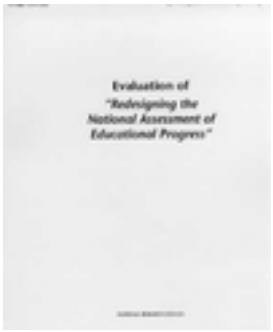

Evaluation of "Redesigning the National Assessment of Educational Progress"

Committee on Evaluation of National and State Assessments of Educational Progress, National Research Council

ISBN: 0-309-56282-1, 36 pages, 8.5 x 11, (1996)

Evaluation of "Redesigning the National Assessment of Educational Progress"

Committee on Evaluation of National and State Assessments of Educational Progress
Board on Testing and Assessment
Commission on Behavioral and Social Sciences and Education
National Research Council

National Academy Press
Washington, D.C. 1996

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competencies and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. William A. Wulf is interim president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. William A. Wulf are chairman and interim vice chairman, respectively, of the National Research Council.

The work of the Committee on the Evaluation of National and State Assessments of Educational Progress is supported by the U.S. Department of Education under Grant No. EA95083001.

ISBN 0-309-05587-3
Additional copies of this report are available from:
National Academy Press
2101 Constitution Avenue, N.W.
Lock Box 285
Washington, DC 20055
Call 1-800-624-6242 or 202-334-3313 (in the Washington metropolitan area).
This report is also available on-line at **http://www.nap.edu/nap/online**.

Printed in the United States of America

COMMITTEE ON THE EVALUATION OF NATIONAL AND STATE ASSESSMENTS OF EDUCATIONAL PROGRESS

JAMES W. PELLEGRINO *(Chair),* Peabody College of Education and Human Development and Learning Technology Center, Vanderbilt University
GAIL BAXTER, School of Education, University of Michigan, Ann Arbor
NORMAN M. BRADBURN, National Opinion Research Center, University of Chicago
ALLAN COLLINS, Bolt Beranek and Newman Inc., Cambridge, Massachusetts
STEPHEN B. DUNBAR, College of Education, University of Iowa
THOMAS H. FISHER, Bureau of Curriculum, Instruction, and Assessment, Florida Department of Education, Tallahassee
LARRY V. HEDGES, Department of Education, University of Chicago
SHARON JOHNSON-LEWIS, Department of Research, Development and Coordination, Detroit Public Schools, Michigan
RODERICK J.A. LITTLE, Department of Biostatistics, University of Michigan
ELSIE G.J. MOORE, College of Education, Division of Psychology in Education, Arizona State University
NAMBURY S. RAJU, Institute of Psychology, Illinois Institute of Technology, Chicago
MARLENE SCARDAMALIA, Ontario Institute for Studies in Education, Toronto, Canada
GUADALUPE VALDES, School of Education and Department of Spanish and Portuguese, Stanford University
SHEILA W. VALENCIA, Curriculum and Instruction, College of Education, University of Washington, Seattle
LAURESS L. WISE, Human Resources Research Organization, Alexandria, Virginia
MICHAEL J. FEUER, *Director, Board on Testing and Assessment*
KAREN J. MITCHELL, *Senior Program Officer*
HOLLY WELLS, *Administrative Assistant*

BOARD ON TESTING AND ASSESSMENT 1995-1996

Contents

Summary

The National Assessment Governing Board has proposed a redesign of the National Assessment of Educational Progress (NAEP). A central premise of the document is that NAEP must be simplified; over time the assessment has been asked (National Assessment Governing Board, 1996:i; reprinted in appendix):

> . . . to do more and more beyond its central purposes. Additions have been made without changing its basic design, making the National Assessment overly complex and costly.

The Governing Board seeks to initiate a process for streamlining NAEP's design in core subject areas, increasing the usefulness and timeliness of reports, and maintaining the accuracy, reliability, and validity of NAEP data.

This report comments on the May 1996 draft redesign plan, as requested by the U.S. Department of Education. It is part of the congressionally mandated evaluation of NAEP by the National Research Council (NRC). The NRC's Committee on the Evaluation of National and State Assessments of Educational Progress is charged with reviewing NAEP generally and evaluating the developmental state assessments, student performance levels, and the extent to which results are reasonable, valid, and informative to the public.

The committee concludes that the motivation for the redesign of NAEP is sound. During the past 25 years, policy concerns about educational opportunity, human resource needs, and school effectiveness have driven the program in varied and, in some cases, conflicting directions. Without changing NAEP's basic design, structural elements have been added and features have been changed in response to the diverse interests of the growing constituency for assessment in America's schools. We applaud the Governing Board for initiating this important redesign process and are pleased to see that the Commissioner of Education Statistics is hastening to address the issues raised in the redesign proposal.

The committee's chief conclusion is that the proposed redesign is at once too ambitious and not ambitious enough. It is too ambitious in the sense that it tries to be responsive to the interests of all kinds of users as well as to myriad criticisms of NAEP. The very degree of responsiveness militates against the overall goal of simplification and streamlining. At the same time, and more importantly, the current redesign proposal does not go far enough in addressing the root cause of the problems that motivated the redesign. We attribute much of the dilemma surrounding NAEP today as flowing from the multiplicity of purposes for this assessment program that has accrued over the years. What we do not see in the redesign proposal is a clear sense of priorities from which decisions about audience, information needs, measurement design and administration design would flow.

The committee also finds that the proposed changes in the NAEP design, administration, analysis, and reporting schemes are largely unspecified. Not only does the proposal not suggest choices—or at least recognize the need for choices—among many and varied design principles, it does not describe the process by which principles and plans would be implemented. The proposal does not specify mechanisms for deciding among conflicting program elements and ensuring the coherence and integrity of the assessment.

Finally, several aspects of the proposal involve suppositions about the future program for which there is a limited empirical base. Some of the proposal's basic premises are stated without data on feasibility or discussion

of design tradeoffs. Proposed actions such as annual administrations are predicated on cost savings derived from simplification of NAEP's measurement and administration designs. Whether or not a simplified assessment can be realized is as yet unknown. Should real cost savings not materialize, several redesign tenets would either have to be dropped or be saved at the expense of presently unspecified components of the program.

These considerations lead us to recommend that the National Assessment Governing Board and the National Center for Education Statistics, which has responsibility for developing the instruments and carrying out the National Assessment of Educational Progress, view the current redesign process as an interim solution. The implicit correlate is that a fundamental rethinking of NAEP is needed. Hence, care needs to be taken that decisions or choices made at this time do not compromise a more ambitious reconceptualization of the National Assessment of Educational Progress.

Evaluation of "Redesigning the National Assessment of Educational Progress"

BACKGROUND

The origins and evolution of the National Assessment of Educational Progress provide the necessary backdrop to our analysis of NAEP's mission and measurement objectives, its design, and its governance and management structure.

NAEP's Origin and Evolution

In 1963 Francis Keppel, then U.S. Commissioner of Education, appointed a committee to explore options for assessing the condition and progress of American education. The committee's chair, Ralph Tyler (1966:95), described the need for a base of information to help public officials make decisions about education:

> . . . dependable information about the progress of education is essential. . . . Yet we do not have the necessary comprehensive and dependable data; instead, personal views, distorted reports, and journalistic impressions are the sources of public opinion. This situation will be corrected only by a careful, consistent effort to obtain data to provide sound evidence about *the progress of American Education* [italics added].

In 1966 the Keppel committee recommended that a battery of tests be developed to the highest psychometric standards and with the consensus of those who would use it. NAEP was conceived to provide that information base, to monitor the progress of American education (National Center for Education Statistics, 1974; U.S. Congress, 1992)

The design of the original battery reflected the political and social realities of the time (National Assessment Governing Board, no date). Prominent among these was the resistance of state and local policy makers to a national curriculum; local leaders feared federal erosion of their autonomy and voiced concern about pressure for accountability. NAEP's designers responded by defining testing objectives for NAEP that were too expansive to be incorporated in any single curriculum. They specified that results be reported for specific test exercises, not in relation to broad knowledge and skill domains. Tests were developed for and administered to 9-, 13-, and 17-year-olds rather than to individuals at specific grade levels. These features, combined with matrix sampling—which distributed large numbers of items broadly across school buildings, districts, and states, but limited the number of items given to individual examinees—thwarted perceptions of NAEP as a federal testing program addressing a nationally prescribed curriculum. Indeed, NAEP's design provided nationally and regionally representative data on the educational condition of American youth while avoiding any implicit federal standards or state, district, and school comparisons. NAEP was described as the nation's education barometer.

