NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

The Working Paper Series was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series.

U. S. Department of Education
Office of Educational Research and Improvement

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

NAEP RECONFIGURED:
An Integrated Redesign of the National Assessment of Educational Progress

Working Paper No. 97-31 October 1997

Contact: Steven Gorman
Assessment Group
(202) 219-1937
e-mail: steven_gorman@ed.gov

U. S. Department of Education
Office of Educational Research and Improvement

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Ricky T. Takai
Acting Assistant Secretary

National Center for Education Statistics
Pascal D. Forgione, Jr.
Commissioner

Assessment Group
Gary W. Phillips
Associate Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Office of Educational Research and Improvement
U.S. Department of Education
555 New Jersey Avenue, NW
Washington, DC 20208

Suggested Citation

U.S. Department of Education. National Center for Education Statistics. *NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress*, Working Paper No. 97-31, by Eugene G. Johnson, Stephen Lazer, and Christine Y. O'Sullivan. Project Officer, Steven Gorman. Washington, D.C.: 1997.

October 1997

Foreword

Each year a large number of written documents are generated by NCES staff and individuals commissioned by NCES which provide preliminary analyses of survey results and address technical, methodological, and evaluation issues. Even though they are not formally published, these documents reflect a tremendous amount of unique expertise, knowledge, and experience.

The ***Working Paper Series*** was created in order to preserve the information contained in these documents and to promote the sharing of valuable work experience and knowledge. However, these documents were prepared under different formats and did not undergo vigorous NCES publication review and editing prior to their inclusion in the series. Consequently, we encourage users of the series to consult the individual authors for citations.

To receive information about submitting manuscripts or obtaining copies of the series, please contact Ruth R. Harris at (202) 219-1831 or U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Ave., N.W., Room 400, Washington, D.C. 20208-5654.

Samuel S. Peng
Acting Director
Statistical Standards and Services Group

This page intentionally left blank.

NATIONAL CENTER FOR EDUCATION STATISTICS

NAEP Reconfigured:

An Integrated Redesign of the National Assessment of Educational Progress

Eugene G. Johnson
Stephen Lazer
Christine Y. O'Sullivan

In collaboration with:

Nancy Allen
John Barone
Johnny Blair
Henry I. Braun
John Burke
Nancy W. Caldwell
James E. Carlson
John Donoghue
Elizabeth Durkin
Carol Errickson
John J. Ferris
John Fremer
David S. Freund
Lauren G. Fried
James Green
Elissa Greenwald
Lynn Jenkins
Graham Kalton
Debra L. Kline
Barbara M. Klish
Robert Linn
Rosemary A. Loeb
John Mazzeo
Robert J. Mislevy
Dori Nielson
Norma Norris
Paul Ramsey
Linda L. Reynolds
Keith Foster Rust
Terry L. Schoeps
Juliet Popper Shaffer
Gerald Shelton
Brent Studer
Bradley J. Thayer
David Thissen
William C. Ward
Kim R. Whittington
Gita Z. Wilder
Paul L. Williams

August 1997

Prepared by Educational Testing Service and Westat under a cooperative agreement with the National Center for Education Statistics

U.S. Department of Education
Office of Educational Research and Improvement

This page intentionally left blank.

- TABLE OF CONTENTS -

CHAPTER 1
INTRODUCTION

EXECUTIVE SUMMARY

This chapter discusses the general purpose of this report, which is to outline the potential plans for the redesign of the National Assessment of Educational Progress (NAEP). We view NAEP as an integrated system, where a change in any one of the functional areas—cognitive measurement, contextual questionnaire development, sampling, data collection, scoring, analysis, and reporting—will have impact on the others. Thus, we argue that any successful redesign effort must consider NAEP as a whole. Our report considers overall NAEP designs and discusses the implications that each of these designs have for the various functional areas.

This page intentionally left blank.

CHAPTER 1
INTRODUCTION

- Eugene G. Johnson / Stephen Lazer -

For 27 years the National Assessment of Educational Progress (NAEP) has served as the nation's primary indicator of what students know and can do. Based on state-of-the-art measurement techniques, integrated use of cognitive and background questions, and representative national samples, NAEP has served as the country's best provider of reliable, objective information on student performances and on trends in academic achievement. NAEP data and reports are currently used in a variety of arenas and have informed the various debates about educational reform in the United States.

