DOCUMENT RESUME

ED 231 878 TM 830 482

TITLE	NAEP Design.
INSTITUTION	National Opinion Research Center, Chicago, Ill.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	3 Nov 82
CONTRACT	400-82-0019
NOTE	42p.; Recommendations for redesign of NAEP.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Educational Assessment; Elementary Secondary Education; *Evaluation Methods; *Federal Programs; Latent Trait Theory; Maximum Likelihood Statistics; Measurement Techniques; Methods Research; *Program Descriptions; *Research Design; Research Methodology
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

To fulfill its mandate and its potential the National Assessment of Educational Progress (NAEP) must be a multi-content annual assessment of priority learning areas. This report describes a procedure under which measures of attainment can be charted over time in a broad range of skill domains, on item-invariant item response theory scales, using data gathered in highly efficient multiple-matrix sampling designs. A new sampling design is proposed that would allow annual reporting in the priority learning areas and would greatly increase the precision of the estimates made from NAEP data. The design calls for retention of many of the central features of the current design, thus maintaining comparability, but proposes two major innovations: substantial increases in the number of pupils tested, and use of a method of sampling known as "rotation sampling." Specific proposals for strengthening NAEP's linkages with state and local assessments to enhance their ability to improve education are presented in three major categories: structures and staff, assessment activities, and supplementary activities. Primary type of information provided by report: Program Description (Operating Policies); Procedures (Conceptual). (PN)

NAEP DESIGN

A Report
Submitted to:
NATIONAL INSTITUTE OF EDUCATION

Contract Number
400-82-0019

Submitted On:
November 3, 1982

Submitted By:
The National Opinion Research Center
6030 South Ellis Avenue
Chicago, Illinois 60637

TABLE OF CONTENTS

INTRODUCTION

In our preliminary proposal we presented our view of the contributions and shortcomings of the National Assessment of Educational Progress to date:

NAEP's Contributions

- NAEP has provided national data on change in educational achievement which are both valid and accurate in terms of comparisons over time, and which cover a wide variety of achievement content areas.

- NAEP has modeled new concepts and procedures for state assessments and other testers. (This has been primarily accomplished through technical assistance efforts and the exhibition of the NAEP methodology in publicly available reports.) As a consequence of this modeling NAEP has achieved greater curricular relevance and balance in state assessments and in some local assessments; lower pupil response burden through matrix sampling; lower costs per information unit through innovations in sampling design.

- NAEP has provided achievement data for secondary analysis by the educational research community (though use of this data has been somewhat limited).

NAEP's Shortcomings

- NAEP has failed to compete for public attention with less representative alternative indicators (e.g., the yearly SAT scores) because information available every four years cannot possibly compete effectively with information available every year.

- NAEP reports of achievement results have had limited relevance and comprehensibility, thus limiting meaningful evaluation of the educational system and not providing a basis for modifying that system at the local, state, or federal level.

- NAEP lacks adequate linkage to politically responsive units with responsibility for action to maintain or improve the quality of education.

- NAEP has depended too much on professional groups with expertise and judgment limited to particular learning areas for determining the content priorities of the Assessment. NAEP needs broader political, economic, and general policy input for establishing content-area priorities and information needs for the 1980s.

- NAEP lacks provisions for specific policy-focused data collection and analysis to supplement the basic achievement results.

From this analysis we developed two major recommendations for the redesign of NAEP:

Recommendation 1: Modify NAEP to replace the SAT as the major publicly accepted index of national educational quality.

Recommendation 2: Strengthen NAEP's linkages with state and local assessments to enhance their ability to improve education.

During the preliminary grant period we focused on the NAEP shortcomings that were most directly related to the recommendations we had made. In the process we confirmed our central belief about the design of NAEP: to fulfill its mandate and its potential NAEP must be a multi-content annual assessment of priority learning areas. We also refined and elaborated our two major recommendations. The material presented in chapters 1 and 2 supports recommendation 1, that in chapter 3 supports recommendation 2 (though there is, of course, some overlap). The following is a summary of the three chapters that follow.

Through the use of multiple-matrix sampling, an innovation developed largely in the context of NAEP, the national assessment has been able to provide economical feedback on the levels of attainment throughout the nation over a broad range of skills. Unfortunately, the analysis of trends in NAEP data has been hampered by the inadequacy of classical psychometric theory for reporting these results; the familiar percent-correct scores, averaged over all items in a given skill area cannot be compared from one assessment to the next as the item pool evolves, yet comparisons based on only common items must discard over half the information accumulated.

This problem is solved in principle by item response theory, which allows for the estimation of attainment levels on invariant scales despite changes in the item pool. Until recently, however, IRT methodology as it had developed in the context of individual measurement could not be applied to the sparse (at the level of individuals) multiple-matrix samples of data that characterize modern assessment designs. The recent introduction of marginal maximum likelihood techniques to item response theory by Bock and Aitkin (1981) has set the stage for the combined exploitation of the efficiency of multiple-matrix sampling theory and the flexibility of item response theory. Section 1 of this report describes a procedure under which measures of attainment can be charted over time in a broad range of skill domains, on item-invariant IRT scales, using data gathered in highly efficient multiple-matrix sampling designs.

In section 2 we present a proposal for a new sampling design for NAEP. This design would allow annual reporting in the priority learning areas and would greatly increase the precision of the estimates made from NAEP data. The design calls for retention of many of the central features of the current design, thus maintaining comparability, but proposes two major innovations: substantial increases in the number of pupils tested, and use of a method of sampling known as "rotation sampling."

We firmly believe that the full potential for national-state

interaction in NAEP, and the resulting benefits for both parties, is far from fully realized. To support our second major recommendation--that these linkages be strengthened--we present quite specific proposals for strengthening the NAEP/state-assessment linkages in section 3. These are divided into three major categories: structures and staff, assessment activities, and supplementary activities.

This report was prepared under the direction of R. Darrell Bock, Celia E. Homans, and David E. Wiley, representing respectively the three institutions in the NORC consortium--the University of Chicago, the National Opinion Research Center, and Northwestern University.

1. CONCEPTS IN THE ANALYSIS OF ASSESSMENT DATA

Until relatively recently, the theory and practice of educational measurement have been devoted entirely to determining the attainment levels of individual pupils. The purpose of such measurement, based largely on objective achievement tests, has been to grade pupils for purposes of advancement and qualification and for prediction of their future scholastic success.

That there could be a form of educational measurement directed not at individual pupils but at the groups to which the pupils belong became recognized only in the 1960s', when Tyler began to advance the idea of assessing educational progress at the national level.

The possibility of applying results in matrix-sampling theory (Hooke, 1956) for this purpose was investigated by Frederic Lord (Lord and Novick, 1968). He came to the somewhat surprising conclusion that levels of attainment in the aggregate could be estimated efficiently by sampling pupils from the population and presenting each with an independent sample of items from the content domain of interest. In fact, the most efficient design on a per-response basis proved to be one in which each randomly selected pupil takes one randomly selected item. This result opened the way to multiple-matrix sampling designs, now widely employed in educational assessment, which make possible the evaluation of attainment in many content domains simultaneously without an excessive burden on any one pupil.

A typical assessment design with items for evaluating, say, 30 different subdivisions of content, might consist of perhaps 25 forms, each made up of 30 items, one from each content domain. When the results of administering the different forms to different pupils are aggregated to the population level, a comprehensive profile of attainment in the 30 domains, each represented by 25 items, emerges. The status or progress of education for the system as a whole can thus be evaluated in greater detail and with much less cost than is possible in traditional "every-pupil" achievement testing programs. Our recommendation of a multiple-content annual assessment is based on the principle of multiple-matrix sampling.

Data obtained with matrix-sampling designs are, of course, much different from measurement data in which each subject is presented with all items of each test. Scores for individual subjects cannot be obtained in the subdivisions of a content area because any given subject responds to only a few items from the subdivision. Although each pupil responds to sufficient items in the content area as whole to provide a score, the forms have entirely different items and are not strictly parallel; comparisons between pupils receiving different forms are therefore not possible even for the subject-matter area as a whole. In fact, any attempt to score pupils in assessment should be avoided. Only at the level of populations, subpopulations, or groups in which all forms have been administered and all items represented are comparisons of attainment levels possible.

1.1 The Structure of the Matrix Sample

Although the matrix sample is often represented simply as an array in which rows correspond to pupils and columns to items, both rows and columns can, and usually do, have a formal structure. The arrangement of rows will typically reflect the administrative units into which the population of pupils is divided. The arrangement of columns will be structured according to prevailing theories of school curricula, the categories of which are to a large extent historical, expedient, and conventional. For the reporting of assessment results, the row and column structures are of paramount importance; they determine the levels at which the data are aggregated and summarized.

1.1.1 Levels of Aggregation in the Population of Pupils

To the question "at what levels in the sample of pupils shall the assessment data be summarized and reported?", the answer depends on both the purpose of the program and the density of sampling. In state assessments conducted for purposes of accountability and resource allocation, results can be reported for administrative units to the lowest level that the sampling will support. In the California Assessment, for example, all pupils of the state who are in school on the assessment day are tested. Results can therefore be reported at the classroom and school levels, although only the latter is included in the California program. The same result can be accomplished by testing every pupil in small schools and randomly sampling pupils in larger schools; but with machine-scored objective tests the marginal cost of adding pupils in large schools is so small that sampling procedures are hardly warranted. The school results can, of course, be readily aggregated to the district, county, and state levels by computing average scores for these administrative units weighted by the numbers of pupils in the schools.

