance should be expressed in these census-like enumerations of relative numbers of persons who attain a certain standard of performance.

This type of information requires, however, estimation of attainment levels in individual pupils, not merely the aggregated responses of groups of students to items. Regrettably, the matrix-sampled assessment as it was developed and used by NAEP up to 1985 did not supply any type of score for individual students, and so there was no possibility of defining standards of individual attainment and estimating numbers of students who had achieved those standards. Indeed, the climate of opinion at the time militated against any such setting of standards. Eventually, however, the minimum competency and accountability movements led to more favorable attitudes toward describing educational outcomes in these terms.

Thus, when the conduct of NAEP passed from the Education Commission for the states to Educational Testing Service, an attempt was made in the Reading area to include enough items per form to enable scale scores in reading ability to be estimated for individual respondents. This was in fact done, but unfortunately, the matrix-sampling design selected was not well suited to this application and many students were administered too few items (only nine items in many of the forms) for accurate measurement. Moreover, because the items were often poorly positioned with respect to difficulty, too many students responded correctly to all items, thus limiting the distinctions that could be made between the students. As discussed below, special features of instrument design, including provision for two-stage testing, are required for accurate measurement of individual student achievement in the context of matrix-sampled assessment.

The ETS developers of a NAEP reading scale were conscious of the need for census-like reporting and assigned labels to certain points on the scale that implied reading standards (NAEP,

1985). Five levels of reading proficiency were arbitrarily defined, and typical reading passages that could be understood with 80 percent mastery at those levels were exhibited. Although a step in the right direction, this method of locating the threshold points lacked a practical motivation: the behavioral significance of the levels was not elaborated in a way that would have direct implications for policy.

We can only conclude that a more suitable matrix sampling design, one that will allow accurate estimation of individual attainment, combined with more objective methods of defining scale standards that correspond to practical behavioral criteria, is essential to the effective policy use of NAEP data.

1.3 RESEARCH

During the period when NAEP was conducted by the ECS considerable emphasis was placed on the preparation of user tapes to make NAEP data available for secondary analysis. It was hoped that educational researchers in academic settings, graduate students, and private educational and social research organizations would find in the NAEP results empirical answers to their theoretical questions. In the ten years that such tapes were available, however, they were not extensively used for this purpose. Sebring and Boruch (1982), reviewing secondary uses of NAEP data, found only two formally published, substantive studies based on the NAEP user tapes. Numerous tapes were distributed, but they were used primarily by the states for test and curriculum development. There was little application of NAEP data to basic educational, psychological, and sociological investigation.

Case-by-case reporting. Although this paucity of secondary uses might be attributed to deficiencies in the dissemination effort, a more fundamental reason is that the matrix-sampling data produced by NAEP at that time was not in a form that secondary researchers were prepared to handle. Both the statistical method-

ology and the computer routines available to such workers were then, and still are, based almost entirely upon data in which the variables are organized on a case-by-case basis. That is, all respondents in the analysis must have values for all responses and background variables in question. As we have seen, this is not the form of matrix-sampling data: different cases respond to different items and there are no summary scores or individual respondents. Such data are unsuited to conventional statistical procedures such as correlation analysis, regression analysis, factor analysis, or analysis of linear structural relationships. Although certain types of random effects analysis can be carried out on the NAEP data (see Mislevy, 1985), the advanced nature of these techniques, and the lack of commercially available computer routines for their execution, limits their use for the present time.

It might be thought that the percent correct on selected single items or small sets of items could be related to background variables in the same manner that an attitude item from a social survey would be analyzed. However, individual test items are merely random representatives of the content domain from which they are drawn, and cannot be treated in the same absolute and fixed manner as an opinion on some controversial topic or a voting intention. Experience shows that there is a great deal of interaction between the background variables and specific features of cognitive test items. These item-specific effects severely attenuate the valid information that the test supplies about the relationship in question. Only by aggregating responses to numerous items from the same class can these interactions be averaged away and sufficient generalizability obtained to make secondary analysis productive.

During the ECS period, an attempt was made to circumvent these difficulties by using average percent correct for groups of items as reporting statistics. Although this aggregation over items solves the problem of generalizability, it leads to intractable statistical problems. The complexities of the matrix sampling scheme coupled with the cluster sampling of respondents make the calcu-

lation of error estimates and confidence intervals extremely complex. In fact, most of these estimates can be obtained by the empirical "jackknife" method, which, though serviceable, is not readily adapted to conventional statistical techniques available to secondary researchers.