Over the three decades of its existence, the National Assessment has become one of the most innovative and successful surveys regularly conducted in the United States. NAEP has been asked to meet a wide variety of goals and priorities, and these have imposed constraints and demands faced by no other educational assessment program. The National Assessment has been called on to measure student knowledge of broad content domains and to gather in-depth contextual information, at the same time minimizing the burden faced by individual participants. NAEP has pioneered the use of performance assessment methodologies in large-scale settings, and NAEP staff have determined ways to use computerized image-processing technologies to score performance exercises in a cost-effective and statistically reliable manner. Psychometricians working on NAEP have developed procedures that allow for the combination of multiple-choice and performance measures into integrated scales. National Assessment analysts, programmers, and authors have developed artificial intelligence systems that generate computer-written natural-language reports for states that participate in NAEP. Overall, NAEP has become a gold standard: a model of innovation and accuracy.

However, it is perhaps NAEP's very successes that have created some of the strains that have led to the current redesign initiative. Because NAEP has shown a

consistent ability to satisfy program goals, new priorities have arisen—priorities that have often been in conflict with other program imperatives. In the late 1980s and early 1990s, NAEP was simultaneously asked to increase its use of performance assessment exercises *and* to test larger numbers of students as parts of state samples. New definitions of assessment content defined in National Assessment Governing Board (NAGB) *Frameworks* necessitated assessments involving both multiple-choice and constructed-response questions *and* the combination of these item-types in core reporting scales. NAEP was called on to measure trends *and* to reflect the best and most up-to-date curricular practices. NAEP was asked to provide timely information for policymakers *and* to allow in-depth analyses by education researchers in various subject disciplines. The publication of *America 2000: An Education Strategy* and the related work of the National Education Goals Panel increased the relevance of NAEP data and led to demands for more timely and frequent reporting; these demands came precisely at the time that the National Assessment was becoming more complex and expensive to administer. These developments and imperatives tended to interact and intensify: NAEP's increasing visibility and proven record of success led policymakers and educators to view the National Assessment as a vehicle of curricular reform and to demand even greater innovation in instrument design.

Overall, NAEP was called on to do more in an era of level funding. Concomitantly, the program's new priorities in no way excused it from its historical imperatives of providing the American public with statistical data of the highest quality, minimizing individual respondent burden, protecting participant confidentiality, responding rapidly to changes in policy, and allowing Department of Education policymakers maximum flexibility in their decision-making processes.

Despite the many and varied challenges that NAEP has faced, the program has continued to meet the majority of its goals. NAEP instruments, sampling designs, administration procedures, and psychometric methodologies have become models of innovation, yet have remained operationally and analytically feasible. NAEP reports have served the needs of a wide variety of audiences. The program's expansion to the

state level has made NAEP the benchmark against which the success of educational reform efforts are measured and new programs are planned. NAEP's management and implementation have fostered a flexibility that has allowed time that was once needed for test development to be spent, instead, on building consensus about what is to be measured and how it is to be measured. In addition, this flexibility has enabled assessments to evolve within the context of maintaining trend data. And NAEP's matrix-sampled design has allowed content, rather than testing methodology, to be the driving factor in the construction of NAEP instruments.

However, these successes have come at a price, where trade-offs have had to be made. With flexibility in schedule and instrument design has come analytic complexity. With performance testing have come further complications of analysis, difficulty in trend determination and, especially in the state program, significant expense. Evolutionary changes in assessments have required special bridging studies whose analyses must fall on the critical work path. New assessment *Frameworks* have invariably posed new developmental and psychometric challenges. Together, these changes in the assessment have prevented NAEP from realizing the efficiencies associated with the operational consistencies of most testing programs. All these factors have tended to add both complexity and cost to NAEP. With complexity has come lengthy reporting schedules. With expense has come limits on the number of assessments that can be administered.

The realization that trade-offs are inherent in the design and conduct of the NAEP program has led many associated with NAEP to begin asking fundamental questions about the program's future directions. For example, do assessments that necessitate extensive use of performance testing, while feasible if administered to small national samples, prove prohibitively expensive in a state-level program? Can a National Assessment that is expensive to administer and score serve the state linking function which many now envision for it? Is it possible that re-crafting the current integrated NAEP structure in favor of a modular design—in which certain inexpensive instruments represent an assessment core while other, more innovative modules could

be given as needed and analyzed off the critical reporting path—might better serve the program's new missions? In general, can one instrument satisfy all the publics who may wish to use its results? These and other questions began to suggest to many that if the basic purposes and structures of NAEP were not reexamined, the program might not be able to fully meet all of its conflicting imperatives.