In state programs that sample pupils in the system by means of sampling schools, as in the state of Connecticut, the sample density permits only reporting at the county level and for the state as a whole. From the point of view of a state department of education, this type of sampling gives sufficiently detailed information, but it does not provide the kind of school-by-school diagnostic information that allows district superintendents, school principals, and school teachers to evaluate their accomplishments in relation to other schools in the state, especially those in similar communities. It is to provide this kind of detail that the California Assessment Program tests all pupils annually in grades 3, 6, and 12.

National assessments, both in the United States and abroad, have had to rely on relatively low sampling rates and accurate results can be reported only for rather large geographic units. The smallest unit that can be accurately reported in the sampling plan for the United States is the region (i.e. Northeast, Southwest, Midwest, and West). The design we recommend (see section 2) assumes primarily the national-level reports, but, as in the case of the present NAEP, national reports can be broken out by various demographic classes based on characteristics of pupils or of schools. The main result reported in the National Assessment then, in addition to the estimates of total population attainment and change of attainment at each grade level, is the relationship of attainment and change to a variety of background variables collected in the pupil, teacher, and principal surveys that accompany the

assessment testing. Statistical methods for extracting information about these relationships from the matrix-sample design in section 2 are discussed in the technical appendices to this report.

Even with the sparse sample of schools in a national assessment, however, the numbers of pupils tested within schools can be large enough to justify the reporting of results to the participating schools. The relatively small marginal cost of testing additional pupils within selected schools makes this possible. The sampling plan described in section 2, which requires data from moderately large random numbers of pupils within schools in order to exploit the greater precision of the rotation design, will also permit reports to the participating schools. These reports, forms of which are included in an appendix, should provide incentive for the schools to participate in the assessment and help pupils and supervisors to justify the classroom time required for the assessment testing. Attainment levels in the content areas included in the assessment can be reported to each school relative to the levels for other schools throughout the country and for schools with characteristics similar to the school receiving the report.

1.1.2 Levels of Aggregation in the Item Content

In Tyler's early proposals for the assessment, no summarization of the item content was envisaged (Tyler, 1968). Certain selected items were to be released after each assessment and the percent correct and change in percent correct of these items reported. This approach, which was patterned on public opinion surveys where the percent of persons in the sample endorsing one or the other side of an issue is reported, had the pupose of stating the results as concretely as possible and avoiding the abstraction represented by the standardized scores in conventional achievement testing programs. This "fixed-item" form of reporting failed to recognize, however, that items that assess attainment actually are random and represent samples from content areas. They do not have the unique status of the public issues that inform the items in opinion surveys. To present the assessment results in fixed-item terms places on the reader the burden of inferring to what wider content the result can be generalized or to what future performance they predict for the pupils. In most cases the reader is less well prepared to make this inference than the assessment analyst. If, in contrast, many items are reported in order to convey something of the generality of the results, the detail is too great for the typical reader to absorb. Only specialists in test construction want this amount of information, and they can obtain access to the item statistics for released and unreleased items that would better serve their purposes.

The reporting levels actually employed in most NAEP publications aimed at a general audience are at the furthest extreme from individual items. They consist, in many cases, of the average-percent-correct for all items in an annual assessment covering one subject matter area, as, for example, average percent correct on all reading or mathematics items. For some purposes, especially where different classes of content interact with classes of pupils, average-percents-correct for major subdivisions of the content are also reported. Biological science and physical science items, for example, show contrasting profiles of attainment in the two sexes that are only revealed when separate domain scores are reported.

Exhibiting and reporting on specific items remains necessary, of course, to clarify and add interest to the assessment reports. Most of the results for individual items appear in NAEP publications purely for these illustrative purposes. To provide items for this purpose, provision for the continuing release of a certain number of items is a necessary part of the assessment design.

1.1.3 Defining the Content Domains

The reporting of results by classes of items, such as biological and physical science, makes the inference to a larger domain that the reader cannot make from individual items. Two main problems must be solved, however, before items can be assigned to such content domains. The first problem is that of choosing the level of detail on which items should be classified. No one would argue with the division into the main subject-matter areas--reading, mathematics, written expression, science, social studies, art, and so forth. The first three of these areas are mandated for separate measurement in the enabling legislation of the National Assessment; the others at various times have been the objects of all or part of an annual assessment. Within these areas, however, knowledgeable persons, such as teachers, curriculum specialists, and textbook writers, can make additional distinctions that appear in various ways in the organization of teaching programs and materials. Assessments designed to be of maximum help to school curriculum planners and classroom teachers need as much of this detail as possible. Reports showing the strength or weakness of a school or school system in specific topics or skills point directly to changes needed in the instructional program. In the California Assessment, for example, main content areas such as reading, mathematics, and written expression are divided into as many as 35 different content elements for reporting purposes. An effort is made to define elements that correspond to distinct curricular elements that would tend to be emphasized or deemphasized as a whole when instructional resources are reallocated.

The degree of detail that may be appropriate for assessments tied directly to state curriculum guidelines does not seem desirable in a national assessment in a system where educational policy making is decentralized. At the national level, a smaller number of divisions can be used, corresponding, on the one hand, to generally accepted distinctions within subject-matter areas along lines of learning processes and skill development that are recognized by curriculum specialists and educational psychologists, and, on the other, to established categories of content that exist in most subject-matter areas. On the general principle that both content and skills become more differentiated as children grow older, the assessment design we recommend would provide at the 9-year level for 16 content classes to be divided among 4 subject-matter areas, at the 13-year level for 30 domains to be divided among 5 areas, and at the 17-year level for 36 domains to be divided among 6 areas.

The definitions of these content domains, and the specifications for constructing items within them, are the responsibility of the assessment content committees in the several subject-matter areas. The intention is that the results of the assessment will be summarized first at the level of content domains within areas; then, for higher levels, the domain summaries will be averaged to provide indices of attainment for the subject-matter areas or subareas as required for communication to a general audience. For example, an

index of overall reading attainment, or perhaps indices that distinquish literal reading from inferential reading skill, might be presented in reports to the media, whereas more detailed results, such as attainment in reading graphical material might be reported for the benefit of persons concerned with reading education. Because the choice of the initial reporting categories represented by the content domains will greatly influence the usefulness of the assessment results, considerable effort to obtain a broad consensus on their definitions is an essential part of assessment policy. These definitions can, of course, evolve along with the items that make up the domains, but changes must be deliberately paced so that any given content domain will have time to prove its worth to the assessment.

1.1.4 Checking the Assigment of Items to Domains

Whether every item constructed to the specifications of a content domain in fact belongs in that class is an empirical question that cannot always be decided by expert judgment. There is a need to verify psychometrically the homogeneity of items within each domain. We recommend that special pretest forms, in which items within domains appear in balanced incomplete block designs that make possible item-factor analyses, be administered in a sample of those schools that are being retired from the rotation sample. These pretest data also provide information about the difficulty levels of the items that can be used to make the forms in the assessment instruments as comparable as possible.

1.2 Choice of Indices for Reporting Attainment Levels in the Content Domains

The assessment design we recommend does not preclude the reporting procedures now used by NAEP. For the immediate future, at least, the assessment design, even after it is shifted to multiple-content annual assessment, could continue to use the unreleased items from earlier assessments. These items might be classified differently in the new structure, but the form of the items and the conditions of administration should be sufficiently comparable to allow trends in average percent correct on these items to be followed for a period of years before and after the changeover. Any discontinuity in the trends due to noncomparability of item administration will be quickly discernible as the yearly data begin to accumulate. Alternatively, special administrations under old and new conditions would be carried out to equate the scores from the two procedures.

For reasons that have been discussed by Bock, Mislevy, and Woodson (1982), however, total reliance on average-percent-correct reporting is not advisable. Despite their appealing simplicity, average item-percent-correct statistics have severe limitations as a reporting medium. A major defect is the lack of any intrinsic meaning of the average percentage of correct reponses to an arbitrary set of items. The value of this statistic depends entirely on the difficulties of the items, and these difficulties can change unpredicably if new items are substitued for old, or if minor changes are made in item wording. The only useful information this approach can provide is the change in average-percent-correct when exactly the same items are given in two or more assessments.

But changes in item-percent-corrects have their own problems. A severe limitation is their lack of comparability when the absolute

percentages from which they ares calculated fall at substantially different levels. Thus, for example, if boys increase in science knowlege from 52 to 54 percentage points between assessments, and girls increase from 22 to 23 points, their is no basis for claiming that the effectiveness of science intruction has improved more for boys than for girls. The difference of the changes may merely reflect the fact that a response to variation of an independent variable is almost always greater toward the middle of the percentage scale. Yet it is just such comparisons of differences in percentages for groups with widely differing attainment levels that make up the bulk of present NAEP reporting.

A second problem with average item percent corrects is their inherent vagueness when broad content areas are summarized. What meaning should the reader attach to a reported increase of 0.5 percent in the proportion of correct reponses to the items in the reading assessment? Is this a large or small change, and what does it imply for the material that children at this age can or cannot read? Because the absolute percentages have no meaning that is invariant with respect to altered item composition in the assessment, and because the degree of change is affected by the absolute level, the reader has no opportunity to learn a consistent system of numerical units on a fixed scale. Nor can the reader compare changes in any meaningful way from one subject-matter area to another. As is true for differant groups of pupils, the average level of percent correct in different areas may differ substantially because item difficulty cannot be accurately controlled in different types of content. These differences in level preclude pontentially interesting comparisons of progress in different subject matters as curricular emphasis changes. Multivariate analysis of subject-matter scores is ruled out by this approach.