Under ETS, an attempt has been made to provide scores for individual respondents—in reading proficiency, for example—that can be analyzed by conventional statistical methods. Scaling techniques based on item response theory (IRT) have been used in the reading assessment to compute scores for each respondent based on the small number of reading proficiency items in each booklet. Unfortunately, the form of spiraled, balanced incomplete block design used for the reading assessment contained, for many respondents, too few items to obtain case-by-case scores suitable for secondary analysis. As mentioned above, too large a proportion of respondents answered all items correctly, or nearly all, of the small number of items correctly, leaving the score distributions unsatisfactory for parametric statistical methods. Although a case-by-case user tape was constructed by means of a Baysian attribution technique (Rubin, 1978), the many assumptions involved in these attributions are likely to discourage secondary use. A different solution to the problem of extracting individual student scores from matrix data is presented in Section 3.

One of the few areas of research that can directly use item-percent-correct statistics is the cognitive study of the influence of specific format features or task components on item difficulty. Accounting for variation in some transformation of the item percents correct provides a test of the model for the underlying cognitive process. The NAEP user tapes containing individual item statistics are suitable for these studies, although more advanced work may require the actual case-by-case item responses (see Embretson, Schneider & Roth, 1986).

The school as second-level sampling unit. Another limitation of the NAEP design, both under ECS and ETS, is the failure to col-

lect data in such a way that classroom and school can be employed as a unit of analysis. The experience of the International Educational Evaluation Association, especially in the Second Mathematics Study, has been that many educationally relevant findings are forthcoming from analysis of random classroom and school variation, in addition to individual variation within classrooms (Wolfe, 1986). In research primarily concerned with the contribution of the school to attainment, analysis of assessment results should be based on heirarchical models including random components at several levels. The background characteristics of the school and its program, together with aggregate measures of the achievement of its children, then constitute the multivariate data for investigating factors associated with variation in levels of attainment. Statistical techniques are now available for analyzing these higher level relationships simultaneously with the more psychologically oriented investigations of cognitive performance of students within classrooms and schools (Aitkin & Longford, 1986; Goldstein, 1986).

Hierarchical regression analysis is especially important in adjusting for differences in economic and demographic characteristics when comparing the productivity of schools, districts, states or other administrative units. Such adjustments are essential to fair comparisons of the effectiveness of instructional programs that have differing resources. The regression equations for making these adjustments must be those appropriate to the level of the population hierarchy in question. (Between-school effects, for example, do not appear in regression models for variation between students within schools.) If the hierarchical model includes between-student variation, the analysis would, of course, have to be based on case-by-case data with attainment scores for individual students. A design for NAEP that would provide data suitable for such analyses would expand greatly the scope for secondary analysis.

2 ASSESSMENT OF CHANGE: THE NAEP LONGITUDINAL DESIGN

From its inception, NAEP has been viewed primarily as a study of trends in national educational outcomes on the scale of years and decades. The term "progress" in its title refers to this conception and more specifically to an anticipated upward trend. To measure trend according to Tyler's (1968) original conception of single-item reporting, the initial item pool in the matrix sample was to be large enough, and the rate of releasing items slow enough, that performance on unreleased items could be compared for many years into the future. Indeed, it is still possible to make such comparisons with NAEP data. When questions of the generalizability of single-item reporting necessitated the adoption of average percent correct reporting, however, the steadily shrinking size of the set of unreleased items, and the fact that the initial set of items was somewhat too difficult at some age levels became critical problems. New items were therefore introduced into the pool around 1975, with the result that the base for reporting the average percent correct scores changed. When data for the decade were displayed later (see Burton and Jones, 1982), discontinuities appeared in the time trend graphs at the point the item pool was refreshed.

IRT scaling. To avoid these troublesome problems of updating average percent correct measures, Bock, Mislevy and Woodson (1982) proposed the use of methods for scaling assessment results. Because the IRT procedures estimate, for large samples of item responses, the location of each item on an internally defined proficiency contiuum, and because each item contributes independently to locating the respondant on that scale, items can be added or deleted from the scale without biasing the estimated attainment of the respondent. Thus, scale scores estimated before an update of the item pool are commensurate with scale scores computed after the updating. Adapted to group-level scoring (Bock & Mislevy,

1981), these methods have been used with good success since 1980 by the California Assessment Program for measuring the performance of each school in the state on a wide range of curricular objectives.