The NAGB/NCES Redesign Initiative

Realizing that NAEP faced questions about its basic mission and design, NAGB began a careful and thorough examination of the nature and purposes of the National Assessment and the ways it could be redesigned to better meet its goals. This process involved input from Board members and staff, an independent Design/Feasibility Team[1] made up of eminent psychometricians, and hundreds of concerned citizens. The result of this initiative was the NAGB adoption, on August 2, 1996, of the *Policy Statement on Redesigning The National Assessment of Educational Progress.* This statement argues that NAEP should have three core objectives that would serve as the means for accomplishing its legislatively-mandated purpose of providing a fair and accurate presentation of educational achievement. These objectives are:

(1) to measure national and state progress toward the third National Education Goal[2] 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 as what students should know and be able to do

[1] The results of this team's work, published as *Design/Feasibility Team: Report to the National Assessment Governing Board,* had an important influence on the NAGB *Redesign Statement* and have also played a major role in organizing the work in this redesign planning effort.

[2] The third National Education Goal, called "Student Achievement and Citizenship," states that: "By the year 2000 all students will leave grades 4, 8, and 12 having demonstrated competency over challenging subject matter including English, mathematics, science, foreign languages, civics and government, economics, arts, history, and geography, and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our nation's modern economy." *The National Education Goals Report: Building a Nation of Learners.* National Education Goals Panel. (1996). Washington, DC.

(3) to help states and others link their assessments with the National Assessment and use National Assessment data to improve education performance[3]

The *Policy Statement* also identifies the primary audience for NAEP as, "The American public, including the general public in states that receive their own results from the National Assessment."[4] Perhaps most importantly, it states what NAEP 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.[5]

Having identified the three core program objectives, the central audiences for NAEP results, and some of the program's limitations, the *Policy Statement* then moves to identify redesign goals and policies related to each objective. The objectives, policies, and goals identified in the *Policy Statement* are unified by their calls for an increase in the predictability, frequency, timeliness, and efficacy of NAEP testing and reporting, and a simplification of NAEP analysis and design. The *Policy Statement* also reaffirms the importance of achievement levels and grade samples to NAEP reporting, and argues that NAEP must aggressively move to make use of computer-based assessment technologies. Finally, all changes to the NAEP system must, according to the *Policy Statement*, be implemented in a way that protects the integrity of NAEP data.

Once the NAGB adopted its policy on redesign, NAEP staff at the National Center for Education Statistics (NCES) affirmed their support for the goals of the

---

[3] National Assessment Governing Board. (1996). *Policy statement on redesigning the National Assessment of Educational Progress*, p. 3. Washington, DC: Author.

[4] Ibid., p. 3.

[5] Ibid., p. 4.

redesign initiative[6] and determined to launch a number of their own efforts to ensure that the NAGB policy would be faithfully and meaningfully implemented.

The Unifying Themes for this Report

This report represents one of those activities. As a result of the redesign planning competition, Educational Testing Service (ETS), Westat, and National Computer Systems (NCS) were awarded a cooperative agreement to develop designs for future NAEP. The product of that cooperative agreement is this report, which is unified by two core themes.

First, we believe that any successful NAEP design must be explicitly based on an understanding of which of the many program goals has primacy. This is because there are trade-offs inherent in any NAEP design; an assessment built to meet one set of goals may likely prove less than optimal for meeting others. Only if we keep these facts in mind can a rational redesign of NAEP be conducted.

Our second unifying theme is that NAEP is an integrated assessment system. One cannot view program areas in isolation. Changes in one area (such as cognitive instrumentation) will force changes in others (such as analysis and reporting). We therefore feel that investigating individual program areas separately is not the best way to rethink NAEP. Rather, we believe that NAEP must be viewed globally, and that the implications of changes in one program area must be viewed through the prisms of all other areas. In addition, we believe it important to consider systemic plans, and to gauge how such plans might be implemented throughout the program.