Over the following decade, the educational landscape changed. Schools across the United States developed new programs to respond to various federally sponsored education initiatives. The Elementary and Secondary Act of 1965 established mechanisms through which schools could address the learning needs of economically disadvantaged students. In the ensuing years, federal support expanded to provide additional resources for students with limited English proficiency, for example, and students with disabilities. As federal initiatives expanded educational opportunities at the local level, however, they fostered an administrative imperative for assessment data to help gauge the effect of these opportunities on the nation's education system.

NAEP's original design could not accommodate the increasing demands for data about federal education innovations. Its reporting scheme, for example, allowed for the measurement of change on individual exercises, but not on the broad content domains that were evolving. Furthermore, age-level (rather than grade-level) testing made it difficult to link NAEP results to state and local education policies and school practices. Increasingly, NAEP was asked to provide more detailed information so that government and education officials would have a stronger basis for judgments about school effectiveness; NAEP's constituents were seeking information that, in many respects, conflicted with the basic design of the program.

Redesign of the Original Plan

The first major redesign of NAEP was implemented in 1984, when its development and administration moved from the Education Commission of the States to the Educational Testing Service. The design for NAEP's second generation (Messick et al., 1983), with its changes in sampling, objective-setting, exercise development, data collection, and analysis, reflected the growing federal role in American education. The introduction of balanced incomplete block designs for matrix sampling, model-based approaches to item scaling within content domains and across age and grade cohorts, and statistical adjustments based on collateral information about examinees afforded NAEP much greater flexibility in responding to policy demands as they evolved.

Almost concurrently, *A Nation at Risk* (National Commission on Excellence in Education, 1983) warned that America's schools and its students were performing below expectation. The report's publication spawned a wave of state-level education reform. As states invested more and more in their education systems, they sought information about the effectiveness of their efforts. In the face of rising costs and multiple demands, policy makers looked to NAEP for guidance on the effectiveness of alternative practices. The National Governors' Association then called for state-comparable achievement data, and a new report, *The Nation's Report Card* (Alexander and James, 1987), recommended that NAEP be expanded to provide state-level results. This recommendation was a dramatic departure from the original NAEP model.

Soon thereafter, participants in the 1989 Education Summit in Charlottesville, Virginia, challenged the prevailing assumptions about national expectations for achievement in America's schools. President Bush and the nation's governors established six national goals for education *(America 2000,* 1991). Goal three specified the subjects and grades in which progress should be measured with respect to national and international frames of reference. By design, these subjects and grades paralleled NAEP's structure. The governors called on educators to hold students to "world-class" knowledge and skill standards. The governors' commitment to high academic standards included a call for the articulation of NAEP results by achievement levels and performance standards. In addition to describing what students know and can do, NAEP was being asked for judgments about the adequacy of observed performance. In the governors' terms, NAEP was asked to test not only what students currently know and can do, but also what young people should know and be able to do.

Current Design and Governance

NAEP surveys the achievement of students at ages 9, 13, and 17 and in grades 4, 8, and 12. The current program calls for assessment in geography, reading, writing, mathematics, science, U.S. history, world history, the arts, civics, and other academic subjects. Three subjects are tested at each biennial administration. As many as 26 different nonparallel test booklets are used at each age and grade level. During the 1990s, two subjects will have been tested twice in the main assessment, six subjects once, and two subjects not at all. At each administration, three sets of batteries are given: main NAEP, trend NAEP, and state NAEP. Between 150 and 170 distinct subsamples are drawn for each NAEP administration.

The characteristics of score distributions are estimated with complex statistical methods, such as conditioning and multiple imputation of plausible values, which are based on sophisticated scaling models. Results are reported in terms of scaled scores, percentiles, anchor points with exemplar items, and NAGB achievement levels with exemplar responses. The 1992 NAEP mathematics report included seven volumes and over 1,800 pages. Given this complexity, it is perhaps not surprising that anomalies have arisen in recent assessments (U.S. General Accounting Office, 1992) and that controversy has plagued the development and reporting of results using performance standards (National Academy of Education, 1993; U.S. General Accounting Office, 1993).

NAEP's multiplicity of purpose has resulted not

only in its complicated design, but also in an increasingly complex governance structure. Amendments to the authorizing statute for NAEP in 1988 established the present structure. Under the structure, the Commissioner of Education Statistics, who heads the National Center for Education Statistics (NCES) in the U.S. Department of Education, retains responsibility for NAEP operations and technical quality control. NCES procures test development and administration services from cooperating private companies; currently, they are the Educational Testing Service and WESTAT.

The program is governed by the National Assessment Governing Board (NAGB or Governing Board), appointed by the Secretary of Education but independent of the department. The Governing Board, which is authorized to set policy for NAEP, is designed to be broadly representative of NAEP's varied audiences. It selects the subject areas to be assessed and ensures that content is planned though a national consensus process; the Governing Board currently contracts with the Council of Chief State School Officers for national consensus development. In addition, the Governing Board identifies achievement standards for each subject and grade tested, in conjunction with its contractor, the American College Testing Program; it also develops guidelines for reporting. Previously, many of these functions were carried out by advisers to NCES's cooperative test development agencies. NAGB's authority to oversee NAEP and give direction to NCES and the cooperative agencies parallels that of the Commissioner of Education Statistics to direct and execute the program.

The U.S. Department of Education recently commissioned a review of NAEP's management and methodological procedures. That review concluded that confusion over the management structure of NAEP has complicated the program, slowed operations, and increased assessment costs (KPMG Peat Marwick LLP and Mathtech, Inc., 1996). Tension between NAGB and NCES and consensus-based decision making also were said to contribute to these problems.

PROPOSED REDESIGN

NAEP has chronicled educational performance for over a quarter of a century. It has been an unparalleled source of information about the academic proficiency of U.S. students, providing among the best available trend data on the academic achievement of elementary, middle, and secondary students in core subject areas. In addition, NAEP has distinguished itself in setting an innovative and rigorous agenda for conventional and performance-based testing. Because NAEP has been a leader in American testing, it is imperative that its redesign honor this tradition of excellence.

In its redesign proposal, the Governing Board concludes that "in its current form, the National Assessment provides too little information, too infrequently and too late." Our committee agrees with this conclusion. We believe three problems drive these difficulties: its unattainably broad measurement agenda, a resultingly complicated design, and confusion over management and oversight responsibility. The committee notes that these problems have been described in various commentaries by other professional groups concerned about the redesign of NAEP (e.g., Forgione, 1996; Glaser et al., 1996; Johnson, 1996; KPMG Peat Marwick LLP and Mathtech, Inc., 1996; Porter and Kilgore, 1996).

The guiding principles for the NAEP redesign listed in the May 1996 Governing Board draft proposal state that the new assessment should:

- test annually according to a publicly released schedule,
- provide state-level results in reading, writing, math, and science at grade 4 and grade 8 according to a predictable schedule,
- use performance standards for reporting whether student achievement is "good enough,"
- use international comparisons where feasible,
- help states and others link their assessments with the National Assessment,
- vary the amount of detail in testing and reporting,
- simplify the National Assessment test design,
- keep test frameworks and specifications stable for at least 10 years,
- simplify how student achievement trends are reported,
- emphasize grade-based reporting over age-based reporting,
- make use of innovations in testing and reporting, and
- use an appropriate mix of multiple choice and performance test questions.

(See the appendix for the full draft of the NAGB proposal; a slightly modified version was adopted by NAGB on August 2, 1996.)

EVALUATION OF THE PROPOSED REDESIGN

We commend the National Assessment Governing Board for reviewing and seeking to improve the current program. The committee supports a number of elements in the Governing Board's redesign proposal. We agree with the desire to accelerate the reporting schedule after testing and to provide more comprehensible results to policy makers and the public. We also agree that the availability of public and predictable schedules for the main, trend, and state assessments is important for planning in numerous policy arenas. We applaud the intention to strengthen the high school data collections. And, like NAGB, we see merit in exploiting new technologies for NAEP to increase the efficiency and accuracy of the assessment.