The special properties of IRT scale score estimation make possible systems of test maintenance that preserve score comparability in the presence of replacement of a fraction (usually 20 percent) of the item pool annually. They also provide provision for inclusion of "variant" items that are calibrated by extension from the active items, but are are not initially used in computing scale scores; only after the properties of these variant items have been judged satisfactory are they merged into the active item pool. These systems can also detect and correct for so-called "item-parameter drift", i.e. changes in the relative difficulty of items due to changes in curricular emphasis during the period that items remain in the pool. Computer based systems of this type are presently under development by the Department of Defense for accessions testing and could readily be adapted for purposes of educational assessment (See Green et al., 1984).

Inasmuch as ETS has already reanalyzed the ECS data using IRT methods, current results expressed in scale scores can be compared with earlier years. Plots showing reading trends for the past 15 years appear, for example, in the 1985 NAEP reading report; they do not have the break at the earlier item updating that is seen in the average percent correct reports (NAEP, 1985). All discussion of assessment design in the remaining sections of the present paper assumes IRT scaling for purposes of longitudinal analysis of the NAEP data.

3 A DESIGN FOR BROADER USE OF NATIONAL ASSESSMENT DATA

Could the assessment instruments used to collect information on student attainment be designed in such a way that all of these

potential uses of the data—evaluation, policy formulation, and research—are satisfied simultaneously? There are very good prospects that they could. Recently, Bock and Mislevy (1986) introduced for the purposes of state assessment programs a new type of assessment instrument, called the "duplex design" which, when used in conjunction with two-stage testing, provides accurate estimates of students' proficiencies in main content areas, while at the same time measuring the progress of classrooms, schools, or larger units in attaining detailed curricular objectives.

The duplex instrument employs multiple test forms with distinct items, as in matrix sampling, but the item arrangement in the forms is such that responses can be aggregated within forms to measure specific student proficiencies, and can be aggregated across forms to measure curricular objectives at the classroom or school level. By use of IRT methods, the attainment of individual students responding to different forms can be estimated on the same scale, in comparable units, and with uniform accuracy. One such form, requiring 45 minutes of administration time, is capable of measuring, with good accuracy, three distinct proficiencies, provided two-stage testing is employed. In two-stage testing, the student is assigned a second-stage test tailored to his level of attainment, as provisionally estimated either from the results of a brief first-stage test, or from teacher's knowledge of the student's previous performance. The final estimate of the student's proficiency level is unaffected by which particular second-stage test he takes, but the accuracy is optimal when he responds to a test suited to his level of performance.

Similarly, IRT methods make possible the estimation of scores for classrooms or schools by aggregating responses of different students to different items representing the same curricular objective. There can be as many objectives measured as there are items per form, typically 30 to 50; thus, an instrument consisting of 20 to 30 such forms yields good generalizability of results at the classroom or school level, assuming a classroom size of 25 or more. This

number of objectives is sufficient for measuring objectives in the level of detail typically required in evaulating alternative curricula or instructional methods and materials.

3.1 An example of a duplex design

A field trial of a duplex design in eighth grade mathematics is presently being carried out in Illinois and California under the auspices of the NIE Center for Student Testing, Evaluation and Standards. The layout of a form from this design is shown in Table 1. The proficiencies and content categories for this design were arrived at by consensus of the mathematics curriculum specialists of the California and Illinois assessments with the assistance of mathematics educators from the University of Chicago and Illinois State University. The design was replicated in 8 random forms, each of which consists of 3 second-stage booklets, one a higher level of difficulty, one intermediate, and one at a lower level of difficulty. To link the proficiency measures in each form by IRT scaling, there was an overlap of four common items between the first and second booklet and between the second and third booklet in each of the three proficiency areas.

Items for constructing the forms were drawn from pools provided by the California and Illinois Assessment Programs, and the Second Mathematics Study of the International Educational Achievement Association. In the absence of sufficient items representing *irrationals, inequalities, other systems of measurement*, and *experiments and surveys*, these topics were omitted from the final instrument. One item from the item pool representing each of the remaining 45 cells of the design was then drawn to constitute a form (test booklet). The item pools were stratified by item difficulty to produce forms at suitable levels of difficulty for the second-stage testing. The first-stage test, which was common for all pupils, consisted of 12 items especially selected for uniform spacing of difficulty and high validity.