[6] National Center for Education Statistics. (1996, November 4). *An operational vision for NAEP—Year 2000 and beyond (draft).* Washington, DC: Author.

CHAPTER 2
AN INTEGRATED APPROACH TO THE REDESIGN OF NAEP

EXECUTIVE SUMMARY

This chapter tracks the evolution of the NAEP program and argues for an integrated approach to the redesign of NAEP. The following arguments and recommendations are made in this chapter:

- The NAEP system has evolved over time to meet a number of important priorities, including flexibility of instrument design and measurement precision, but the price of this evolution has been complexity and expense.

- Redesigning NAEP necessarily involves the setting of priorities as a number of goals of a future NAEP may be contradictory; that is, trade-offs are inherent in any design.

- Each of the functional areas of NAEP reflects potentially conflicting priorities within that area (for example, the design of precise samples involves tradeoffs between accuracy and cost).

- There are also trade-offs between functional areas (for example, limiting student cognitive testing time necessitates complex analysis procedures); it is not possible to have a NAEP program which is optimal for all functional components or possible program goals of NAEP.

- To be successful, a redesign effort must consider the NAEP program as a whole, because changes in one area will affect all other aspects of the program.

This page intentionally left blank.

CHAPTER 2
AN INTEGRATED APPROACH TO THE REDESIGN OF NAEP

- Eugene G. Johnson -

This redesign will mark the third major design for NAEP in its almost 30-year history. The initial NAEP was created in the 1960s to satisfy a particular set of priorities and specifications. In his 1996 paper on the history of NAEP,[1] Lyle Jones provides 15 different specifications that collectively determined the structure and operating characteristics of the nascent National Assessment. It is important to note that all specifications were jointly considered in designing the initial NAEP. As Jones notes, during the planning phase, and again during the startup operational phase, decisions were made that were decidedly suboptimal from the standpoint of one or more of the individual criteria, in order to produce a better overall design for the assessment.

The next major redesign of NAEP was implemented in 1984.[2] As in the initial NAEP, this was a comprehensive design that simultaneously considered all facets of the assessment. The new structure was also a significant departure from the original design in many ways, since it was intended (and developed) to accomplish different goals. The design for the subsequent assessments, up to and including the 1998 assessment, evolved from the 1984 design.

Making Choices Among Conflicting Goals

The process of redesigning NAEP necessarily involves the setting of priorities. And a number of possible goals of a future NAEP may be contradictory. It is important

[1] Jones, L. (1996, October). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25,* 15-21.

[2] Beaton, A., Messick, S., and Lord, F. (1983). *National Assessment of Educational Progress: A new design for a new era.* Princeton, NJ: Educational Testing Service.

to understand that, in any design, trade-offs must be made. Much of the complexity of current NAEP has risen from the fact that policymakers have been willing to achieve some important program goals at the cost of increasing the complexity of the system.

While the current assessment largely meets the many, often conflicting goals, it has done so at the price of increased complexity, cost, and reporting time. According to the National Assessment Governing Board (NAGB) Policy Statement on the Redesign of the National Assessment of Educational Progress, reducing these is now a central priority, as the goals of the program move more clearly in the direction of timely reporting and more frequent assessments. An assessment system designed to be simple, inexpensive, and capable of rapid reporting would look rather different from the current assessment.

Integration Rather Than Local Optimization

In the request for proposals for the cooperative agreement competition, the National Center for Education Statistics identified seven invitational priority areas for consideration in the redesign of NAEP. Somewhat restated from the request for proposals, the areas are:

- **Measuring Cognitive Skills** within the subject areas determined by the National Assessment Governing Board in ways specified by the *Frameworks,* while, as much as is possible, reducing student and school burden.

- **Measuring Contextual Information** such as socioeconomic status, instructional experiences, and home and community educational emphasis to provide context for the assessment results, while, at the same time, minimizing burden for the students, teachers, principals and other respondents.

- **Sampling** schemes designed to collect the needed data while, as much as possible, minimizing burden, encouraging high response, and striking an acceptable trade-off between the cost of the sample and the precision of the results.

- **Data Collection** procedures that obtain the required assessment information while, as much as is possible, minimizing burden on students, teachers, and schools, and, as much as is possible, maximizing the response rates and the motivation of the assessed students.

- **Scoring** procedures for constructed-response items that achieve high scorer reliability while being, as much as possible, cost effective.