The program described by the Governing Board's redesign proposal is laudable in many respects. Overall, however, the document is an amalgam of disparate needs and elements, and it places an inordinate faith in the undefined concept of simplification. Although it recognizes that NAEP has been asked to do "more and more beyond its central purposes," it refrains from serious discussion of the hard political and technical choices that are needed. Our concern is less with any specific element than with the assemblage of elements, less with any given goal than with the lack of clear priorities and the lack of detail about how the goals might be achieved. This finding is the basis for the committee's recommendation that redesign measures undertaken now be considered interim solutions—steps along the way to a fundamental rethinking of NAEP.

Multiple and Varied Purposes

In a minority statement to the Alexander and James report (1987), Linda Darling-Hammond presaged our reaction to the 1996 redesign proposal:

> The effort to make NAEP data useful for a greater range of purposes will undermine the assessment's capacity to perform its basic mission effectively. There is a delicate balance between developing a first-rate assessment of what the nation's students know and can do and attempting to negotiate a multi-purpose testing and data collection effort that may satisfy many objectives superficially but none of them well (p. 31).

It is something of a truism to say that testing has become the victim of its own success. Federal, state, and local policy makers, education administrators, curriculum specialists, educators, researchers, the business community, media, parents, students, and the general public all have legitimate interests in the status of U.S. education. Student achievement data have become the indicator of choice—to gauge the impact of federal and state investment in education, to make judgments about teacher effectiveness or school quality, to review the effectiveness of programs, to evaluate educational innovations, as accountability measures, for classroom feedback, for individual credentialing, and for international comparisons. Congress, the Department of Education, the National Center for Education Statistics, and the National Assessment Governing Board have all succumbed to the growing desire for more and more information about student achievement, and Darling-Hammond's cautionary advice notwithstanding, they are asking NAEP to provide it all.

The committee concludes that this underlying desire cannot be met by NAEP: the universe of possible interests cannot be served simultaneously and well by the same assessment.

In the most recent reauthorization of the National Assessment (Improving America's Schools Act 1994, P.L. 103-382), Congress mandated that it should:

> . . . provide a fair and accurate presentation of educational achievement in reading, writing, and other subjects included in the third National Education Goal, regarding student achievement and citizenship.

To implement this charge, the Governing Board adopted three objectives for NAEP:

- to measure national and state progress toward the third National Education Goal and provide timely, fair and accurate data about student achievement at the national level, among states, and in comparison with other nations;
- to develop through a national consensus, sound assessments to measure what students know and can do as well as what they should know and be able to do; and
- to help states and others link their assessments to National Assessment and use National Assessment data to improve education performance.

This agenda has been constructed over the last 8 years. It is the crux of the problem. While each of these objectives is in itself a worthy goal, collectively they have produced a testing program that everyone admits is overburdened and excessively complex. The failure to adopt a workable set of priorities—plus the addition of ambitious plans for annual administrations and additional subjects—suggests that the redesign has

the potential to continue and perhaps exacerbate the problems it is seeking to solve.

The tension between assessment for national and state education goals epitomizes the committee's concern about the many and diffuse purposes of the national assessment. Without question, there is great public interest in the progress of states toward self-determined education goals and standards, and this interest has, over the years, moved NAEP in a direction that better serves the states in their pursuit of information to evaluate progress. At the same time, the addition of the trial state assessments to NAEP necessitated significant accommodations in the design, scheduling, and reporting of assessments and their results. It is the relative cost and benefit of accommodations of this sort that the committee believes must receive more careful scrutiny in the redesign of NAEP.

For example, the sampling framework to support inferences about state-level performance differs from that required for inferences about the nation as a whole, resulting in separate samples being drawn for national and state NAEP. A natural response to this apparent duplication is discussed in the May draft; it proposed developing new sampling methods so that both kinds of inferences can be supported by a single sample. Several alternatives for such combined sampling have been reviewed in the interim by the Design and Feasibility Team commissioned by NAGB. Its evaluation of alternatives accentuates technical difficulties that would arise in the areas of equating, content sampling, and participation for national versus state NAEP (Forsyth et al., 1996). In the final policy statement, NAGB steps away from combined sampling as a viable solution to the problem of trying to reconcile two very different, demanding objectives. This does not resolve the dilemma, however, of trying to serve both national and state needs adequately.

Other conflicts between needs for national and state assessments are less easy to anticipate, but important to consider. To the extent that NAEP moves toward greater focus on the states, one might expect increased interest in the degree to which the curriculum frameworks developed at the national level accurately characterize state curricula, education goals, and standards. Although the procedures in place for developing curriculum frameworks yield broadly representative specifications for test content, they are not designed for alignment with particular curricula and standards. Whether states will want them to become so aligned is a matter of state education policy, but that alignment is a critical component of the validity of inferences based on NAEP.

The juxtaposition of competing values for national and state assessment focuses attention on the need for informed discussion of the central purposes for national assessment. It also makes manifest the complexity of further extensions of NAEP to achieve, for example, valid international benchmarks for performance or of including in the assessment some populations (e.g., students with various disabilities) that may require selection of exercises and other alterations to accommodate their special situations.

Insufficient Detail

The very general nature of the Governing Board's redesign document makes it difficult to evaluate its feasibility. The redesign proposal lacks specificity and detail with respect to the new assessment's design, administration, analysis, and reporting schemes. Fruitful debate about the redesign objectives will require more information about the feasibility and likely psychometric characteristics of the assessment envisioned. The proposal does not specify how the redesign objectives will be achieved or at what cost —in terms of validity, reliability, and timeliness, as well as, funding

To give but one example, detail is lacking on the means by which trend data would be collected through the main assessment. The redesign proposal states that it may be impractical and unnecessary to operate two separate testing programs and that a plan should be developed to allow the main assessment to become the primary way to measure trends in reading, writing, mathematics, and science. It does not explore in any depth options for combining the main and trend data collections, nor does it describe a process for analyzing and deciding among alternatives. It does, however, state the Governing Board's intention to do away with the trend assessment after ". . . a carefully planned transition . . ." (p. 7).

Collapsing the main and trend data collections would be very difficult for a number of reasons. The most obvious obstacle is that the content frameworks for the two assessments are different. The likelihood of even minor changes in frameworks and items raises questions about the validity of trend lines that would be based on the main assessment. The proposal to combine the trend and main assessments would thus jeopardize NAEP's continuity over time and thereby undermine what is, in the committee's view, one of

NAEP's unique and most valuable features. Once broken, the chain of evidence is irretrievable.

The Committee recommends that a separate collection of NAEP trend data be continued. This recommendation comports with our recommendation that the contemplated redesign be viewed as a set of limited, interim solutions, including no elements that could compromise a more ambitious reconceptualization in the future.

Both the Department of Education and the National Assessment Governing Board will be aided in thinking about how the redesign proposal might be operationalized by the recent report of the Design and Feasibility Team (Forsyth et al., 1996). This group of leading measurement experts was asked by the Governing Board to suggest operational alternatives for the redesign objectives. It is important to note that the focus of the study was not on the feasibility and impact of specific changes—or more importantly, the constellation of changes—considered by the Design and Feasibility Team. These are as yet largely unknown. The research undertaken by NAGB and NCES in coming months will need to address these questions.

Insufficient Empirical Base

NAEP's managers seek to simplify the program in ways that would release funds for more frequent assessment, additional subject testing, accelerated reporting, and other enhancements. However, streamlining NAEP's measurement and administration designs in accordance with the proposal will be exceedingly difficult, both conceptually and technically. A number of issues work against parsimony. First, the policy framework for NAEP is unclear; parsimony rests, at least in part, on clarity of purpose. Second, NAEP's measurement and design properties are very complex, so that many of NAEP's findings derive from sequential calculations. Third, the expansion of NAEP's client base to state testing offices intensifies, rather than simplifies the burdens of NAEP for sampling, administration, analysis, and reporting.