Table 1
A GRADE 8 MATHEMATICS DUPLEX DESIGN

		Proficiencies		
	Content Categories	a. Procedural Skills[a]	b. Factual Knowledge[b]	c. Higher Level Thinking[c]
10.	*Numbers*			
	Integers	11a	11b	11c
	Fractions	12a	12b	12c
	Percent	13a	13b	13c
	Decimals	14a	14b	14c
	Irrationals	15a	15b	15c
20.	*Algebra*			
	Expressions	21a	21b	21c
	Equations	22a	22b	22c
	Inequalities	23a	23b	23c
	Functions	24a	24b	24c
30.	*Geometry*			
	Figures	31a	31b	31c
	Relations & Transformations	32a	32b	32c
	Coordinates	33a	33b	33c
40.	*Measurement*			
	English & metric units	41a	41b	41c
	Length, area & volume	42a	42b	42c
	Angular measure	43a	43b	43c
	Other systems (time, etc.)	44a	44b	44c
50.	*Probability & Statistics*			
	Probability	51a	51b	51c
	Experiments & surveys	52a	52b	52c
	Descriptive Statistics	53a	53b	53c

[a]Calculating, rewriting, constructing, estimating, executing algorithms.

[b]Terms, definitions, concepts.

[c]Proof, reasoning, problem solving, real-world applications

This particular duplex design provides individual student scores for three broad types of mathematics proficiency—namely, procedural skills, factual knowledge, and higher order thinking processes. These three scores can be combined into a general measure of the student's mathematics proficiency if desired. At the same time, performance of the school can be measured in each of the 57 (or in this case 45) cells of the content by proficiency classification. This detailed information can, of course, be aggregated over schools to obtain regional or national indices with respect to the same objectives. The objectives are sufficiently specific to allow strengths and weaknesses of curricular emphasis, instructional methods, or educational materials to be appraised at all levels of aggregation from the classroom upward, including special groupings of schools by program or background characteristics. Although combined with accurate measurement of individual achievement, the evaluation function of matrix sampling is retained intact in the duplex design.

3.2 Comparison of the duplex design to the present NAEP item design

While not all cells of the duplex design are necessarily represented in the NAEP item pool, the item content of a duplex design such as shown in Table 1 is not inconsistent with the present NAEP mathematics items. The design is distinctive from the present NAEP assessment instrument only in the highly structured arrangement of the material within the forms, the greater number of items per test booklet, and the provision of three levels of difficulty among the test booklets for purposes of two-stage testing. The duplex design is motivated much more by substantive educational and cognitive considerations than the BIB (balanced incomplete block) spiraled design used in the 1985 NAEP reading assessment (Messick, Beaton & Lord, 1983). Whereas the BIB design has elaborate provisions for estimating item-response in-

tercorrelations, the duplex design will provide for such analyses only within test booklets. Correlations between the measures of curricular objectives will be possible at the school level of the hierarchical design but not at the student level. As mentioned above, where the BIB design for reading failed most notably, was in the ability to estimate scores for individual students with acceptably high and uniform levels of reliability. Although the attribution procedures applied to the data was a sophisticated data-salvage effort, the original BIB design was obviously not well suited to the needs of the secondary data analyst.

3.3 Rotation sampling of schools

Admittedly, the duplex design requires testing times more comparable to those of traditional achievement testing batteries than those of the present NAEP matrix sampling instrument. This has implications for allocation of available classroom time and for the sampling of schools. One could imagine the program set up in a basic two-year cycle in which two class periods per school are required for measuring, say, Mathematics and Reading in even numbered years and Science and Writing in odd numbered years. Cooperation of the schools to take on this somewhat greater burden of testing could be encouraged in two ways. First, if all students were tested in classroom settings at the target grade level, then useful reports could be returned to the school both for individual students in the proficiency measures, and for the classrooms and or schools for the curricular objectives. Because national data would be available, the students and the school personnel whould find results expressed in comparison with the nation interesting and informative. This would satisfy an important condition of good testing practice: that the participants be motivated by feedback of meaningful results for their efforts in taking the tests. This condition is not satisfied in the present NAEP design.