Finally, and this point is stressed by Bock, Mislevy, and Woodson, there is no possibility with average-percent-correct reporting to update the assessment instruments while maintaining a consistent scale for measuring change. Because items tend to become obsolete over a period of years, and others must be released in order to illustrate the assessment results and to inform the public about the content of the assessment, the set of items that can be held constant for purposes of reporting change becomes smaller and smaller. The generalizibility of the reportable results deteriorates to a point, now nearly reached by NAEP, that the suite of comparable reports has to be broken. There is no possibility in this system to develop the type of long-term statistical time series for educational progress that is typical of other indicies of social and economic productivity.

1.2.1 The Item Response Theoretic (IRT) Method of Reporting Attainment Levels

Beginning with a paper by Lawley in 1943, and continuing in the work of Lord, Samijima, Bock, Andersen, and others, a new approach to test scoring has developed, based on responses to individual items rather than the number-right score for the test as a whole. In the modern approach, a model giving the probability of a correct response to each item as a function of a scale score for the examinee is fitted to item data from a large sample of persons drawn from the population of interest. The scale score of a particular examinee can then be estimated from his or her pattern of correct or incorrect respones to any set of items that have been fitted in this way. All such estimates are on a scale with common origin and units; thus they are directly

comparable, and items may be added to or deleted from the set without affecting comparability. In contrast to number-right and percent-correct scores, which do not have this property, the resulting scale score is called "item invariant." The item-invariance property effectively solves the problem of updating a test instrument as items are retired and new items are added. It also provides a method of equating scores of separate instruments, such as those of the national and state assessments, through the medium of linking items. Applications of IRT methods for this purpose have been reviewed recently by Lord (1981).

1.2.2 The Concept of a Linearly Ordered Content Domain

That the probability of correct reponses to each item in the set defining the scale can be calculated from the examinee's scale score gives these scores a much more concrete interpretation than percent correct scores. One can say that a person with a scale score of X on a given kind of item has mastered that type of item because he or she has, say, eighty chances in one hundred of answering such items correctly. The fitted item-response models identify classes and orderings of items that have similar response probabilities at the same scale score levels. To the extent that items in these classes have similar characteristics, one can infer the type of knowledge or skill represented by the score level. (An example of this type of content referencing appears in Bock and Mislevy, 1981.)

The scale on which the scores are expressed corresponds to an ordering of classes of items from high probability of correct reponse (easy items) to low probability (hard items). In the context of measuring levels of school attainment, we expect these orderings to arise from the order in which pupils encounter and master various domains of curricular content. There are good reasons for believing that subtantially homogeneous orderings exist within many content domains even at the national level. Much of the ordering arises from the way textbooks and other educational materials are graded and sequenced. Especially in the more structured subject matter, such as mathematics and grammar, the same topics tend to be introduced at given grade levels and in similar order within grade. Materials for subjects such as reading and spelling are often ordered according to the results of the same statistical studies of children's reading materials.

Further ordering is imposed by the hierarchical nature of cognitive tasks: multiplication requires addition; division requires multiplication and subtraction; clauses are constructed from phrases, sentences from clauses; and so forth.

A still higher level of order grows out of the developmentally increasing capacity of children as they mature. As short- and long-term memory expand, longer and more complex material can be presented. As social awareness increases, more thoughtful analyses of history and literature become possible. As the capacity for abstract reasoning develops in adolescence, higher level mathematical and scientific topics can be introduced.

Pupils in different schools and different pupils in the same school will, at a given age, have progressed to different points along the linearly ordered content domains. When we estimate a scale score for a given pupil, we identify the point that is most probable given the item responses we have

observed from that pupil. Typically, we locate the pupil at the point where he is just on the threshold of mastering the class of content at that point—the threshold of 50 or 80 percent mastery may be chosen for this purpose. The pupil's score is then said to be "domain referenced," meaning that the content of the domain can be divided at the threshold into that which has been mastered and that which is still to be mastered. Once the item classes have been identified and ordered on the scale, the scores admit of this concrete and intuitively meaningful interpretation.

1.3 Consistent Item-Invariant Attainment Scales

The domain definitions established by the assessment content committees determined the qualitative nature of the attainment scales, and the item analysis in which the IRT models for the items are fitted would determine the ordering of content on the scale. The origin and unit for the scale are, however, arbitrary, and it is advantageous to set them so that the mean and standard deviation of the scale scores in the population have convenient values in some base year. Since the scale used by Educational Testing Service, having a mean of 500 and standard deviation of 100, is now widely known, this convention would be a good choice for the assessment base year.

If the domain attainment levels are reported on such a scale for a number of years, and the origin and unit remain constant, the numbers will begin to take on a more and more definite meaning. The size of typical year-to-year changes will become known; the increase in average scale score between grades in the assessment design will show the effect of a year of instruction at different grade levels; the differences between demographic groups expressed in standard deviations of the total population will be available; most important, the referencing of the scale to typical item types at various points will convey the behavioral significance of the scale score levels.

With scale score reporting, graphs can be drawn showing progress in each domain and subject-matter area for an indefinite number of years (see the discussion of scale maintainence in section 1.6). These graphs will provide visual comparisions of levels and trends for the various subpopulations and for different types of content. (With average-percent-correct reporting, such graphs can extend only over the number of years that sufficient unreleased items remain from the original set. If new items are introduced and overall difficulty levels change, the graphs will show a break between years. See, for example, the figures in Burton and Jones, 1982). Fortunately, methods of estimating scales scores that can be applied retrospectively to the existing NAEP data are now available. In those content domains for which sufficient items exist in the NAEP assessments already completed, it will be possible to calculate scale scores for an articulated graphical display of trends in attainment from the first years of the assessment and continuing into the future.

1.4 Item Calibration and Scale Score Estimation in Assessment Data

IRT methods developed for scoring individual subjects (see Lord, 1981) do not apply directly to assessment data because the matrix sample provides only a few responses from the same person within the content domains to be scored. Variants of these methods suitable for matrix samples have been developed, however: Bock, Mislevy, and Woodson (1982) describe the method

used in the California Assessment Program based on direct estimation of scale scores for schools. The assessment instrument for CAP is so constructed that only one item in each of the narrow content domains called "elements" appears in each form. Thus, each pupil answers only one item per element, and the reponses to the items that make up the element are independent within schools. This means that, with respect to the sample of responses within school, the number of correct responses to each item is binomially distributed. A response model, similar to the standard IRT model for Poisson variables, can therefore be constructed and used in item calibration and scoring directly at the school level. Great economy of computation results because the data file consists of the number of attempts and number correct for each item in each school. The size of this file is of the order of the number of schools rather than the number of pupils.

This one-item-per-form approach to matrix-sample data has been applied by Reiser (1980) and by Mislevy, Reiser, and Zimowski (1982) to existing NAEP data, which is not reported by school, by assuming independence of responses within the subclasses of the high-order demographic classification of pupils in the national sample. By using only one item per form within a content domain, they are able to estimate main effects and interactions in the demographic design without resorting to scores for individual pupils.

Inasmuch as the sample design outlined in section 2 provides for moderately large samples of pupils within the selected schools, the California solution could be applied to the resulting data. For a multi-purpose national assessment, however, that approach has the distinct disadvantage of not providing any information about variation or covariation among pupils within schools. It is limited, for example, to describing the statewide variance between schools, rather than the total score variance including between-school and within-school components. Similarly, it can examine correlations of school characteristics with school attainment levels, but not the correlation of pupil characteristics with pupil attainment within schools.

Because within-school effects will be of interest in the national assessment, a more general method of analysis is needed. Such a method is available in the approach to analysis of incomplete data studied by Dempster, Laird, and Rubin (1977), Dempster, Rubin, and Tsutakawa (1981), and others. Bock and Aitkin (1981) and Mislevy and Bock (1982b) have shown how these methods can be applied to calibration of any parametric form of response model. In special cases, Andersen and Maddsen (1977) and Sanathanan and Blumenthal (1978) have applied similar methods to the estimation of group or population scale score distributions. Mislevy (1982) has generalized the latter to apply to any form of response model and various characterizations of the group or population distribution.

These methods, based on marginal maximum likelihood estimation, are ideal for scaling NAEP data because they provide information at many levels for a variety of audiences. Most important are the estimated school means, with large-sample standard errors, for each content domain, and means for the main classes of pupils within each school. Thus, for example, a mean can be calculated for all nine year olds in any given school sample, for pupils in grade 3 or grade 4, for nine-year-old boys or girls, or for nine-year-old boys and girls in the major sociocultural groups. The averages of these means, weighted by the number of pupils in age group or grade in the school, and by

the school weights for the probability sample, provide the population-level statistics describing the current status of attainment in the assesment year. At the same time, the change estimates can be computed from these means in two successive years, or trends over three or more years can be estimated. From the standard errors of the means, the school weights, and the between-year correlations, standard errors for the population means and change means can be calcuated to show the precision of estimation.