Answers to some of these outstanding issues may come during the next 6 months when several commissioned papers on critical features of the present NAEP will be issued. However, the current schedule for letting contracts to carry out elements of the plan may preclude the possibility for the redesign to be influenced by three such studies: *Quality and Utility: The 1994 Trial State Assessment in Reading (1996) and Capstone Report of the National Academy of Education Panel on the Evaluation of NAEP* (in press), both coming from the National Academy of Education, and NAEP Validation Studies white papers (National Center for Education Statistics, no date).

Another important document, the KPMG Peat Marwick LLP and Mathtech, Inc. (1996) *Review of the National Assessment of Educational Progress: Management and Methodological Procedures,* was only recently released: it addresses feasibility issues with direct bearing on the redesign proposal, but was not available when NAGB's proposal was being written. Among its important conclusions was the fact that several of NAGB's widely discussed revisions are unlikely to provide cost savings. Given the importance of assumed cost reductions through simplification to the expansive elements in the redesign proposal (e.g., more frequent testing), this conclusion is troubling. Decisions about simplification should represent an informed balance of estimated costs and proposed benefits, and they should be based on understanding of the implications for the technical integrity of the test instruments in light of the purposes currently espoused.

The committee recommends that the U.S. Department of Education evaluate the cost implications of specific aspects of the redesign proposal.

Despite the impressive work of the Design and Feasibility Team, KPMG Peat Marwick and Mathtech analysts, and others, several basic premises of the proposed redesign are as yet unsupported by empirical data. While the committee recognizes that NAEP cannot lie dormant while all relevant research is conducted, the feasibility of major changes should be explored prior to their approval. Such specific issues as sample size, test content, length, item format mix, and scoring procedures merit empirical investigation. Provisional estimates of the effects of various proposed design changes can be obtained by analyses of existing NAEP data. Among the questions that would reward empirical analysis are the following:

- How would recommended changes affect the reliability and validity of cross-sectional data?
- How would changes in the numbers and types of items affect the reliability and validity of trend information?
- Would combining trend and cross-sectional assessments compromise the measurement of long-term trends?
- Are results for the developmental achievement levels sufficiently accurate and informative that they should be operationally adopted?

In addition, there are a number of other questions for which research has long been needed. Alternatives for enhancing student motivation on a low-stakes exam warrant study, as do strategies for increasing the value of NAEP information to various users. As accountability and assessment become more integrally linked in state testing programs, questions about possible changes in test preparation and test performance also become important. Extending assessment to students with special learning needs and those with limited English proficiency presents challenges in test development, analyses, and reporting. Answers to these questions will be critical to a useful redesign of NAEP.

The committee understands that NAGB intends to move ahead with strategies for revision of NAEP while planning to contract for research on the questions for which empirical data are needed. This approach raises concern that fundamental policy issues about national and state assessment would be decided in negotiations with a variety of contractors. Although such an approach might yield creative solutions, it also affords little protection against conflicts of interest and does not allow for a broad view and high-level policy attention to the cohesiveness of the program and the integrity of NAEP.

FUNDAMENTAL RECONCEPTUALIZATION OF NAEP

We recommend that the National Assessment Governing Board and the U.S. Department of Education consider the NAGB redesign proposal as a range of possible interim measures to alleviate some of the immediate pressures on NAEP while undertaking a more fundamental rethinking of NAEP's goals and character.

The advantages of thinking in terms of interim solutions are several. First, this approach defines some bounds for the redesign in terms of time, expenditures, and expectations. A deliberate decision to work on interim solutions would suggest avoiding changes that could inadvertently damage or constrain future options when a comprehensive redesign is undertaken. Such goals as shorter reporting times, a known, regular schedule of administrations, and simpler, more comprehensible reports would seem to satisfy this criterion. Second, viewing the redesign as an interim measure helps clarify what is workable at this stage. The redesign proposal does not make any major adjustments to the avowed purposes of NAEP; as a consequence, any simplifications in test design, sampling, or administration should be limited to those that will not compromise the quality of the data that federal and state policy makers assume is present. And finally, accepting such limited goals provides time in which to engage in a fundamental rethinking of NAEP and its purposes.

We urge a modest approach in the current design, but we are also strongly convinced of the need for the National Assessment Governing Board, the National Center for Education Statistics, and the Congress to embark on a process of rethinking the National Assessment of Educational Progress from the ground up. Paramount among issues that require close examination are the assessment's purpose and measurement objectives. NAEP has been called on to inform policy debates about the academic achievement of U.S. students, equality of educational opportunity, human resource issues, school effectiveness, and the attainment of forward-looking performance standards. There is general agreement that this is a mandate no single testing program can fulfill. Officials in Congress, the Executive Branch, the states, and members of the Governing Board need to make choices. These political choices need to be informed by a careful weighing of technical options and information needs.

For example, for policy makers to decide if they should make it a top priority for NAEP to be focused on state testing, questions about linkage—to state frameworks, state assessments (both conventional and performance based), and state reporting requirements—would be pressing. The means by which links could be made are unknown. More extensive research on linking or comparability would be needed, in combination with deliberation about the types and levels of service NAEP could support under various funding assumptions. Moreover, problems associated with participation rates and desired inferences to smaller sampling units would need to be specifically addressed.

The use of technology to streamline certain aspects of the national assessment should be considered in expanding the redesign initiative. For example, NAEP has successfully implemented new scanning technologies to create image databases that increase the efficiency of scoring open-ended exercises. Other such uses of technology to streamline operations have obvious appeal with regard to costs and timely reporting. Less obvious, however, is the gain to be achieved by computerized test delivery and adaptive testing, the former because of the comparative low cost of paper-and-pencil tests and the latter because adaptive methods are not well suited for summarizing aggregate performance across domains. Just how technology can be

used to cut costs and improve measurement in NAEP requires careful evaluation and planning.

Decisions about the long-term future of NAEP also need to take account of likely new directions in testing and assessment. For example, understanding of knowledge structures and the way individuals acquire and represent knowledge is very different from what it was when NAEP and other large-scale testing programs began (Glaser et al., 1992; Gifford and O'Connor, 1992; Wittrock and Baker, 1991). Scientific information about learning and cognition is only beginning to be applied to test design in the United States. But there is every reason to think that the burgeoning sciences of learning and cognition will have a significant impact on education and, therefore, on assessment as well. In keeping with its role as a leader in assessment, NAEP in the twenty-first century should grow out of the science of learning.

CONCLUSION

A common perception of those who have watched NAEP over the years is of an ever-expanding black box with contents that are thoroughly understood by an ever-shrinking number of specialists. The Governing Board's proposal adjusts the dimensions of the box, shakes its contents, and smoothes some of its rougher edges. However, it stops short of examining what belongs inside. The committee believes that NAGB's proposal to redesign NAEP should represent an initial step in reconsideration of the fundamental purposes and practices of the national assessment.

NAEP's prominence in American education, its allure as a palliative in addressing education's ills, and the diverse interests of legitimate stakeholders all operate to make its mission diffuse. NAEP's complex design and cumbersome management and governance structure mirrors its many purposes. The committee believes that NAEP's reconceptualization should directly address the fundamental tensions among measurement purposes.

References

Alexander, L. , and H. James . 1987 *The Nation's Report Card.* Washington, D.C. : National Academy of Education .

1991 *America 2000: Excellence in Education Act.*

Forgione, Jr., P.D. 1996 Memorandum to Mark Musick on NCES's Review of NAGB Redesign Paper . July 1, 1996 .

Forsyth, R. , R. Hambleton , R. Linn , R. Mislevy , and W. Yen . 1996 *Design/Feasibility Team Report to the National Assessment Governing Board.* Washington, D.C. : National Assessment Governing Board .

Glaser, R. , R. Linn , and G. Bohrnstedt . 1996 Letter to Roy Truby on the future of NAEP . February 23, 1996 ,

Glaser, R. , K. Raghavan , and G. Baxter . 1992 *Cognitive Theory as the Basis for Design of Innovative Assessment: Design Characteristics of Science Assessments.* CSE Technical Report No. 349 . Los Angeles : University of California, National Center for Research on Evaluation, Standards, and Student Testing .