Another way to enhance cooperation would be to sample schools

in four-year rotation panels, with half the national sample rotating every two years. This has many advantages in terms of the economies of recruiting schools for the study, establishing a committed relationship over a period of time, and having the possibility of examining change within the school over the rotation period. Assuming the above testing cycle, each school would then complete the cycle twice during the four year rotation. Although there is some danger that "Hawthorne" effects would inflate performance during later years, the presence of these effects could be detected by comparing the average performance level of schools in the first cycle of testing with those in the second; expected within school effects due to the testing program itself, could then be estimated and corrected in the data. A very important advantage of these rotation designs, and the main reasons for their use in other social survey research, is that where change over time is of primary interest, the precision of estimating change from longitudinal data within units is much better than estimates from independent cross-sectional samples. On the conservative assumption that results between years will correlate 0.70 or higher at the school level, the estimate of average gain based on 100 schools in a rotation design has the same precision as at least 154 schools in a cross-sectional design. When focusing on the measurement of progress and change, NAEP could enhance its sampling efficiency appreciably by the use of rotating four-year panels of schools rather than the present independent samples.

4 THE CONTRIBUTION OF NAEP TO STATE TESTING PROGRAMS

The mission of NAEP should be viewed in the context of the extensive state-wide student testing that is in place or authorized at the present time. Forty-seven of the 50 states have some form of mandated testing, typically minimum competency testing (Winfield, 1986). Thirty-nine states require the administration of tests

of basic skills at several benchmark grade levels, and in twenty three of the states, satisfactory performance on the test at the twelfth grade level is a requirement for graduation. The standard for passing performance is, in most cases, set arbitrarily by the state legislature, state department of education, or school district authorities. In many states the minimum competency test is produced by the state department of education; in others, the school districts have the option of preparing their own tests or choosing from among various commercial tests.

All of these tests are traditional individual achievement measures, but many are pitched at lower levels of attainment and are not useful for measuring the full range of student attainment. For this reason, and also because of the variety of tests used by different states, there is little possibility of using the minimum competency testing results for comparisons between states. The commercial achievement tests are better in these respects, but because they tend to be used when districts have local testing options, representative state results may not be available.

Building on the NAEP model, a number of states have implemented sampling assessment programs oriented toward curriculum evaluation and policy formulation. Because, only a sample of schools and students are tested, however, these types of assessment do not serve the needs for school management or any form of student guidance or competency certification. For this reason, most such programs are being converted to an every-student testing (in Illinois and Michigan, for example). The California Assessment Program is a prototype for those programs that test all students in the state at the benchmark grade levels and provide each school in the state with a computerized report on strengths and weaknesses in attaining specified curricular objectives. The need for testing in the states to serve the purposes of guidance, certification, school and program evaluation, local and state level management of schools, and broader state educational policy formulation necessitates continued every-student testing. Because NAEP is a

sampling assessment, it cannot serve the purposes of these types of programs and cannot replace them.

Recently, however, the unique value of data on which between-state comparisons of average student attainment can be based has been recognized. This type of information is often sought by companies searching for new industrial sites, and states that cannot supply such figures may have a disadvantage in such searches. In these cases, the student attainment information may be needed both for the state as a whole and for particular regions or school districts that are candidates for economic development. What would best serves this need is a way of expressing their state test results on a common scale so that the state average and regional and district averages within the state can be compared from one state to another. NAEP could play an important role in making such comparisons possible.

Modern test theory provides a variety of methods of equating separate tests to a common scale based on some agreed-to standard test. Although one of the more popular commercial achievement tests might play this role, various considerations of proprietary rights to normative data, possible allegations of unfair competition, and the non-public nature of the construction of such tests has discouraged this approach.

Another possibility is that a voluntary association or consortium of the states could pool materials from the separate testing programs and produce an agreed-upon standard test. The Council of Chief State School Officers, which has recently endorsed the principle of between state comparisons of educational attainment, is now investigating this possibility. The main impediments are the great variety of state programs, the difficulty of obtaining agreement with such a large number of governmental units, and the costs of creating a new organizational base to support such an effort.

Perhaps the most attractive approach is to modify the NAEP assessment instruments to serve this purpose. That is, the choice

of content and proficiencies to be measured in the various subject matter areas by the NAEP instrument could be moved in the direction of a model state instrument. This would require the cooperation, but not necessarily the separate endorsement, of the various state testing programs. The effort here would be similar to the test development carried out by the International Educational Evaluation Asscciation for the purposes of international comparisons of student attainment. (That the international effort was broadly successful encourages the belief that similar cooperation could be obtained among the fifty states.) The task is not as difficult as it may first appear: most of the decisions about subject-matter contents and performance skills are necessarily based on curriculum theory, textbooks, and instructional practices, all of which are determined much more at the national level than at the state level. Although the terminology in which objectives are expressed from state to state may differ, the intent is much the same. The actual item content of the tests is even more similar: in many cases the items are drawn from from the same nationally available pools, including the NAEP retired items.