- **Analysis** procedures that estimate the achievement characteristics of specified populations of students in ways that are statistically and psychometrically defensible, accurate, and reliable, while, as much as is possible, minimizing the processing time.

- **Reporting** techniques that inform educators, parents and leaders to take actions to improve their schools, while, at the same time, being statistically rigorous.

These seven areas (along with assessment *Framework* design and achievement-level setting) represent the functional pieces of NAEP. They therefore represent logical areas on which to focus the search for savings and efficiencies. Note that each of the areas listed above represents a potentially conflicting priority. For example, the sampling procedures are called on to permit the collection of the necessary data for the assessment while minimizing burden and striking an acceptable trade-off between the cost of the sample and the precision of the results. It is a fact from sample design theory that the precision of a sample increases as the sample size increases. It is also a fact that the cost of the sample increases as the sample size increases. Therefore, the desire to have a cost-effective, yet precise, sample involves a trade-off between the desired precision and the cost, both of which increase as the sample size increases. Similar tradeoffs are indicated within each of the other areas. It is simply not possible to have a program that is optimal for all competing priorities. Decisions have to be made in each case as to which competing priority has primacy. As noted by the Committee on the Evaluation of National and State Assessments of Educational Progress, there is a need

for a "clear sense of priorities from which decisions about audience, information needs, measurement design and administration design would flow."[3]

But the task of designing an assessment is even more complex than simply satisfying the conflicting constraints within one specific functional area. It is crucial to regard the NAEP program as an integrated system that includes framework development, cognitive and background instrument construction, sampling, administration, scoring, analysis, and finally, reporting.

As noted by the Design/Feasibility Team, a proposed design intended to improve a given component of the program may seem logical and efficient when viewed alone, but may be globally suboptimal for the program as a whole. For example, imagine increasing individual student assessment time to four hours in the eighth-grade assessment of United States history. This would allow NAEP to have every sampled student work with a variety of item types and cover all content needed for a thorough survey of the domain. As a result, NAEP might have one four-hour assessment book in history, rather than the current 33 books that each take approximately one hour to complete. In addition, an assessment of this length would likely prove reliable enough to eliminate the need for many of the sophisticated analyses currently a part of NAEP. On the other hand, such a system might introduce fatigue and motivation effects of a severity great enough to invalidate all findings. In addition, the burden of such a long test would likely reduce school and student participation rates to a point far below acceptable levels. Alternatively, the test could be split into four one-hour segments administered on concurrent days; however, students' participation would still be an issue, particularly if they are absent on one of the days. Student motivation may also still be an issue. Additionally, school participation would be compromised by such an extensive administration. So a change that seems to be an improvement from one perspective may, in fact, harm the system as a whole.