Gifford, B. , and M. O'Connor , eds. 1992 *Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction.* Boston, Mass. : Kluwer Academic Publishers .

Johnson, S.T. 1996 Letter to Ray Fields on behalf of the Design and Analyses Committee . June 19, 1996 .

KPMG Peat Marwick LLP and Mathtech, Inc. 1996 *A Review of the National Assessment of Educational Progress: Management and Methodological Procedures.* Washington, D.C. : U.S. Department of Education .

Messick, S. , A. Beaton , and F. Lord . 1983 *National Assessment of Educational Progress Reconsidered: A New Design for a New Era.* Princeton, N.J. : Educational Testing Service .

National Academy of Education 1993 *Setting Performance Standards for Student Assessment.* Stanford, Calif. : Stanford University, National Academy of Education .

1996 *Quality and Utility: The 1994 Trial State Assessment in Reading.* Stanford, Calif. : Stanford University, National Academy of Education .

in press *Capstone Report of the National Academy of Education Panel on the Evaluation of NAEP.* Stanford, Calif. : Stanford University, National Academy of Education .

National Assessment Governing Board 1996 *Redesigning the National Assessment of Educational Progress: Draft for Public Comment* (May 1996) . Washington, D.C. : National Assessment Governing Board .

no date *The National Assessment of Educational Progress: Demands, Realities and Assumptions.* Work Group on Planning . Washington, D.C. : National Assessment Governing Board .

National Center for Education Statistics 1974 *NAEP General Information Yearbook.* Washington, D.C. : U.S. Department of Education .

no date NAEP Validation Studies for 1996 . Unpublished paper . U.S. Department of Education .

National Commission on Excellence in Education 1983 *A Nation at Risk: The Imperative for Educational Reform.* Washington, D.C. : U.S. Government Printing Office .

Porter, A. , and S. Kilgore . 1996 Letters to Mr. William Randall on behalf of the Advisory Council on Education Statistics . April 23, 1996 ; June 27, 1996 .

Tyler, R. 1966 The development of instruments for assessing educational progress . Pp. 95-101 in *Proceedings of the 1965 Invitational Conference on Testing Problems.* Princeton, N.J. : Educational Testing Service .

U.S. Congress, Office of Technology Assessment 1992 *Testing in American Schools: Asking the Right Questions.* Washington D.C. : U.S. Government Printing Office .

U.S. General Accounting Office 1992 *National Assessment Technical Quality.* GAO/ PEMD-92-22R . Washington, D.C. : U.S. General Accounting Office .

1993 *Educational Achievement Standards: NAGB's Approach Yields Misleading Interpretations.* GAO/ PEMD-93-12, June 1993 . Washington, D.C. : U.S. General Accounting Office .

Wittrock, M. , and E. Baker , eds. 1991 *Testing and Cognition.* Englewood Cliffs, N.J. : Prentice Hall .

Appendix: Redesigning the National Assessment of Educational Progress: Draft For Public Comment

A slightly modified version was adopted on August 2, 1996.

National Assessment Governing Board
National Assessment of Educational Progress

Redesigning The National Assessment of Educational Progress
Draft for Public Comment

The National Assessment Governing Board seeks your comments on this proposed policy to change the National Assessment of Educational Progress. The National Assessment provides the American public with information about student achievement nationally and state-by-state. The proposed policy describes changes that will make the National Assessment a more effective monitor of student achievement and make it more useful to the public.

Written comments should be submitted to Ray Fields, National Assessment Governing Board, Suite 825, 800 North Capitol Street, NW, Washington, DC 20002 for receipt by June 28, 1996. Your comments will be used to help refine the proposed policy before final action by the Nation Assessment Governing Board on August 3, 1996.

Prepared by The Work Group on Planning
Mark Musick, Chair
Marilyn McConachie
Jason Millman
Richard Mills
William Moloney
Michael Nettles
William Randall
Staff: Daniel Taylor, Ray Fields

800 North Capitol Street, N. W.
Suite 825, Mailstop 7583
Washington, D.C. 20002-4233
Phone: (202) 357-6938
Fax: (202) 351-6945

Redesigning the National Assessment of Educational Progress

OVERVIEW

The National Assessment is the only means for the American public to know with accuracy how its students are achieving nationally and state-by-state. However, in its current form, the National Assessment provides too little information, too infrequently and too late. Over the years, the National Assessment has been asked to do more and more beyond its central purpose. Additions have been made without changing its basic design, making the National Assessment overly complex and costly.

The National Assessment must be changed in order to provide the public the information it needs about student achievement within the funding available. The National Assessment must be simplified. Its funding should be focused on its central purpose: reporting on student achievement in ten required subjects. Useful but less essential activities should be cut back or carried out by others.

The audience for the National Assessment is the American public. Reports should be timely, easy to use, understandable, and widely available. Results should describe both changes over time and whether student achievement on the National Assessment is "good enough."

While change is in order, many current policies should not change. For example, reliability, validity, and accuracy of the data will remain hallmarks of the National Assessment. Students who are tested will be as representative as possible of the students in that grade; exclusions because of disability or limited English proficiency will be kept to a minimum.

The proposals recommended below will make the National Assessment more useful and allow it to report on more subjects, more frequently, and more quickly. The recommendations include the following:

- test annually according to a publicly released schedule
- provide state-level results in reading, writing, math and science at grade 4 and grade 8, according to a predictable schedule
- use performance standards for reporting whether student achievement is "good enough"
- use international comparisons where feasible
- help states and others link their assessments with the National Assessment
- vary the amount of detail in testing and reporting
- simplify the National Assessment test design
- keep test frameworks and specifications stable for at least ten years
- simplify how student achievement trends are reported
- emphasize grade-based reporting over age-based reporting
- make use of innovations in testing and reporting
- use an appropriate mix of multiple choice and performance test questions

Redesigning the National Assessment of Educational Progress

A Better Way to Measure Educational Progress in America

An effective democracy and a strong economy require well-educated citizens. A good education lays a foundation for getting a good job, leading a fulfilling life, and participating constructively in society.

But is the education provided in your state and in America good enough? How do our 12th graders compare with students in other nations in mathematics and science? Do our 8th grade students have an adequate understanding of the workings of our constitutional democracy? How well do our 4th grade students read, write, and compute? The National Assessment of Educational Progress is the only way for the public to know with accuracy how American students are achieving nationally and state-by-state.

The National Assessment tests at grades 4, 8 and 12. By law, it covers ten subjects, including reading, writing, math and science. The National Assessment has performance standards that indicate whether student achievement is "good enough." The National Assessment is not a national exam taken by all students. In fact, only several thousand students are tested per grade, comprising carefully drawn samples that represent the nation and the participating states. Since its first test in 1969, the National Assessment has earned a trusted reputation for its quality and credibility. That reputation must be maintained.

The National Assessment is unique because of its national, state-by-state, and 12th grade results. State and local test results cannot be used to provide a national picture of student achievement. States and local schools use different tests that vary in many ways. The results cannot simply be "added up" to get a national score nor can state scores on their different tests be compared. Virtually no state tests 12th graders, so the only source of information about 12th grade achievement is the National Assessment. College entrance tests such as the ACT and the SAT are taken only by students planning on higher education; the results do not represent the achievement of the total 12th grade class. Twelfth grade achievement is important to monitor because it marks the end of elementary and secondary education, the transition point for most students from school to work, to college, or to technical training.

While there is much about the National Assessment that is working well, there is a problem. Under its current design, the National Assessment tests too few subjects, too infrequently, and reports achievement results too late--as much as 18 to 24 months after testing. Testing occurs every other year. During the 1990's, only reading and mathematics will be tested more than once using up-to-date tests and performance standards. Six subjects will be tested only once and two subjects not at all during the 1990's.

Why is the National Assessment testing so few subjects and fewer subjects now than years ago? Over the years, the National Assessment has become increasingly complex. Its quality and integrity have led to a multitude of demands and expectations beyond its central purpose. Meeting those expectations was done with good intentions and seemed right for the situation at the time. However, additions to the National Assessment have been "tacked on" without changing the basic design, reducing the number of subjects that can be tested and driving up costs.