Assuming that the NAEP assessment instrument could be reformed in a manner that would make it acceptable as a standard test in the main subject matter areas, the technical task of equating state attainment scales to NAEP scales is relatively straight forward. The key to the equating procedure is the national coverage of the NAEP sample. In those states with every-student testing programs in the same benchmark grades in which NAEP tests (i.e., fourth, eighth, and eleventh grade), there are students who have taken both the NAEP tests and the state tests. (In many cases this testing occurs at approximately the same time of the year, usually between late February and early May). Then, because the schools keep the coded identification of the respondents to the NAEP tests, it is possible for each state to match, student by student, the state scale scores for their tests with the corresponding NAEP scale scores. From these paired records, the

relationship predicting the NAEP score from the state score can be established by simple statistical regression methods.

Alternatively, student records can be matched virtually uniquely within schools and grades by such information as first letter of last name, sex and date of birth. This method has been used by the California Assessment Program to obtain school-level prediction equations for equating CAP *Reading Comprehension Skill* to the College Board *Degrees of Reading Power* test. The correlation of the equated school scores is extermely high.

Although technical provisions such as corrections for the time of testing and interpolation of results for state grade levels not included in the NAEP benchmark grades may be necessary, these adjustments could be made with data obtainable from the state testing programs. Provision for counter balancing test administration to control order effects , and the possible use of more than one state scale to best predict each of the NAEP scales, may also be desirable. These technical matters should be quite manageable within the scope of modern statistical and cognitive test theory.

Under ETS management, NAEP has provided the states the option of purchasing an increase in the size of the NAEP sample within the state in order to make comparisons with the national results possible. From a financial point of view, this is a relatively inexpensive way of obtaining information on the national standing of average state attainment. But it also has certain disadvantages. One of these is the redundancy of additional testing if the state is already carrying out every-student testing in the same subject matter areas. This means that the cost and, perhaps equally important, the classroom time consumed by the additional testing are an unnecessary duplication of expenditures. Another disadvantage is that the sampling assessment gives only a statewide average results, whereas district and regional results may be more germane to educational planning or economic development. The alternative of linking the states' every-student results to the NAEP scales would provide for national and between state comparisons

at the district and local school system level as well as at the state level. That this multilevel data would be much more useful than state-level results alone argues for the role of NAEP as provider of the standard test by which state results could be equated.

5 SUMMARY OF DESIGN FEATURES FOR AN IMPROVED NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

The redesign proposed here for NAEP can be summarized in eight points:

1. Application of the duplex design to the assessment instrument in order to provide measurement of individual student achievement in broad subject-matter proficiencies, simultaneous with evaluation of curricular objectives at the level of classrooms, schools or higher levels of aggregating the data.

2. Use of two-stage testing for reliable measurement of individual student achievement in three main proficiencies per subject-matter area.

3. IRT scaling of achievement and evaluation measures in order to provide comparability of scores for long-term trend analysis and to allow periodic updating of the item pools.

4. Annual testing and a two-year cycle of the main subject-matter areas as follows:

Year 1	Year 2
Reading	Science
Mathematics	Writing

   This cycle could be supplemented by occasional testing in more specialized topics.

5. Four-year rotation sampling of schools, with half of the schools replaced every two years.

6. Retention of the classroom and school identification of students for purposes of multi-level heirarchical data analysis.

7. User-oriented reporting of assessment results:

   a) Census-like reporting results in terms of criterion referenced standards for purposes of policy formulation and for communicating with the media and public.

   b) Conventional case-by-case reporting of student proficiency scores for purposes of secondary analysis.

   c) Reporting of group statistics for evaluation of curricular objectives.

   d) Reporting of item statistics for purposes of test development and cognitive research.

8. Cooperation with state testing agencies in providing standard scales in terms of which state test scores can be expressed for purposes of between-state comparisons.