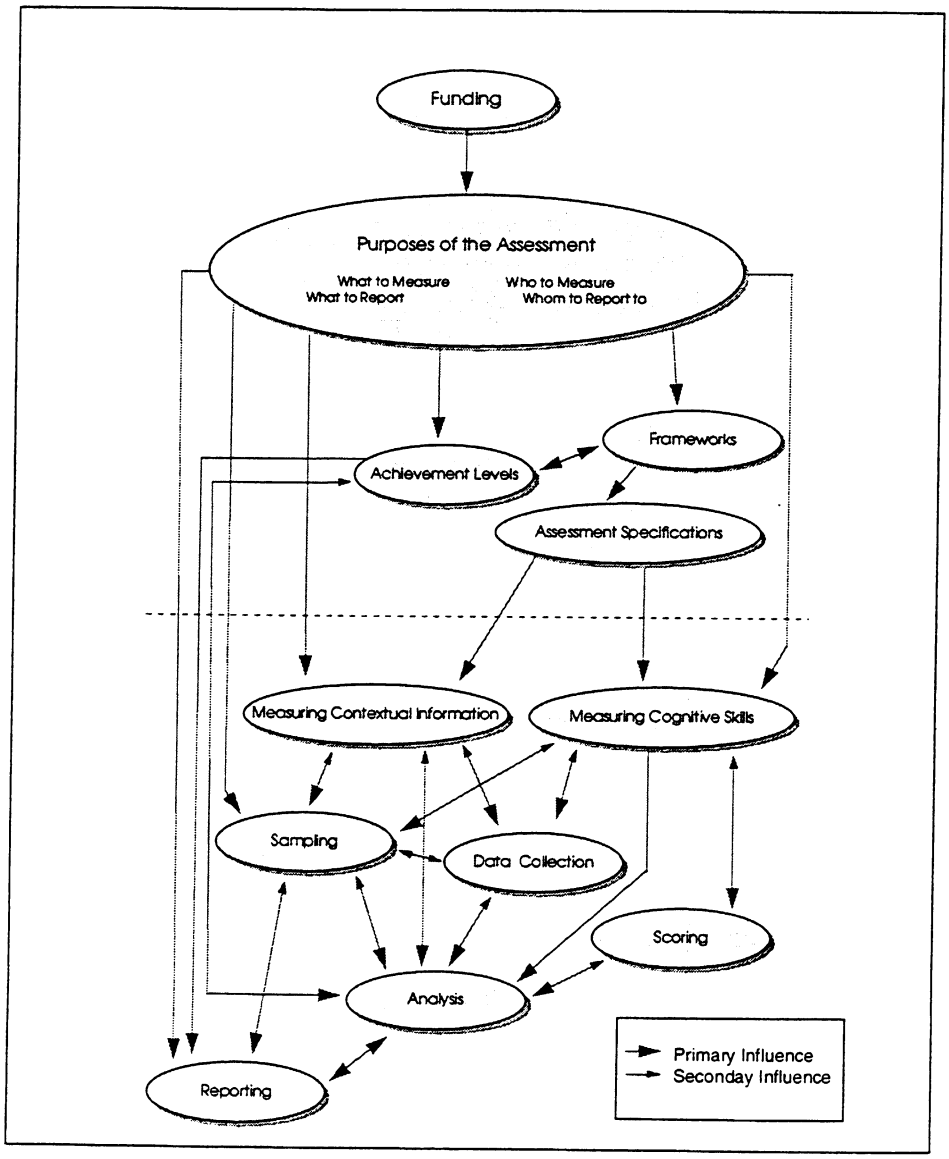
---

[3] Committee on Evaluation of National and State Assessments of Educational Progress, National Research Council. (1996). *Evaluation of "Redesigning the National Assessment of Educational Progress"*, p. 1. Washington, DC: National Academy Press.

The components of the NAEP system are deeply and inextricably interconnected so that a change in one area will necessarily affect the rest of the system. Changes in cognitive instrumentation will always lead to changes in analysis, scoring, and reporting. In many cases they will also lead to changes in sampling. Ideas that may reduce cost and complexity in one area may well increase them in others.

Figure 2-1 provides a graphical display of the interdependence of the functional areas of NAEP. Each of the ovals in the figure indicates a key functional area of NAEP.

Figure 2-1: The Interdependence of Functional Areas in NAEP



The areas are arranged from the top to the bottom of the figure in an approximate temporal order; thus, in a typical assessment cycle, the areas nearer to the top of the figure are a priority earlier than those nearer to the bottom. In addition to the seven functional areas indicated earlier, the figure identifies other key areas that have a fundamental effect on what NAEP is. For example, the first critical area is the amount of **Funding** available for the assessment. The next critical areas are the **Purposes of the Assessment**. The arrows in the figure indicate which areas influence other areas and the degree of the influence. Thus, the arrow between **Funding** and **Purposes**, with the large arrowhead pointing at **Purposes**, indicates that the available funding directly influences the four purposes of the assessment:

- **What to Measure**, meaning what subjects are to be assessed and how they are to be measured.
- **Who to Measure**, meaning what subgroups are to be assessed, including decisions about the inclusion of students not testable under standard situations or the exclusive assessment of certain subgroups consisting of students with specific experiences.
- **What to Report**, meaning the level of detail and mechanism for reporting the results.
- **Whom to Report to**, meaning the target audience or audiences of the assessment results.

As indicated by the large arrowhead, the **Purposes of the Assessment** directly influence the **Framework** which "shall describe the overall assessment and methodology, and the content to be assessed at each grade level."[4] Consequently, these purposes can be considered to be the overarching controllers of the entire assessment. The **Purposes of the Assessment** also influence the consensus process that leads to the preliminary **Achievement Level** descriptions and the **Framework**. These inform each other, as indicated by the double-headed arrow. Through its influence on the entire assessment, the **Framework** determines the assessment instrumentation and,

[4] National Assessment Governing Board. (1997). *Policy on framework and specifications development: General policy principles.* Washington, DC: Author.

eventually, the final assessment results that become the basis for determining the final achievement level descriptions and setting the final achievement levels. The **Framework** directly determines the **Assessment Specifications,** which provide the blueprint for the assessment.