For example, where a single 120 page mathematics report once sufficed, mathematics reporting in 1992 consisted of seven volumes totalling almost 1,800 pages, not including individual state reports. Also, there are now two separate testing programs for reading, writing, math and science. One monitors trends using tests developed during the 1970's; the other reflects current views on instruction and uses performance standards to report whether achievement is good enough. In addition, there are separate samples for reporting national and state results, even when the state samples may be adequate for some national reports.

The current National Assessment design is overburdened, inefficient and redundant. It is unable to provide the frequent, timely reports on student achievement the American public needs. The challenge is to supply more information, more quickly, with the funding available.

To meet this challenge, the National Assessment design must be changed, building on its strengths while making it more efficient. The design of the National Assessment must be simplified. The purpose of the National Assessment must be sharply focused and its principal audience clearly defined. Because the National Assessment cannot do all that some would have it do, trade-offs must be made among desirable activities. Useful but less important activities may have to be reduced, eliminated, or carried out by others. The National Assessment must "stick to its knitting" in order to be more cost-effective, reach more of the public, provide more information more promptly, and maintain its integrity.

[On the pages that follow are preliminary proposals for new policies for the National Assessment being offered for public comment by the National Assessment Governing Board. The intent of these proposals is to specify purposes, audiences, and changes that will make the National Assessment a more effective monitor of student achievement.]

Purpose of the National Assessment of Educational Progress

The purpose of the National Assessment is stated in its legislation:

> to provide a fair and accurate presentation of educational achievement in reading, writing, and the other subjects included in the third National Education Goal, regarding student achievement and citizenship.

Thus, the central concern of the National Assessment is to inform the nation on the status of student achievement. The National Assessment Governing Board believes that this should be accomplished through the following objectives:

(1) to measure national and state progress toward the third National Education Goal and provide timely, fair and accurate data about student achievement at the national level, among the states, and in comparison with other nations;
(2) to develop, through a national consensus, sound assessments to measure what students know and can do as well what students should know and be able to do; and
(3) to help states and others link their assessments with the National Assessment and use National Assessment data to improve education performance.

The Audience for the National Assessment

The primary audience for National Assessment results is the American public, including the general public in states that receive their own results from the National Assessment. Reports should be written for this audience. Results should be released within 6 months of testing. Reports should be understandable, jargon free, easy to use, and widely disseminated.

Principal users of National Assessment data are state policymakers and educators concerned with student achievement, curricula, testing and standards. National Assessment data should be available to these users in forms that support their efforts to interpret results to the public and to improve education performance.

What the National Assessment Is Not

The National Assessment is intended to describe how well students are performing, but not to explain why. The National Assessment only provides group results; it is not an individual student test. The National Assessment tests academic subjects and does not collect information on individual students' personal values or attitudes. Each National Assessment test is developed through a national consensus process. This national consensus process takes

into account education practices, the results of education research, and changes in the curricula. However, the National Assessment is independent of any particular curriculum and does not promote specific ideas, ideologies, or teaching techniques. Nor is the National Assessment an appropriate means, by itself, for improving instruction in individual classrooms, evaluating the effects of specific teaching practices, or determining whether particular approaches to curricula are working.

Recommended Changes to the National Assessment

To provide the American public with more frequent information in more subjects about the progress of student achievement, changes must be made in the way that the National Assessment is designed and the results are reported. Many current policies should continue. Reliability, validity, and quality of data will remain a hallmark of the National Assessment. The sample of tested students will be as representative as possible, keeping to a minimum the number of students excluded because of disability or limited English proficiency. Tests and test frameworks will be kept stable to measure progress in student achievement over time.

The recommended changes relate to the three objectives outlined above. Current contracts for conducting the National Assessment extend through 1998. Changes can be incorporated in assessments in the year 1999 and thereafter. Where feasible, these recommendations should be used to guide decisions under, current contracts.

OBJECTIVE 1: To measure national and state progress toward the third National Education Goal and provide timely, fair and accurate data about student achievement at the national level, among the states, and in comparison with other nations.

Test all subjects specified by Congress: reading, writing, mathematics, science, history, geography, civics, the arts, foreign language, and economics

The gap must be closed between the number of subjects the National Assessment is required to test and the number of subjects it can test under the current design. By law, the National Assessment is required to test ten subjects and report results and trends. In order to chart progress and report trends, subjects must be tested more than once. However, during the 1990's only reading and mathematics will have been tested more than once using up-to-date tests and performance standards to report how well students are doing.

Recommendations:

- the National Assessment should be conducted annually;

- reading, writing, mathematics and science should be given priority, with testing in these subjects conducted according to a publicly released 10-year schedule adopted by the National Assessment Governing Board;
- history, geography, the arts, civics, foreign language, and economics also should be tested on a reliable basis according to a publicly released schedule adopted by the National Assessment Governing Board.

Vary the amount of detail in testing and in reporting results

More subjects can be tested if different strategies are used. But each time the National Assessment is conducted, it uses a similar approach, regardless of the nature of the subject or the number of times a subject has been tested. This approach is locked-in through 1998 under current contracts. Under this approach, a larger number of students is tested in order to provide not just overall results, but fine-grained details as well (e.g. the achievement scores of 4th grade students whose teachers that year had five hours or more of in-service training). The National Assessment also collects "background" information through questionnaires completed by students, teachers, and principals. The questionnaires ask about teaching practices, school policies, and television watching, to name a few. Data analyses are elaborate. Reports are detailed and exhaustive, involving as many as seven separate reports per subject. Although the National Assessment has been praised for this thoroughness, it comes at the cost of testing more subjects, more frequently, with more timely reporting.

The different strategies needed might include several approaches to testing and reporting. For example, these approaches could take the form of "standard report cards," "comprehensive reports," and special, focused assessments. A standard report card would provide overall results in a subject with performance standards and average scores. Results for standard report cards would be reported by sex, race/ethnicity, socio-economic status, and for public and private schools, but would not be broken down further. This may reduce the number of students needed for testing and may reduce associated costs. Student, teacher and principal survey questionnaires, if collected at all, would be limited and selective, with reports of results focused on only the most essential issues. Generally, subcategories within a subject (e.g. algebra, measurement and geometry within mathematics) would not be reported. However, data from the National Assessment would continue to be available to state and local educators and policymakers for additional analysis. Most National Assessment reports would use this strategy.

Comprehensive reports, like the current approach, would be an in-depth look at a subject, perhaps using a newly adopted test framework, many students, many test questions, and ample background information. In addition to overall results using performance standards and average scores, subcategories within a subject could be reported. Results would be reported by sex, race/ethnicity, socio-economic status, and for public and private schools, and might be broken down further as well. In some cases, more than one report may be issued

in a subject. However, comprehensive reporting would occur infrequently, perhaps once in ten years in any one subject.

Special, focused assessments in a subject would be scheduled as needed. They would explore a particular question or issue and may be limited to particular grades. Generally, the cost would be less than the cost of a standard report card. Examples of these smaller-scale, focused assessments include: (1) assessing subjects using targeted approaches (e.g. 8th grade arts), (2) testing special populations (e.g. in-school 12th graders vs. out-of-school youth), and (3) examining skills and knowledge across several subjects (e.g. readiness for work).

Recommendations:

- National Assessment testing and reporting should vary, using standard report cards most frequently, comprehensive reporting in selected subjects about once every ten years, and special, focused assessments as needed;
- National Assessment results should be timely, with the goal being to release results within 6 months of the completion of testing.

Simplify the National Assessment design

The current design of the National Assessment is very complex. No student takes the complete set of test questions in a subject and as many as twenty-six different test booklets are used within each grade. Students, teachers, and principals complete separate questionnaires and may submit them for scoring at different times. Scores are not calculated directly from the test booklets, but are estimated using statistical procedures known as "conditioning," "drawing plausible values," and "imputation." The estimates are calculated in part by using the questionnaire data collected from the students, teachers, and principals, in addition to the student answers to the test questions. Although using these procedures helps make the data accurate, it also increases the possibility of mistakes. Under these procedures, each time a problem arises in analyzing the data, everything must be redone. It is not unusual for data to be re-calculated hundreds of times. The current complex design of the National Assessment lengthens the time from testing to reporting and adds significantly to its cost.