Another level of analysis, aimed at a more technical audience, is available in the variance and covariance component estimates that can be computed within and between schools. Because all possible pairs of domains appear together in a balanced manner in the matirx sample, covariances between domain scores can be obtained at the pupil level by marginal maximum likelihood estimation. In this way, domain covariance matrices within subject-matter areas can be computed, either for single schools (when the school sample is large enough) or for homogeneous groups of schools. (Covariation between subject-matter areas probably would not be of interest, but could be obtained if necessary.) These covariance matrices would not necessarily be positive-definite, but could be made so by reconstructing estimates of them from an unweighted-least-squares factor analysis solution.

Because data from the matrix sample are complete at the school level, the between-school covariance matrix can be estimated by standard methods for one-way designs with unequal numbers of cases per subclass (see Searle, 1971). In the present context, the subclasses are schools, and the school weights, incorportating the number of pupils per school, play the role of numbers of cases within subclasses.

From the within- and between-school variance-covariance components, the variances and covariances for the total population of pupils or for subpopulations can be estimated even though scores for individual pupils were not calculated at any point in the analysis. Methods for this purpose are discussed in Mislevy (1982). This is an example of the potential for matrix-sampled assessment data of recent advances in the analysis of incomplete data by means of marginal maximum likelihood estimation.

Yet a third level of analysis is accessible when collateral information about schools or pupils is brought into play. Again, because data are complete for schools, the analysis of relationships between attainment as measured by domain or area mean scores and quantitative or qualitative information about schools is entirely straightforward. With some modifications to allow for school weights when relationships for the population as a whole are desired, standard univariate and multivariate least-squares methods can be applied. These analyses are economical to carry out because only the school summary file, consisting of perhaps 450 schools at each age level, is required. Both the ease of anlysis and the relatively extensive collateral information that can be obtained for schools from the principal's report and other sources should make the school-level analyses a popular and productive form of research.

Relationships between collateral information and pupil attainment within schools, in contrast, will have to be investigated by the marginal maximum likelihood methods that provide estimates of means, variances, and covariances from incomplete data. Some of these results will be contained in

the estimated means for types of pupils, such as boys and girls, within each school. Similar results could be obtained for other classifications of pupils, such as type of program pursued in high school. By related techniques, correlations between quantiative information, such as number of years at present residence, can be estimated for pupils within given schools or homogeneous groups of schools. These analyses make use of the pupil file and will necessarily be more costly than between-school analysis if the full sample of perhaps 24,000 pupils per age level is employed. For most purposes, however, sampling of the pupil file would be quite satisfactory. In fact, the marginal maximum likelihood methods can be applied to any aggregation of pupils, ignoring schools, by using case weights in the calculations. The resulting analysis would be similar to the analyses of NAEP data that have been carried out up to this time, in which smaller numbers of pupils are sampled within schools and school membership is ignored in calculating the case-weighted average item percent correct for content areas.

1.5 Choice of Response Models, Item Analysis Procedures, and Computer Programs

The marginal maximum likelihood procedures that for the first time allow the application of item response theory to efficient item-sampling designs like that of NAEP can be used with any IRT model proposed to date, including multidimensional and multiple-category response models. This fact permits great flexibility in choice of item response models for the analysis of National Assessment data.

The selections should be driven by two considerations. First, models must be selected that exploit the information inherent in the data. While the more familiar models for dichotomous data will be approporiate for the majority of items designated for attainment measures, the formats of other items clearly call for a more encompassing model. The ratings generated in the Writing assessment and Likert-scale attitude measures require models for ordered response categories; codings that reflect distinct but unordered classifications require nominal categories models. Second, within the class of models appropriate to a particular class of items, the principle of parsimony should guide final selection. Statistical tests of model fit will provide guidelines for selecting models that capture the essential features of NAEP data without overparameterization.

Computer programs embodying marginal maximum likelihood procedures have been developed for use in the assessment setting with one-, two-, and three-parameter logistic models for binary data (Mislevy and Bock's BILOG, 1982a) and for ordered and nominal logistic models for multiple-category data (Thissen's MULTILOG, 1982). (The former is currently used on a production basis in three large-scale assessments, namely, the California Assessment Program, High School and Beyond, and the Second International Study of Mathematics.) Procedures requisite to NAEP are considered below.

Measurement scales must be established in the first assessments employing our revised domain specifications, through the estimation of item parameters. Both BILOG and MULTILOG are able to provide estimates of item parameters from item-sampling designs in which each subject is presented as few as the five items per scale anticipated in the NAEP design--a capability not possessed by any program that must, in the course of estimating item

parameters, also estimate scale scores for individual subjects (e.g., LOGIST, by Wood et al., 1976). Both programs provide global tests of model fit, by which the comparative fit of more and less parsimonious models can be examined, and item diagnostics, by which flawed or misclassified items can be identified.

It may be noted that the traditional procedures employed for selecting items in the setting of individual measurement, such as difficulty and item-test correlation, are not optimal in the assessment setting. Content coverage and scale definition, as determined by expert judgment of content-area specialists, as well as model fit, are more important than discrimination among persons. Indeed, person measurement is proscribed by intent; NAEP is charged with the estimation of population attributes rather than individual differences.

To provide measurement on invariant scales over time, it is necessary to obtain estimates of item parameters from previous and subsequent assessment on the scales established in the initial assessment in each content area. Again, both BILOG and MULTILOG share the capability of estimating parameters of new items from joint patterns of response to new items and previously calibrated items.

Two approaches have been widely proposed in the literature for linking new item parameters into an existing scale. The first would be to calibrate items separately in data from each assessment, then determine from common items the linear transformation that produces the best match of item parameters among common items. This approach suffers from the multicollinearity associated with the estimation of item parameters in producing response curves in multi-parameter IRT models. A second approach uses only data from the second assessment, estimating parameters for new items while the parameters of previously calibrated items are held fixed. This approach is faulted in that these parameter values are not known values but imperfect estimates. A modification of this latter approach, already implemented in the BILOG program, corrects this fault: item parameters from both new and previously calibrated items will be estimated from the responses to the new assessment, but with Bayes priors on the parameters of the previously calibrated items. In this way the information about the parameters of these items from previous assessments can be incorporated, but with their precision appropriately taken into account. We recommend the use of BILOG, which is commercially available, for NAEP because of the program's unique capabilities.

Marginal maximum likelihood algorithms to estimate the distribution of attainment within the populations and subpopulations of interest have also been developed, and have been employed in the Second International Study of Mathematics. As required in NAEP, even the most sparse item-sampling data may be utilized; it is never necessary to estimate a score for any individual. The statistical theory upon which these programs are based is presented in Mislevy (1982). It will be noted that these procedures not only handle distributions of attainment in a single domain, but also extend to the joint estimation of attainment in multiple domains, and between attainment domains and other pupil characteristics such as attitude and background measures.

The routine calibration of item parameters, linking of assessments,

and estimation of attainment distributions in NAEP will of course require a large number of program setups and job runs. If the NORC consortium is conducting NAEP, these tasks will be automated to a large extent. Macro-level programs will handle interface bewteen the NAEP database and the special-purpose psychometric programs through NORC's SIR database management system. Neither excessive time nor psychometric expertise will be required of the staff members who carry out the routine analyses; only specifications of items, assessment years, skill domains, and subpopulations of interest need be specified.

The programs described above could be made available in one form or another to prospective users of NAEP data. NAEP staff would regularly conduct workshops and seminars on the analysis of NAEP data at the assessment center we would propose to establish (see section 3), at professional meetings such as that of the American Educational Research Association, and at invited conferences.

1.6 Maintenance of Item Domains

As discussed above, a technology based on item response theory promises to provide measures of attainment on invariant scales over time in the context of evolving item pools and intense item-sampling data collection. The mere fact that IRT models are used, however, does not guarantee in and of itself that this promise will be fulfilled. If the requisite assumptions upon which the IRT models are based are not suitably satisfied, the desired invariant scales will not be forthcoming. Indeed, large-scale educational assessments employing IRT in Los Angeles and in Great Britain have recently been subject to severe criticism (e.g., Goldstein, 1980) for just this reason. This section briefly reviews the thrust of those criticisms, then outlines our approach to maintaining the integrity of scales.

An item response model is able to provide measures on an invariant scale, regardless of which particular items in the scale are used for measurement, only when patterns of response to all the items in the scale can be suitably explained in terms of a single hypothetical variable, namely, the scale score. One implication of this assumption is that for a scale to maintain its integrity over time, changes in performance must be roughly proportional (in the log-odds scale) over all the items in the scale. If two substantially discrepant skills were calibrated together--mathematical reasoning and arithmetic operations, for example--and performance over time increased in one skill but decreased in the other, no single score would be able to explain both trends. Trend analyses would thus be degraded because of a failure of the assumption of a single underlying variable (or, equivalently in this case, the failure of the assumption of local independence). The single combined scale is poorly defined and ill-suited to the purpose for which it was intended.

This problem of scale instability can also be explicated in terms of item parameter estimates. If the IRT model holds, then the parameters for each item in the scale will be identical within the limits of precision of estimation over all time points they are presented. If the scales are not well-defined, however, different trends among different subclasses of items will lead to substantial discrepancies among item parameter estimates from different time points--that is, item parameter drift.