The top portion of Figure 2-1 (above the dotted line), consisting of the **Purposes of the Assessment**, the **Framework**, the **Achievement Levels**, and the **Specifications**, represents the goals and the specifications of the assessment. The bottom portion of the figure (which consists of the seven operational areas discussed in this report) represents the functional areas which actualize the assessment. It is important to note that, with the sole exception of the feedback loop between analysis results and final Achievement Levels (indicated by the smaller-headed arrow), the procedures needed to actualize the assessment have little influence on the goals and specifications of the assessment. Rather, the operational characteristics of the functional areas which actualize the assessment are explicitly determined by the priorities and specifications provided by the **Purposes, Framework, Achievement Levels**, and **Specifications.**

Nevertheless, Figure 2-1 shows that the seven operational areas are largely interconnected in the sense that a change within one area is likely to have an impact on the conduct of another area. For example, the way that **cognitive skills** are measured can have substantial impact on other functional areas since the use of performance tasks has a profound effect on nearly every facet of NAEP, adding cost and complexity to instrumentation, sampling, administration, scoring, psychometrics, and reporting. In addition, NAEP subject-area *Frameworks* have been developed wholly within the context of a particular subject. As discussed in the next paragraph, this can lead to sampling and data collection inefficiencies.

Sampling influences analysis and reporting, since these are dependent on the data gathered. It also impacts data collection, since the design of the sample determines, in part, how the administration of the assessment will proceed. Sampling also affects database complexity. On the other hand, the *Frameworks*, instrumentation, and reporting requirements all influence sample design. For example, an assessment in

which each student takes half of all items will necessarily have a smaller national sample than an assessment in which each student takes, for example, 10 percent of the entire item pool. Assessments that allow different subjects to be spiraled together allow for more efficient sampling designs than do those that require these subjects to be administered in separate sessions.

Because **data collection** procedures generate the basic information of NAEP, they fundamentally affect analysis and reporting. They also influence sampling and instrument development in that certain kinds of samples or assessments may be impractical to implement or administer within the constraints of budget and time. Data collection is the most reactive of the seven program areas. One should never first design a data collection system and then force a survey to fit it. Rather, one must identify the goals of an assessment, including the kinds and nature of the reports that will be needed, and then determine practical ways to collect the necessary data. All other facets of the total assessment will affect data collection. For example, it is possible to design high-quality means of cognitive measurement that will add prohibitive expense and complexity to data collection procedures.

Currently, **analysis** cannot begin until **scoring** has been completed. Scoring affects analysis, since it provides the data needed for the analysis. Analysis has a feedback-loop influence on scoring, since, if the level of inter-rater reliability is unacceptable, the scoring must be repeated.

The statistical and psychometric procedures needed to analyze the data are determined by the instrumentation, administration, and sampling, which are, in turn, determined by the *Frameworks* and other statements of the goals and priorities of NAEP. The statistical and psychometric procedures used affect **reporting**, since these procedures determine the results that can be reported. In addition, they affect instrumentation, sampling, and data collection in that Balanced Incomplete Block (BIB) designs, necessary for both broad content coverage and reduced burden, impose some complexity.

The end result of the assessment are the reports which are the actualization of the "What to Report" and "Whom to Report to" components of the **Purposes of the Assessment.** The reports are clearly affected by all of the functional areas.

About the Following Chapters

The previous discussion has demonstrated that, to be successful, a redesign effort must consider the NAEP program as a whole because changes in one area will affect all other aspects of the program. Given the integrated nature of the assessment, it is necessary to consider the interplay of the functional areas on specific global assessment designs. This will be the topic of Chapter 3, where three distinct designs are considered in terms of their impact on each functional area.

In addition, there are areas where local improvements or efficiencies in one of the functional areas can be beneficial to the system as a whole. Chapters 4 through 10 describe improvements to each of the seven areas which could be profitably implemented as a part of one or more of the integrated designs. Each chapter discusses one of the seven invitational priorities and outlines recommendations for change. Sometimes these recommendations are for concrete changes based on available research. At other times, recommendations are for future research that should be implemented within the NAEP program.