Recommendation:

- options should be identified to simplify the design of the National Assessment and reduce reliance on conditioning, plausible values, and imputation to estimate group scores.

Simplify the way the National Assessment reports trends in student achievement

From its beginning in 1969, monitoring achievement trends has been a central mission of the National Assessment of Educational Progress. Since 1990, the National Assessment has reported achievement trends using two unconnected testing programs. The tests, criteria for selecting students, and reporting are all different. The first program, "the main National Assessment," tests at grades 4, 8 and 12 and covers ten subjects. The tests are based on a national consensus representing current views of each subject. Performance standards are used to report whether student achievement on the National Assessment is "good enough." The schedule of subjects to be tested in the main National Assessment is unrelated to the schedule of subjects tested under the second testing program.

The second testing program reports long-term trends that go as far back as 1970. Only four subjects are covered: reading, writing, mathematics and science. The tests are based on views of the curricula prevalent during the 1970's and have not been changed. Testing is at ages 9, 13 and 17 except for writing, which tests at grades 4, 8 and 11. Trends are reported by average score; performance standards are not used. The long-term trend program has been valuable for documenting declines and increases in student achievement over time and a decrease in the achievement gap between minority and non-minority students.

It may be impractical and unnecessary to operate two separate testing programs. However, it also is likely that curricula will continue to change and that current test frameworks may be less relevant in the future. The tension between the need for stable measures of student achievement and changing curricula must be addressed carefully.

Recommendations:

- a carefully planned transition should be developed to enable "the main National Assessment" to become the primary way to measure trends in reading, writing, mathematics and science in the National Assessment program;
- as a part of the transition, the National Assessment Governing Board will review the tests now used to monitor long-term trends in reading, writing, mathematics and science to determine how they might be used now that new tests and performance standards have been developed during the 1990's for "the main National Assessment." The Governing Board will decide how to continue the present long-term trend assessments, how often they would be used, and how the results would be reported.

Use performance standards to report whether student achievement is "good enough"

In reporting on "educational progress," the National Assessment has, until recently, only considered current student performance compared to student achievement in previous years. Under this approach, the only standard was how well students had done previously, not how

well they should be doing on what is measured by the National Assessment. Although this approach has been useful, it began to change in 1988 from a sole focus on "where we have been" to include "where we want to be" as well.

In 1988, Congress created a non-partisan citizen's group--the National Assessment Governing Board--and authorized it to set explicit performance standards, called achievement levels, for reporting National Assessment results.

The achievement levels describe "how good is good enough" on the various tests that make up the National Assessment. Previously, it might have been reported that the average math score of 4th graders went up (or down) four points on a five-hundred-point scale. There was no way of knowing whether the previous score represented strong or weak performance and whether the amount of change should give cause for concern or celebration. In contrast, the National Assessment now also reports the percentage of students who are performing at or above "basic," "proficient," and "advanced" levels of achievement. Proficient, the central level, represents "competency over challenging subject matter," as demonstrated by how well students perform on the questions on each National Assessment test. Basic denotes partial mastery and advanced signifies superior performance on the National Assessment. Using achievement levels to report results and track changes allows readers to make judgments about whether performance is adequate, whether "progress" is sufficient, and how the National Assessment standards and results compare to those of other tests, such as state and local tests.

Recommendation:

- the National Assessment should continue to report student achievement results based on performance standards.

Use international comparisons

Looking at student performance and curriculum expectations in other nations is yet another way to consider the adequacy of U.S. student performance. The National Assessment is, and should be, a domestic assessment. However, decisions on the content of National Assessment tests, the achievement standards, and the interpretation of test results, where feasible, should be informed, in part, by the expectations for education set by other countries, such as Japan, Germany, and England. This, in turn, should take into account problems in making international comparisons truly comparable. In addition, the National Assessment should promote "linking" studies with international assessments, as has been done with the Third International Mathematics and Science Study, so that states that participate in the National Assessment can have state, national and international comparisons.

Recommendations:

- National Assessment test frameworks, test specifications, achievement levels and data interpretations should take into account, where feasible, curricula, standards, and student performance in other nations;
- the National Assessment should promote "linking" studies with international assessments.

Emphasize reporting for grades 4, 8 and 12

An aspect of the National Assessment design that needs reconsideration is age versus grade-based reporting. At its inception, the National Assessment tested only by age. Current law requires testing both by age (ages 9, 13 and 17) and by grade (grades 4, 8 and 12). Grade-based results are generally more useful than age-based results. Schools and curricula are organized by grade, not by age. Grades 4, 8 and 12 mark key transition points in American education. Grade 12 performance is particularly important as an "exit" measure from the K-12 education system. Grades 4, 8 and 12 are specified for monitoring in National Education Goal 3. Age-based samples may be more appropriate with respect to international comparisons and, given high school drop-out rates, would be more inclusive for age 17 than for grade 12 samples, which are limited to youth enrolled in school. However, assessing the knowledge and skills of out-of-school youth may properly fall under the purpose of another program, such as the National Adult Literacy Survey.

Although grade-based reporting is generally preferable, there is a problem about the accuracy of grade 12 National Assessment results. At grade 12, a smaller percentage of schools and students that are invited actually participate in testing than is the case with 4th and 8th graders. Also, more 12th graders fail to complete their tests than do 4th and 8th graders. In addition, when asked "How hard did you try on this test?" and "How important is doing well on this test?" many more 12th graders, than 4th or 8th graders, say that they didn't try hard and that the test wasn't important. Low participation rates, low completion rates, and indicators of low motivation suggest that the National Assessment may be underestimating what 12th graders know and can do.

One possible reason for low response and low motivation is that schools and students receive very little in return for their participation in the National Assessment beyond the knowledge that they are performing a public service. They do not receive test scores nor do they receive other information from the National Assessment that teachers and principals might wish to use as a part of the instructional program. This should be changed. The National Assessment design should use meaningful, practical incentives that will give school principals and teachers a greater reason to participate and students more of a reason to try harder. The underlying idea is clear: if principals and teachers see direct benefits, they are more likely to agree to participate in the National Assessment. Students may be more likely to take the assessment seriously if they see that their teachers and principals are enthusiastic about participating.

Recommendations:

- the National Assessment should continue to test in and report results for grades 4, 8 and 12; however, in selected subjects, one or more of these grades may not be tested;
- age-based testing and reporting should continue only to the extent necessary for international comparisons and for long-term trends, should the Governing Board decide to continue long-term trends in their current form;
- grade 12 results should be accompanied by clear, highlighted statements about school and student participation, student motivation, and cautions, where appropriate, about interpreting 12th grade achievement results;
- the National Assessment design should seek to improve school and student participation rates and student motivation at grade 12.

National Assessment results for states

In 1988, testing at the state level was added to the National Assessment. Previously, the National Assessment reported only national and regional results. For the first time, the information was relevant to individuals in states who make decisions about education funding, governance and policy. As a result, states now are major users of National Assessment data.

Participation was strong in the first state-level assessment in 1990 and has grown to include even more states. In 1996, 44 states and 3 jurisdictions participated in the math assessments at grade 4 and 8 and the science assessment at grade 8.

Currently, the National Assessment draws a separate sample to obtain national results in addition to the samples drawn for individual state reports. Testing separate national samples increases costs and creates additional burdens on states, particularly small states. If this practice can be discontinued, savings should be possible.

States participate in the National Assessment for many reasons, including to have an unbiased, external benchmark to help them make judgments about their own tests and standards. National Assessment data are used to make comparisons to other states, to help determine if curriculum and standards are rigorous enough, to develop questions about curricular strengths and weaknesses, to make state to international comparisons, and to provide a general indicator of achievement.

There is a strong interest among states to use the National Assessment to get state level information in reading, writing, science and mathematics. The level of interest in using the National Assessment varies with respect to the other subjects. State education officials are most interested in the National Assessment testing at grades 4 and 8. They say that

obtaining cooperation from high schools and 12th grade students is difficult. Also, from their perspective, 12th grade testing comes at the end of compulsory schooling, after which remediation is not feasible within the elementary and secondary system.