6 IN WHAT DIRECTION SHOULD NAEP DEVELOP TO SERVE MOST BROADLY THE NATIONAL EDUCATIONAL EFFORT

At the time of NAEP's inception in the 1960's, conditions were not favorable for a program with strong ties to state departments of education. Neither were NAEP's activities closely related to those of other centers of educational research and testing. As a result, the methods of testing, the objectives assessed, and the manner of reporting results was rather idiosyncratic to the thinking of particular members of the NAEP staff and advisory committees at that time. The assessment program and its reports were not in the mainstream of the national educational effort as represented by the state school systems or by educational research and scholarship in the universities. The educational literature of

that period contains very little information about NAEP and no careful analysis of its purpose and design.

So it not surprizing that, apart from utilization of the pool of released items, the states took little notice of the existence of NAEP as they began to move in the direction of greater accountability and more rigorous evaluation of their instructional programs. Similarly, the educational research community responed more to results from the established national testing organizations, and to the IEA studies, which originated among curriculum specialists within the universities, than to NAEP.

At the present time, the climate of opinion is much more favorable toward a national educational assessment in lively relationship with state testing programs and in active collaboration with national educational research centers. The source of this change in attitude has been, in part, the realization that other technically developing countries with more comprehensive standards of educational excellence have, according to the results of the IEA studies, attained educational outcomes considerably above those of the United States. To meet the educational demands of an industrial and technological society, NAEP must contribute to clarifying and raising educational standards in the United States. To do so it must become more intimately a part of the national educational effort which, in the U.S. system, necessarily means greater cooperation with the states and the state testing programs. At the same time it needs to move closer to the educational research community represented in the universities and other research organizations. It can do this by providing data in a form suitable for secondary analysis, by funding extramural reseach, and by encouraging national publications and conferences devoted to informed discussion of educational problems. These are some of the steps need to expand the community of users of NAEP results—steps that should lead ultimately to better understanding of the curricular, instructional, and organizational factors that make for effective public and private education.

REFERENCES

Aitkin, M & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A.*, **149**, 1-26

Bock, R. D. & Mislevy, R. J. (1981). An item response model for matrix-sampling data: the California Grade-three assessment. In D. Carlson, (Ed.), *Testing in the States: Beyond Accountability. New Directions in Testing and Measurement*, No. 10. San Francisco: Josey-Bass.

Bock R. D. & Mislevy, R. J. (1986). Comprehensive educational assessment for the states: the duplex design. (submitted for publication.)

Bock, R. D., Mislevy, R. J., & Woodson, C. E. M. (1982). The next stage in educational assessment. *Educational Researcher*, **11**, 4-11, 16.

Bock, R. D. & Muraki, E. (1986). Detecting and modeling item-parameter drift. (in preparation).

Burton, N. W. & Jones, L. V. (1982). Recent trends in acheivement levels of black and white youth. *Educational Researcher*, **11**, 10-14.

Embretson, S., Schneider, L. M. & Roth, D. L. (1986). Multiple processing strategies and construct validity of verbal reasoning tests. *Journal of Educational Measurement*, **23**, 13-22.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43-56.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L. & Reckase, M D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, **21**, 347-360.

Haertel, G. D., Walberg, H. J., Jonker, L. & Pascorella, E. T. (1981). Early adolescent sex differences in science learning: Evidence from

the National Assessment of Educational Progress. *American Educational Research Journal*, 18, 329-341.

Harnischfeger, A., Huckins, L. E. & Wiley, D. E. (1977). *The National Assessment of Educational Progress Model: A tool for achievement based Title I fund allocation?* Chicago: ML-Group for Policy Studies in Education, CEMREL, Inc.

Messick, S., Beaton, A. & Lord, F. (1983). *A New Design for a New Era.* Report No. 83-1, National Assessment of Educational Progress. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1985). *Inferences about latent populations from complex samples.* NAEP Research Report, 85-41. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Reiser, M. R., & Zimowski, M. F. (1981). *Scale-score reporting of National Assessment data.* Final report of ECS Contract 02-81-20314. Chicago: International Educational Services.

NAEP (1985). *The Reading Report Card.* Princeton: Educational Testing Service.

Rubin, D. B. (1978). Multiple imputation in sample surveys. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

Sebring, P. A. & Boruch, R. F. (1982). *On the uses of the National Assessment of Educational Progress.* Report No., A-137-5 NIE Grant G-79-0128. Evanston: School of Education, Northwestern University.

Tyler, R. W. (1968). *What is an Ideal Assessment Program?* Sacramento: Bureau of Research Services, California State Department of Education.

Wolfe, R. G. (1986). *The IEA Second International Math Study: Overview and selected findings.* Annual meeting of the American Educational Research Association, San Francisco, April 1986.