Our strategy for maintaining the integrity of measurement scales has been developed over the past five years by our psychometric staff in conjunction with the California Assessment Program, which uses IRT models in its annual assessments of educational attainment. The strategy is to define the scales in which all IRT item calibration and attainment estimation will take place within skill areas defined along the lines of linearly ordered content domains, so as to best satisfy the assumption of unidimensionality and local independence, and to construct global scales as rationally weighted linear combinations of domain scores. If the NORC consortium conducts the NAEP, the five-step plan we have developed with and used in the California Assessment will be carried out:

1. Identify linearly ordered content domains in which all calibration and estimation will take place.
2. Through a balanced braiding design, link each assessment in a given content area to assessments of one, two, and five years previous.
3. Using automated procedures, compare parameter estimates of items obtained in successive assessment years.
4. Split off as separate measurement domains groups of items that exhibit a consistent pattern of change that is opposed to the pattern in the domain in which they have been included.
5. Retire items that show idiosyncratic patterns of drift.

In this manner, consistent measurement in well-defined scales can be guaranteed at the level of skill domains. The manner in which domain results are aggregrated into more global content areas such as Reading or Mathematics must remain, in essence, a political task in the sense that rational selection of weights for combining disparate trends must be agreed upon. This task is charged to the APC committee. In order to reflect new national emphases, the committee may on occasion deem it necessary to delete certain domains, add new domains, or reweight existing domains that make up the composite defining a content area score.

2. SAMPLE DESIGN

2.1 Introduction

The sampling design that we recommend is substantially more efficient than the design that has been in use under ECS. The efficiency gains allow for several improvements: more frequent assessments; more informative reporting of findings; reduced cost of data collection; and increased precision. The proposed sample design will be described below. First we note some important implications and features of the design:

1. The design will allow annual reporting of achievement in each of the major subject areas (reading, writing, and mathematics). The precision of the annual results will be as high as or higher than the quadrennial results now obtained by ECS. The design allows for the testing of one additional topic each year.

2. Achievement can be reported for the same subgroups of the population as currently done by ECS, and, in addition, for: grades 3, 4, 7, and 8; public schools, private parochial schools, private non-parochial schools; each individual school participating in NAEP.

3. The greater efficiency of the sample design will result in reduced cost and greater precision. Some important aspects of the design are: three-stage stratified probability sampling, in some ways similar to the current design and in some ways different; larger numbers of students tested; fewer schools tested; schools tested two years consecutively (rotation sampling).

Our sample design is one means by which NAEP would be able to deliver more information, more often, and with equal or higher reliability than current practice. The sample would be designed so that the reliability of estimates is as high as possible. We are recommending a design that is in some ways the same and in other ways different from previous practice. The features that are the same include:

- The same grade ranges will be tested
- The overall selection--that is, first geographic areas are selected, then schools within areas, and then pupils within schools
- The number of schools sampled will be approximately the same

The departures from previous practice are:

- Substantial increases in the numbers of pupils tested
- Use of a method of sampling known as "rotation sampling."

Each of these changes by itself will give more precision for about the same amount of money. Over and above these increases in precision there will be other increases, due to the use of modern testing techniques such as item response theory (Section 1). For the most part, we restrict attention in this chapter to increases in reliability coming from more efficient sampling, independent of the fact that we are using additional analytic methods.

2.2 Overview of Sample Design

The sample design we recommend would represent a stratified, three-stage probability sampling with rotation. First primary sampling units (PSUs) consisting of groups of contiguous counties would be sampled. After the first year, each PSU selected would appear in the sample for two consecutive years. (This technique, known as rotation sampling, will be described below more fully.) Second, schools within PSUs would be sampled, with oversampling of private schools for precision in public-private comparisons. Third, students in grades 3, 4, 7, and 8 and 17-year olds in grades 10, 11, and 12 would be sampled from the selected schools.

2.2.1 First-stage Sampling

At the first stage the nation would be divided into geographical units composed of counties or groups of contiguous counties. These units are called primary sampling units or PSUs. The PSUs would be formed to contain a certain number of students and also to contain a heterogeneous collection of schools. From the list of PSUs a stratified random sample is drawn with probability proportional to size measures. The stratification would guarantee adequate numbers of schools to provide precise estimates for all regions of the country, various sizes of communities, and urban and rural parts of the country. To improve overall precision we would stratify PSUs in the rural West by Hispanic composition and in the rural South by black composition.

2.2.2 Second-stage Sampling

In the second stage, for each PSU drawn into the sample, all public and private schools would be listed. Schools would be stratified several ways, including: grade-ranges of students; whether public, private parochial, or private non-parochial; whether urban or rural; racial/ethnic composition of students; and school size. First, schools would be selected from the strata with probability proportional to the number of students in the grades of interest such that all strata were represented. Then additional schools would be selected from the private parochial and private non-parochial strata to provide for precise public-private student comparisons.

2.2.3 Third-stage Sampling

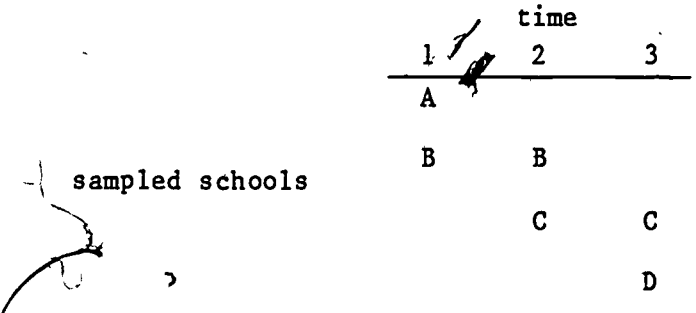
The third stage of sampling would occur during the data collection. Samples of students from grades 3, 4, 7, and 8 would be tested. For small schools all students in the appropriate grades would be tested. In addition,

from each selected school, a list of 17-year olds in grades 10, 11, and 12 would be obtained and a random sample of those students would be drawn. The increased number of students tested (in grades 3, 4, 7, 8 and 12) would allow preparation of reliable reports for the schools participating in the assessment. It is inevitable that some students who otherwise will have been tested would be absent. The names of those students would be listed and schools revisited to test them.

2.3 Rotation Sampling and Increased Efficiency

Although rotation sampling represents an innovation for NAEP--it has never been used by ECS and no other first preliminary proposal even mentioned it--rotation sampling is nonetheless a well-known technique for reducing cost and increasing reliability; see the authoritative sampling textbooks of Kish (1963), Cochran (1977), or Hansen, Hurwitz, and Madow (1953) for discussion. The Census Bureau uses rotation sampling in its Current Population Survey and in its Retail Trade Survey to improve precision both of cross-time comparisons and of single-time estimates.

For a simple illustration of rotation sampling, consider estimating overall achievement levels at time points 1, 2, 3, etc. At time 2 we measure achievement again in school B but replace school A by another school, say C, and measure achievement there. At time 3 we replace school B and D and measure achievement in school C again. Schematically we represent this as follows.



To estimate change in achievement from time 1 to time 2 one could take the change in average achievement in schools B and C at time 2 and subtract the average achievement in schools A and B at time 1. This estimator will be called the simple estimator. The appearance of school B in the sample at both time points greatly stabilizes the estimate of change, the reason being that a school's achievement levels over time are highly intercorrelated. We will use the letter r to denote the correlation between a school's achievement levels in consecutive years. It is well-known that a better estimator than the simple estimator is a composite estimator that weights the change in school B more heavily than the difference in achievement between schools C and A.

Use of rotation sampling in conjunction with either the simple or the composite estimator provides impressive gains in efficiency for estimating yearly change. To make this notion precise, we define relative efficiency as the effective sample size equivalent to 100 schools in the proposed rotation sample design. The efficiency gains are particularly large when the composite estimator is used. Thus, Table 1 shows that if the year-to-year correlation

r is .85 and the simple estimator is used, then 100 schools in the rotation design could yield as much precision as 174 schools in the current ECS design; this means that costs could be reduced by 43 percent (= 174 - 100/174). If the composite estimator is used, then 383 schools are needed in the ECS design to give as much precision as 100 schools in the rotation design; this could mean a cost reduction of 74 percent (= 383 - 100/383).

TABLE 2.1

Relative Efficiency* in Estimates of Yearly Change: Various Levels of Correlation r

Correlation r	Current Design (no rotation)	Rotation Design (composite estimator)	Rotation Design (simple estimator)
.60	100	143	175
.70	100	154	217
.80	100	167	300
.85**	100	174	383
.90	100	182	550
.95	100	190	1,050

*Relative efficiency reflects the effective sample size of 100 schools in the proposed rotation design.

**Empirical evidence indicates r is .85 or higher.

Rotation sampling also improves the precision of achievement estimates for a single point in time. The gain in precision derives from the use of composite estimators. Such estimators represent an innovation for NAEP; however, they have been used for many years in the Current Population Survey to improve the precision of labor force estimates. The gains in efficiency for single-time estimates are not as enormous as for estimates of year-to-year change but they are still substantial. Table 2 shows that if the year-to-year correlation is r then 100 schools in the rotation design yield as much precision as 129 schools in the current ECS design. This means that by using rotation sampling the number of schools sampled could be reduced by 22 percent while the current precision in single-year estimates of achievement was maintained.