States are active partners in the National Assessment program. States help develop National Assessment test frameworks, review test items, and assist in conducting the tests. The National Assessment program is effective, to a great degree, because of the involvement of the states.

Because it is useful to them, and because they invest time and resources in it, states want a dependable schedule for National Assessment testing. With a dependable schedule, states that want to will be better able to coordinate the National Assessment with their own state testing program and make better use of the National Assessment as an external reference point.

Recommendations:

- National Assessment state-level assessments should be conducted on a reliable, predictable schedule according to a 10-year plan adopted by the Governing Board;
- reading, writing, mathematics, and science at grades 4 and 8 should be given priority for National Assessment state-level testing;
- testing in other subjects and at grade 12 should be permitted at state option and cost;
- where possible, national results should be estimated from state samples in order to reduce burden on states, increase efficiency and save costs.

Use innovations in measurement and reporting

The National Assessment has a record of innovations in large-scale testing. These include the early use of performance items, sampling both students and test questions, using standards describing what students should know and be able to do, and employing computers for such things as inventory control, scoring, data analysis and reporting. The National Assessment should continue to incorporate promising innovative approaches to test administration and improved methods for measuring and reporting student achievement.

Technology can help improve National Assessment reporting and testing. For example, reports could be put on computer disc, transmitted electronically, and made available through the World-Wide Web. Test questions could be catalogued and made available on-line for use by state assessment personnel and classroom teachers. Also, the National Assessment could be administered by computer, eliminating the need for costly test booklet systems and reducing steps related to data entry of student responses. Students could answer

"performance items" in cost-effective, computerized formats. The increasing use of computers in schools may make it feasible to administer some parts of the National Assessment by computer under the next contract for the National Assessment, beginning around the year 2000.

Other examples of promising methods for measuring and reporting student achievement include adaptive testing and domain-score reporting. In adaptive testing, each student is given a short "pre-test" to estimate that student's level of achievement. On the basis of the pre-test, higher achieving students are given tougher questions; students who know and can do less are given easier questions. Since the test is "adapted" to the individual, it is more precise and can be markedly more efficient than regular test administration. In domain-score reporting, a subject (or "domain") is well-defined, a goodly number of test questions are developed that encompass the subject, and student results are reported as a percentage of the "domain" that students "know and can do." This is in contrast to reporting results using an arbitrary scale, such as the 0-500 scale used in the National Assessment.

Recommendations:

- the National Assessment should assess the merits of advances related to technology and the measurement and reporting of student achievement;
- where warranted, the National Assessment should implement such advances in order to reduce costs and/or improve test administration, measurement and reporting;
- the next competition for National Assessment contracts, for assessments beginning around the year 2000, should ask bidders to provide a plan for (1) conducting testing by computer in at least one subject at one grade, and (2) making use of technology to improve test administration, measurement, and reporting.

OBJECTIVE 2: To develop, through a national consensus, sound assessments to measure what students know and can do as well as what students should know and be able to do.

Keep test frameworks and specifications stable

Test frameworks spell out in general terms how a test will be put together. The test frameworks also determine what will be reported and influence how expensive an assessment will be. Should 8th grade mathematics include algebra questions? Should there be both multiple choice questions and questions in which students show their work? What is the best mix of such types of questions for each grade? Which grades are appropriate for testing in a subject area? Test specifications provide detailed instructions to the test writers about the specific content to be tested at each grade, how test questions will be scored, and the format for each test question (e.g. multiple choice, essay, etc.).

Test frameworks and specifications are developed through a national consensus process conducted by the Governing Board. The national consensus process involves hundreds of teachers, curriculum experts, directors of state and local testing programs, administrators, and members of the public. The national consensus process helps determine what is important for the National Assessment to test, how it should be measured, and how much of what is measured by the National Assessment students should know and be able to do in each subject.

Through the national consensus process, both current classroom teaching practices and important developments in each subject area are considered for inclusion in the National Assessment. In order to ensure that National Assessment data fairly represent student achievement, the test frameworks and specifications are subjected to wide public review before adoption and all test questions developed for the National Assessment are reviewed for relevance and quality by representatives from each participating state.

An important role of the National Assessment is to report on trends in student achievement over time. For the National Assessment to be able to measure trends, the frameworks (and hence the tests) must remain stable. However, as new knowledge is gained in subject areas and as teaching practices change and evolve, pressures arise to change the test frameworks and tests to keep them current. But, if frameworks, specifications and tests change too frequently, trends may be lost, costs go up, and reporting time may increase.

Recommendations:

- test frameworks and test specifications developed for the National Assessment generally should remain stable for at least ten years;
- to ensure that trend results can be reported, the pool of test questions developed in each subject for the National Assessment should provide a stable measure of student performance for at least ten years;
- in rare circumstances, such as where significant changes in curricula have occurred, the Governing Board may consider making changes to test frameworks and specifications before ten years have elapsed;
- in developing new test frameworks and specifications, or in making major alterations to approved frameworks and specifications, the cost of the resulting assessment should be estimated. The Governing Board will consider the effect of that cost on the ability to test other subjects before approving a proposed test framework and/or specifications.

Use an appropriate mix of multiple-choice and "performance" questions

To provide information about "what students know and can do," the National Assessment uses both multiple-choice questions and questions in which students are asked to provide their own answers, such as writing a response to an essay question or explaining how they solved a math problem. Questions of the latter type are sometimes called "performance items." The two types of questions may require students to demonstrate different kinds of skills and knowledge.

Performance items are desired because they provide direct evidence of what students can do. Individuals confronted with problems in the real world are seldom handed four possible answers, one of which is correct. Although they may be desirable, performance items are more expensive than multiple-choice to develop, administer, and score.

Multiple-choice questions are desired because conclusions are more practical to obtain about the kinds of skills and knowledge assessed by these items, given the time available for testing. However, multiple-choice questions are more subject to guessing than are performance items.

Currently, all students tested by the National Assessment are given both types of questions. Generally, about half the testing time is devoted to each type of question, but the amount of time for each differs based on the skills and knowledge to be assessed, as established in the National Assessment test frameworks. For example, in a writing assessment, all students are asked to write their responses to specific "prompts." In other subjects, the appropriate mix of multiple-choice and performance items varies.

Recommendations:

- both multiple-choice and performance items should continue to be used in the National Assessment;
- in developing new test frameworks, specifications, and questions, decisions about the appropriate mix of multiple-choice and performance items should take into account the nature of the subject, the range of skills to be assessed, and cost.

OBJECTIVE 3: To help states and others link their assessments with the National Assessment and use National Assessment data to improve education performance.

The primary job of the National Assessment is to report frequently and promptly to the American public on student achievement. The resources of the National Assessment must be focused on this central purpose if it is to be achieved. However, the products of the National Assessment--test questions, test data, frameworks and specifications, are widely regarded as being of high quality. They are developed with public funds and, therefore, should be available for public use as long as such uses do not threaten the integrity of the National Assessment or its ability to report regularly on student achievement.

The National Assessment should be designed in a way that permits its use by others while protecting the privacy of students, teachers, and principals who have participated in the National Assessment. This should include making National Assessment test questions and data easy to access and use, and providing related technical assistance upon request. Generally, the costs of a project should be borne by the individual or group making the proposal, not by the National Assessment. Examples of areas in which particular interest has been expressed for using the National Assessment include linking state and local tests with the National Assessment and performing in-depth analysis on National Assessment data. States that link their tests to the National Assessment would have an unbiased external benchmark to help make judgments about their own tests and standards and would also have a means for comparing their tests and standards with those of other states.

Recommendations:

- the National Assessment should develop policies, practices and procedures that enable states, school districts and others who want to do so at their own cost, to conduct studies to link their test results to the National Assessment;
- the National Assessment should be designed so that others may access and use National Assessment test questions, test data and background information;
- the National Assessment should employ safeguards to protect the integrity of the National Assessment program, prevent misuse of data, and ensure the privacy of individual test takers.

NATIONAL ACADEMY PRESS

The National Academy Press was created by the National Academy of Sciences to publish the reports issued by the Academy and by the National Academy of Engineering, the Institute of Medicine, and the National Research Council, all operating under the charter granted to the National Academy of Sciences by the Congress of the United States.