TABLE 2.2

Relative Efficiency* in Estimates of Yearly Achievement: Various Levels of Correlation r

Correlation r	Current Design (no rotation)	Rotation Design with Composite Estimator
.60	100	111
.70	100	116
.80	100	124
.85**	100	129
.90	100	134
.95	100	141

*Relative efficiency reflects the effective sample size of 100 schools in the proposed rotation design.

**Empirical evidence indicates r is .85 or higher.

2.4 Sampling Students in Schools

As described above, schools would be tested in two consecutive years. Also, at the two lower age levels we would recommend testing students at grades 3, 4, 7, and 8. We would recommend testing not just 9 year olds or 13 year olds, but rather samples of every age student in those grades.

For schools in the sample containing the highest age group (17), we would obtain a list of 17 year olds in grades 10, 11, and 12 and randomly sample 17 year olds. For the schools containing the two younger age groups (9 and 13) we would sample students by grade.

We plan to test at least 120 students per age group per school for the two lower grades and 120 seventeen year olds from the high schools. Where there are fewer than 120 students available for testing, we will test all students in the grade (3, 4, 7, or 8) or age group (17 year olds).

Testing so many students per school accomplishes several things: it increases the year-to-year correlations r discussed above, it permits use of item-response models for analyzing and reporting the achievement of a single school grade or age group, and it improves the precision of the estimated achievement level for a school.

As discussed in the appendix, increasing the number of students tested per grade to 120 would increase the precision of school-level achievement estimates by a factor of at least two. Since overall (e.g., national or regional) achievement estimates are weighted averages of school-level estimates even if we did not recommend rotation sampling but instead recommended the same number of schools and PSUs as ECS currently does, the precision of the achievement estimates in any assessment area would be at least double the current precision. However, our estimates will be produced each year for each of the three major assessment areas while ECS's estimates have been produced only every three or four years.

2.5 Sample Sizes

Currently ECS samples approximately 75 PSUs and 450 schools per age level. Use of rotation sampling would allow reduction of the number of schools per age level to 360 and still yield more precision than ECS. The increase in the numbers of students tested per school that we recommend would further increase the precision.

3. STRENGTHENING NAEP'S LINKAGES WITH STATE AND LOCAL ASSESSMENTS TO ENHANCE THEIR ABILITY TO IMPROVE EDUCATION

This was the second of the major recommendations in our preliminary proposal. We will detail our plans for implementing this recommendation in the major proposal to be submitted to NIE later this month. In this report, we outline the structures, staff, and activities that we see as essential to strengthening state linkages.

We recognize that there is a tradition of links between the National Assessment and the states. But, as we said in our preliminary proposal, "the current relations of NAEP with state assessments fall far short of the potential for contributing to the improvement of state and local education systems." We recommend that the existing links--and some of the specific expressions of them, such as the Annual Conference--be maintained. But we recommend further that these links be strengthened and that the state-national assessment partnership be expanded. In our preliminary proposal we spoke of the benefits to the states of association with the National Assessment. We continue to believe that the National Assessment has offered much to the states and can offer much more. But the investigations we have conducted under the preliminary grant have made us aware of what the states have to offer the national assessment effort--and the importance of bidirectionality in the national-state relationship.

We recommend that the following steps be taken to ensure that the links between the national and state assessments are strong, productive, and beneficial to all.

Structure and Staff

1. Increase state participation in activities of the APC
2. Designate an area of senior responsibility for state and local relations within the NAEP management structure
3. Create a State Assessment Clearinghouse
4. Provide an opportunity for state assessment personnel to participate in NAEP research

Assessment Activities

1. Provide states with a range of options for working with the National Assessment
2. Employ methods that will allow greater item sharing between the national and state assessments
3. Develop technical standards acceptable to both the national and state assessment efforts
4. Provide a translation between national and state data for states that participate in the network

Supplementary Activities

1. Expand the scope of the Annual Conference
2. Establish relationships with existing state assessment associations
3. Improve the dissemination of information about state assessments
4. Provide performance feedback to participating schools

We believe that the items listed above constitute a sound basis for strengthening NAEP's linkages with state assessments. We will now discuss each of the items listed in greater detail. Discussing these items in more specific terms sometimes requires that we depart from the language of the report, turning to the language of the proposal. And there will in fact be considerable ovelap between what follows and the proposal we submit to NIE later this month.

3.1 Structures and Staff

3.1.1 Increase State Representation on the APC

An essential part of our full proposal to NIE will be a strengthening of the role of the APC, to allow it to fulfill its legislative mandate in fact as well as form. An important element in our proposal in this regard will be a subcommittee structure to share and support the work of the APC. These plans were in part motiviated by our desire to enhance national-state linkages, and we believe that they will have a salutary effect when implemented. The states are now represented on the Assessment Policy Committee at many levels: a governor, a chief state school officer, and state representatives; and we will propose in addition that a state assessment director be included among the APC members. Because each member of the APC has a dual function--to bring to the committee the collective interests of the group represented and to bring back to that group information about APC decisions and the status of the assessment, this constitutes a strong and bidirectional national-state involvement. But, in spite of this strong representation of state interests at the policy level, it is not possible to represent adequately all state-level interests on the Assessment Policy Committee. The NORC Consortium proposes to include more of those interests in the subcommittee structure. For this reason we will propose a State Assessment Subcommittee as a working resource for that representative on the APC. One of our primary concerns in proposing the subcommittee structure is to provide these working committees with communication channels and resources that are separate from the assessment channels. Through the State Assessment Subcommittee, state-level assessment personnel and interest groups concerned about state assessments will have direct access to the APC, both to provide information on issues before the APC and to obtain information for their constituencies directly from the assessment policy group.

3.1.2 Designate an Area of Senior Reponsibility for State And Local Relations Within the NAEP Mangement Structure

We believe that it is vital that a state and local relations group be one of the major units in the structure of NAEP operations and that a full-time member of the NAEP staff at the senior level have the direction of this group as his or her primary responsibility. We can see no other way that this area, so critical to NAEP's success in operating successfully and in fulfilling its potential, can be adequately attended to.

3.1.3 Create a State Assessment Clearinghouse

Many states have substantial resources already devoted to assessment efforts, but the amount and kind of resources available to states for assessment activities vary widely. We recommend that a State Assessment Clearinghouse be established so that the national as well as other state assessments can learn from state assessments that are leaders in the education assessment field.

A State Assessment Clearinghouse is essential for a number of reasons. It will provide a systematic and ongoing record of the various activities in the states that conduct assessments. It will allow these states--as well as those that do not now conduct assessments--to take advantage of one another's development efforts and to learn from one another's mistakes. And it will array before the national Assessment the full range of state activities. This will inform NAEP's planning for coordination with the states and will allow the NAEP to take advantage of improvements in assessment methodology on the part of states.

The Clearinghouse would, we expect, be established by NAEP staff, under the direction of the Director for State and Local Relations. After the Clearinghouse was established, we would seek additional funding from outside sources for its continued functioning. We envision a library of material, a hotline, and a newsletter as components of its operation. We would also hope to encourage creative uses of the Clearinghouse resource. For example, the Clearinghouse, with technical assistance from the national assessment staff, might enable states with few resources for assessment to borrow items from the ongoing state assessments in other states if the national assessment itself could not be helpful (for example, in content areas not covered by the national assessment).

3.1.4 Provide an Opportunity for State Assessment Personnel to Participate in NAEP Research

State assessment personel have something to offer and something to learn in research using assessment data as well as in assessment operations. The NORC Consortium will propose that a center for the study of assessment be established as a related but separate entity. The center will house fellows interested in NAEP and other assessment data and issues. Some of these positions would be offered to state-level persons, both assessment and administrative personnel. The Fellowships would provide an opportunity for the assessment staff to learn about state-level issues from the points of view of the Fellows, and at the same time give the Fellows an opportunity to think about issues beyond the context of their home states.

3.2 Assessment Activities

3.2.1 Provide States With a Range of Options for Working with the National Assessment

States are already involved in education assessment. About forty states now conduct some sort of pupil assessment, although the extent and methodology of such assessments varies widely from state to state. This variation dictates flexibility in the relaions between NAEP and the state assessments if there is to be both maximal participation and respect for state autonomy. The varying relationships between state assessments and districts are a microcosm of the macrocosm we recommend of varying relationships between the national and state assessments. We recommend that the NAEP develop jointly with the states a series of flexible options so that each state can join the national assessment at the level of its choice. These options might include subcontracting large sections of the state assessment effort to the national assessment for the cost efficiencies involved; augmentation of the national assessment sample in the state to allow state level data to be provided; including some national assessment exercises in state assessment for translation to the national metric; use of assessment instruments, scoring, and data processing by the state; working with state assessment personnel to ensure that the state assessment data is collected in a manner comparable to national data; provision of a translation service between national assessment data and that of the state; participation in the Annual Conference or assessment workshops; and utilizing any of the helpful materials generated by the center for the study of assessment.

3.2.2 Employ Methods That Will Allow Greater Item Sharing Between the National and State Assessments

Our recommendation that the NAEP use IRT will allow the national assessment to provide much more assistance to state assessments. One by-product of our plan, when it is fully implemented, will be a very great expansion of the potential national assessment item pool. This expansion will make it possible to release many more items to states for use in assessments since there is no requirement to hold a large proportion of items in confidential reserve for retests in the future (see section 1 of this report).

3.2.3 Develop Technical Standards Acceptable to Both the National and State Asssessment Efforts

All formal arrangements between the National Assessment and the state assessments should begin with the development of an agreement about mutually acceptable technical assessment standards. The agreement should cover all major technical areas, including sample design, test administration, scoring, and so forth. The guiding principle of these arrangements should, of course, be maximum comparability, but the different needs of the state and national assessments must be taken into account. We believe that both of these things can be done. In the preceding section of this report, for example, we recommended that sampling (and reporting) be done by both age and grade to respond to the different needs of the two kinds of assessments. Test administration is perhaps the area in which most operational differences would arise. We would recommend that the National Assessment continue to employ NAEP test administrators, and we expect that state assessments will continue

to use school personnel for test administration. But we recommend the use of outside administrators primarily for reasons of confidentiality, not because the recommended methods of administration would be difficult for school personnel. We are certain accommodations could be made on all important points involving comparability.

3.2.4 Provide a Translation Between National and State Data for States that Participate in the Network

The design and analytic procedure recommended in sections 1 and 2 of this report will foster comparison of state and national data. Specifically, they will facilitate states' "piggybacking" their assessments onto NAEP--that is, concurrently administering their assessments to a supplemental sample of students to permit state-level reporting. It would be possible for a state to administer an assessment instrument that had no overlap of items with the concurrent NAEP, to a sample of students that had no overlap with the NAEP sample, yet still obtain measures of attainment on the NAEP scales in designated skill domains.

These objectives could be accomplished through the IRT scaling methods upon which we believe measurement in NAEP skill domains should be based. States could obtain estimates of attaiment in these domains on the NAEP scales by administering an assessment instrument that contains items from two sources: previous or concurrent NAEP items and supplemental items in the NAEP domains provided by the state itself.

It is the NAEP items that would guarantee comparability of NAEP and state-level attainment indices. Based on previous or concurrent estimates of the item parameters of these linking items, it would be possible to estimate the locations of supplemental state items on the same measurement scales; from the entire set of items in a given domain administered by the state, then, indicies of pupil attainment could be estimated.

NAEP technical support for states wishing to tie in with the National Assessment should will include workshops and consultations on appropriate analytic procedures. In addition, non-proprietary software developed for internal NAEP use should be made available.

3.3 Supplementary Activities

3.3.1 Expand the Scope of the Annual Conference

We would recommend that the Annual Conference be improved in a number of ways. This most important national assessment gathering has much unrealized potential for the promulgation and explication of NAEP results. Specifically, for using the Annual Conference to strengthen the national-state linkages, we do recommend that the state-level assessment Fellows take a leading role at the Conference, exploiting the experience they gain from working with NAEP data as Fellows and the experience with education at the state and local levels that they bring to their roles as Fellows.

3.3.2 Establish Relationships With Existing State Assessment Associations

As we have said throughout this section, we recognize that the state assessments are a very great resource, and one that should be tapped for the benefit of both the National Assessment and the state assessments themselves. One of the ways to do this is to have the National Assessment establish relationships with already-existing associations, formal and informal, of state assessments. We recommend, for example, that NAEP be represented at the meetings now convened by a dozen or so of the major state assessments. The exchange of ideas and information so fostered would surely benefit both of the parties participating, and could be made to benefit others through publications of the State Assessment Clearinghouse.

3.3.3 Improve the Dissemination of Information about State Assessments

Our major recommendation in this regard is the establishment of a State Assessment Clearinghouse (see 3.1.3, above). But other vehicles, whether part of the Clearinghouse or independent of it, could serve important purposes.

The Consortium recommends technical workshops on assessment measurement issues and techniques. State assessment personnel and Assessment Center Associates and Fellows would both lead and participate in these workshops. A Technical Newsletter could be developed in parallel with this workshop series and circulated to national assessment data users, state assessment personnel, and others interested in technical problems.

The NORC Consortium also recommends the use of existing networks to increase the circulation of assessment information and the implications of assessment results for policy and practice. For example, the State Education Policy Seminars provide an already-developed forum for presentation of assessment findings to a wide range of persons who make education policy decisions and implement those decisions in thirty states.

3.3.4 Provide Feedback to Participating Schools

Under present methods of analyzing the assessment data, no directly useful information is returned to the schools in the sample. The incentive for schools to cooperate is minimal, and schools cannot evaluate results and use them to modify programs. This may eventually affect participation rates. As stated in section 2 of this report, the sample design we recommend would allow NAEP to provide performance feedback to participating schools. As we noted in our preliminary proposal, analytical methods currently used by the California Assessment program for monitoring levels of performance in schools and districts could be adapted to provide reports for schools in the NAEP sample. The reports to each school indicate its position in the distribution of scores for the state as a whole, and also its position among schools in communities with similar socioeconomic characteristics. Relative strengths and weaknesses in various content areas are highlighted in these reports so that school officials and teachers can plan curricular and instructional improvements.

Similar reports could be prepared for the schools in the NAEP sample relative to the U.S. population of the schools. This type of information would not be reported publicly, but it would be of interest to the participating schools for possible curriculum revision and would be an asset in gaining the cooperation of new schools. We have carefully examined the California Assessment feedback materials and believe that they could serve as a model for NAEP feedback. We append and example from CAP. — not appended

The National Assessment is completely dependent on state personnel. We are able to conduct an assessment only with their cooperation. And the assessment can only affect education through the provision of timely and useful information to state officials, superintendents, principals, school board members, and classroom teachers.

APPENDIX 1: Analysis of Precision Increases Due to Rotation Sampling

Rotation sampling is a technique to improve the precision of recurring surveys. A well known technique to sampling experts, rotation sampling has long been used in major government surveys of the U.S. Bureau of the Census, for example in the Current Population Survey (Hanson, 1978). Rotation sampling represents an innovation for NAEP but it is not a new invention. The technique is described in standard textbooks on sampling (Hansen, Hurwitz, and Madow, 1953; Kish, 1963; Cochran, 1977). Rotation sampling is still being introduced into some surveys (e.g., the Census Bureau's Retail Trade Survey, as described by Wolter (1979)). The following discussion illustrates how rotation sampling provides efficiency increases of up to 250 percent or more for estimating year-to-year changes in achievement. Efficiency increases of 25 percent - 30 percent or more may be attained for estimating achievement levels in any single year.

By "rotation sampling" we mean specifically that schools will enter the sample, be tested in two consecutive years, and then will leave (rotate out) of the sample. Technically, this is known as one-level rotation sampling with 50 percent overlap. For clarity of exposition the following discussion will simplify somewhat, in that schools are assumed to have equal numbers of students enrolled and sampled, schools are assumed to be selected by simple random sampling, and all variances and covariances are assumed constant. The calculations of relative efficiency can be shown to be unaffected by these assumptions, which greatly facilitate exposition.

The rotation design will now be described (subject to above simplifications). Let $x(t,i,a)$ refer to the measured achievement level of school i in year t, and a denotes whether this is the first or second consecutive appearance of the school in the sample (a may equal 1 or 2 only). Imagine that $2M$ schools are sampled each year, as follows:

YEAR		
0	1	2
x(0,1,1)	x(1,1,1)	
.	.	
.	.	
.	.	
.	.	
x(0,M,1)	x(1,M,1)	
	x(1,1',1)	x(2,1',2)
	.	.
	.	.
	.	.
	x(1,M',1)	x(2,M',2)
X(0,1''',2)		x(2,1'',1)
.		.
.		.
.		.
x(0,M''',2)		x(2,M'',1)

Note that schools 1, ..., M appear in the sample years 0 and 1; schools 1', ..., M' appear in years 1 and 2; schools 1'', ..., M''' appear in year 0 but not years 1 or 2.

Some rotation is convenient. Let

S^2 = variance of $x(t,i,a)$

r = correlation between $x(t,i,1)$ and $x(t+1,i,2)$.

Calculations based on data from the California Assessment show that $r = .85$. The great advantage of rotation sampling derives from the autocorrelation r.

For estimating change from year 1 to year 2 we may use a so-called composite estimator (Cochran (1977), Hansen (1978), Wolter (1979)).

$$M^{-1}(2-r)^{-1}\left[(1-r)\sum_{i}^{M}\left(x(2,i',1) - x(1,i,2)\right) + \sum_{i}^{M}\left(x(2,i',2) - x(i,i',1)\right)\right] \qquad (1)$$

Similar estimators have been used in the Current Population Surveys since the 1950s. The estimator reduces to the simple average change if $r = 0$, that is

$$1/2\,M^{-1}\left[\sum_{i}^{M}\left(x(2,i'',1) + x(2,i',2) - x(1,i,2) - x(1,i',1)\right)\right] \qquad (2)$$

We will refer to estimator (1) as the composite estimator and estimator (2) as the simple estimator.

The relative efficiency of rotation sampling is defined as the number of schools that would need to be sampled if no rotation sampling were used, if we wanted to attain the same precision attainable with rotation sampling based on 100 schools. Thus, relative efficiency is the effective sample size equivalent to 100 schools in the proposed rotation sample design. The relative efficiency depends on the correlation r and the form of the estimator, i.e., sophisticated or simple. For the sophisticated estimator the relative efficiency is given by the formula

$$\text{relative efficiency} = 100 + 50r/(1-r) \qquad (3)$$

For example (see Table 1, below) if the correlation r = .85 and the sophisticated estimator is used then 100 schools in a rotation design give the same precision as 383 schools in a non-rotation design (i.e., the current design).

TABLE 1

Relative Efficiency* in Estimates of Yearly Change: Various Levels of Correlation r

Correlation r	Current Design (no rotation)	Rotation Design (composite estimator)	Rotation Design (simple estimator)
.60	100	143	175
.70	100	154	217
.80	100	167	300
.85**	100	174	383
.90	100	182	550
.95	100	190	1,050

*Relative efficiency reflects the effective sample size of 100 schools in the proposed rotation design.

**Empirical evidence indicates r is .85 or higher.

The preceding analysis assumed that if the ECS were used that estimates would be produced yearly and that the variance of any school's estimate would be the same as for our procedures. These assumptions are made for comparative purposes. Actually, the ECS design currently cannot produce yearly estimates. Furthermore, as discussed in another appendix, the variances of the variances of the school-level estimates will be lower for our proposed design than for the current design.

Rotation sampling can also improve the precision of estimates of the level of achievement in a single year. This improvement utilizes a composite estimator, such as

$$M^{-1}(4-r^2)^{-1}\left[(2-r^2)\sum_{i}^{M} x(1,i',1) + 2\sum_{i}^{M} x(1,i,2) + r\sum_{i}^{M}\left(x(0,i''',2) - x(0,i,1)\right)\right] \quad (4)$$

for estimating achievement in year 1. Discussion of the rationale behind this estimator may be found in the books by Kish (1963) or Hansen, Hurwitz, and Madow (1953), and its use in the Current Population Survey is described in Hansen (1978).

For estimating the level of achievement in a single year the relative efficiency of the composite estimator, compared to the estimator currently used by ECS, is given by the formula

$$\text{relative efficiency} = 100 + 50r2/(2-r2)$$

For example (see Table 2, below) if the correlation $r = .85$ and the composite estimator is used then 100 schools in a rotation design give the same precision as 129 schools in the current ECS design.

TABLE 2

Relative Efficiency* in Estimates of Yearly Achievement: Various Levels of Correlation r

Correlation r	Current Design (no rotation)	Rotation Design with Composite Estimator
.60	100	111
.70	100	116
.80	100	124
.85**	100	129
.90	100	134
.95	100	141

*Relative efficiency reflects the effective sample size of 100 schools in the proposed rotation designs.

**Empirical evidence indicates r is .85 or higher.

The estimators described above will not be used exactly as specified above because the correlation r will not be known exactly. However, we will be able to specify r fairly closely, with the effect that our tabulated relative efficiencies will still be approximately correct. For example, if we mistakenly used r = .75 in the composite estimator (4) when in fact r was truly .85 then the relative efficiency would be 127 rather than 129. The use of rotation sampling, even without other improvements discussed elsewhere, thus allows us to maintain or improve current levels of precision while reducing the number of schools sampled by at least 20 percent.

APPENDIX 2: Analysis of Precision Increases Due to Testing Larger Numbers of Students

The following discussion illustrates how we will be able to increase efificency to achieve at least twice as much precision (on a per school basis) for the same achievement statistics that are currently produced by ECS. This increased precision is attained despite the fact that our statistics will be produced each year although ECS produces them only once every three or four years.

The key to the gain in precision is the increased numbers of students tested per school sampled. Currently, ECS samples 10 to 35 students per age group in each school, "depending on an estimate of the rate of nonresponse for that school" (Three Assessments of Science, 1969-1976: Technical Summary, p. 41). For calculation purposes, the average number of responding students in the ECS design may be taken as 15 (this is a generous figure, and it in fact is 25 percent higher than the planned numbers for ECS's first two science assessments). We propose to test at least 120 students per age (9, 13, 17) per school (unless fewer students are available, in which case they will all be tested). For the two lower age groups, we plan to sample entire classes of students.

The gain in precision will now be calculated. We typically will use 15 forms per assessment per age level, so that 1/15 of the students tested will receive any given item, i.e., for each age level eight students per school will receive the item. ECS also used roughly 15 forms per assessment per grade level, hence on average only one student per school would receive any item. For any given item the variance of our achievement estimate will be only of the variance of ECS's achievement estimate. This gain will be somewhat offset for the usual achievement estimates for groups of items because

ECS administers four times as many items per assessment group on each form (in those years in which the assessment area is measured). The increased number of items does not cancel out the 1/8 reduction in variance. The reason is that under the ECS design a single student will respond to four times as many items in an assessment area but since any single student's responses will be highly intercorrelated the advantage of using four times as many items is not a factor of 4 but only 1.3 (if a student's intercorrelation is .7) or perhaps 2 (if the intercorrelation is .5). To be conservative, assume the advantage is 2 and thus note that for the proposed design the variance (per school) for the estimated achievement on a group of items will be 2 v 1/8 or 1/4 the variance of ECS's figures.

Essentially, we get our increased efficiency by testing more students per school, which enables us to administer fewer items per assessment area while improving the precision of the results. Needing to administer fewer items per assessment area allows us to increase the number of areas assessed each year. There is a slight cost increase from testing more students but this increase is actually very marginal. Furthermore, use of rotation sampling, described below, will allow us to test in fewer schools while increasing the precision provided by data on each school. Reducing the number of schools in the sample allows for substantial cost reduction. Finally, further increases in precision are obtainable using statistical modeling such as item response theory, but that is discussed elsewhere. It must be emphasized that the gains in efficiency provided by our sample design are not based on assumptions that may or may not hold. Rather, we gain efficiency by taking advantage of correlation structures which we know exist, based on empirical study.

REFERENCES

Andersen, E. B. and Madsen, M. "Estimating the Parameters of a Latent Population Distribution," Psychometrika, 42, 1977: 357-374.

Bock, R. D., and Aitkin, M. "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," Psychometrika 46, 1981: 443-459.

Bock, R. D., and Mislevy, R. J. "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, Number 10. San Francisco: Jossey-Bass, 1981.

Bock, R. D.; Mislevy, R. J.; and Woodson, C. E. "The Next Stage in Educational Assessment," Educational Researcher 11, 1982: 4-11, 16.

Burton, N. W., and Jones, L. V. "Recent Trends in Achievement Levels of Black and White Youth," Educational Researcher 11, 1982: 10-14.

Cochran, W. G. Sampling Techniques 3rd ed. New York: John Wiley and Sons, 1977.

Dempster, A. P.; Laird, N.; and Rubin, D. B. "Maximum Likelihood from Incomplete Data Via the EM Algorithm (with Discussion)," Journal of the Royal Statistical Society 39, Series B, 1977: 1-38.

Dempster, A. P.; Rubin, D. B.; and Tsutakawa, R. K. "Estimation in Covariance Components Models," Journal of the American Statistical Association 76, 1981: 341-353.

Goldstein, H. "Dimensionality, Bias, Independence, and Measurement Scale Problems in Latent Trait Test Score Models," British Journal of Mathematical and Statistical Psychology 33, 1980: 234-246.

Hansen, M. H.; Hurwitz, W. N.; and Madow, W. G. Sample Survey Methods and Theory (2 vols.). New York: John Wiley and Sons, 1953.

Hansen, R. H. The Current Population Survey - Design and Methodology. Technical Paper No. 40. Washington, D.C.: U.S. Bureau of the Census, 1978.

Hooke, R. "Some Applications of Bipolykays to the Estimation of Variance Components and their Moments," Annals of Mathematical Statistics 27, 1956: 80-98.

Kish, L. Survey Sampling. New York: John Wiley and Sons, 1965.

Lawley, D. N. "On Problems Connected with Item Selection and Test Construction," Proceedings of the Royal Society of Edinburgh 61, 1943: 273-287.

Lord, F. M. Applications of Item Response Theory. New York: Lawrence Erlbaum, 1980.

Lord, F. M., and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Mislevy, R. J. "Estimating Population Distributions from Multiple Matrix Sample of Item Responses." Paper presented at the meetings of the American Educational Research Association, March, 1982.

Mislevy, R. J., and Bock, R. D. BILOG: Item Analysis and Test Scoring with Binary Logistic Response Models. Chicago: International Educational Services, 1982a.

Mislevy, R. J., and Bock, R. D. "Implementation of the EM Algorithm in the Estimation of Item Parameters: The BILOG Computer Program." Paper presented at the 1982 IRT/CAT Invitational Conference at the University of Minnesota, July, 1982.

Mislevy, R. J.; Reiser, M. R.; and Zimowski, M. Scale-Score Reporting of National Assessment Data. Report to the Education Commission of the States under Contract #02-81-20314. Chicago: International Educational Services, 1981.

Reiser, M. R. "A Latent Trait Model for Group Effects." Doctoral Dissertation, University of Chicago, 1980.

Sanathanan, L., and Blumenthal, S. "The Logistic Model and Estimation of Latent Structure," Journal of the American Statistical Association 73, 1978: 794-799.

Searle, S. R. Linear Models. New York: Wiley, 1971.

Thissen, D. MULTILOG: Item Analysis with Mutliple Category Response Models. Chicago: International Educational Services, 1982.

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

Wolter, K. M. "Composite Estimation in Finite Populations," Journal of the American Statistical Association 74, 1979: 604-613.

Wood, R. L.; Wi[illegible]sky, M. S.; and Lord, F. M. LOGIST: A Computer Program for Estimati[illegible]inee Ability and Item Characteristic Curve Parameters. Research [illegible]orandum 76-6. Princeton: Educational Testing Service, 1976.