This research will make most sense if conducted as part of a scheduled, predictable implementation plan that is itself designed to fit into the operational NAEP schedule. Such work must be carefully planned and staged so that it is completed in a timely fashion, and yet does not interfere with ongoing NAEP activities. We strongly believe that such an implementation plan is every bit as important as the revised design. Long-term research planning—involving research on both redesign implementation issues and on educational research issues—must become a regular part of the NAEP project management process.

This page intentionally left blank.

CHAPTER 3
POTENTIAL DESIGNS FOR NAEP

EXECUTIVE SUMMARY

This chapter examines three potential designs of the NAEP system. These are: a streamlined version of the current NAEP; a modular assessment design in which surveys are made up of core instruments used to generate reporting scales on a rapid schedule and special modules that may or may not be reported on such schedules; and, the replacement of current matrix-sample designs by systems of parallel test forms. The following arguments and recommendations are made in this chapter.

- In the short and intermediate term, substantial savings and efficiencies may be realized from a streamlined version of the current NAEP structure. In addition, the stability of assessment frameworks can lead to real gains in efficiency.

- In the longer term, modular testing designs likely make the most sense, from both measurement and economic perspectives, especially in programs designed for use at the state level.

- While "short forms" may be useful for certain linking and reporting purposes, it is not possible to replace the current NAEP matrix-sample design with a system of parallel forms. Use of parallel forms would severely limit the content coverage of NAEP instruments, and greatly increase test development and field testing burden, and would likely necessitate increases in student testing time.

This page intentionally left blank.

CHAPTER 3
POTENTIAL DESIGNS FOR NAEP

- Eugene G. Johnson -

Because of the integrated nature of the functional areas of NAEP, it is most useful to consider the interplay between the areas within the context of specific designs. By considering specific designs, the relative trade-offs between functional areas will become clearer. In this chapter, we will compare and contrast the current NAEP with the following three viable designs:

- **A streamlined NAEP** which maintains the basic structure of the current NAEP while adopting certain changes and compromises in order to reduce cost and complexity.

- **A modular NAEP** consisting of a core assessment designed to permit rapid and inexpensive assessment of the core features of the *Framework*, supplemented by one or more modules designed to answer specific research questions, to be presented to a subset of the assessed sample, and to be analyzed off the critical path.

- **A parallel-forms NAEP** consisting of test forms that are statistically parallel, each of sufficient length to measure the core features of the *Framework*.

Each of these variants has its own advantages and disadvantages. In this chapter, we will discuss the goals that each design satisfies, weigh the advantages and disadvantages of each design, and provide an indication of the relative costs of an assessment built to each design. These models have much in common, although they approach NAEP needs in somewhat different ways. In fact, a final design may involve elements from all three models.

In many instances, each of these three designs can serve as a mechanism to collect the information required for the NCES vision of comprehensive, standard, and

short-form assessments (see the Appendix). The NCES definitions[1] of these types of assessments are as follows:

- **Comprehensive** assessments would be the first assessments after a new *Framework* and would fully implement the characteristics of the *Framework* and *Specifications*. These assessments would allow for innovations in the cognitive instrumentation and could include a wide range of questionnaires to permit contextual analyses.

- **Standard** assessments would be subsequent assessments of the same *Framework* and would consist of items first administered in the comprehensive assessment. These assessments could have reduced use of background questionnaires and reduced reporting.

- **Short-form** assessments would be tests consisting of relatively few items designed to represent the main features of the cognitive scales.

Essentially, the NCES assessment types and the designs considered in this document are orthogonal to each other. The NCES vision is largely concerned with the amount of information that will be delivered by a particular type of assessment. The designs, on the other hand, are means to obtain that information. As will be discussed below, a streamlined NAEP design can be used to fulfill the information requirements of either a comprehensive or a standard assessment. The same is true of the modular design.

Before discussing the various designs, it is useful to consider the characteristics of the current NAEP.

The Current NAEP

The current NAEP evolved to satisfy a variety of purposes, and much of its complexity stems from the multiple purposes and goals that have accrued to the program

[1] National Center for Education Statistics. (1996, November 4). *An operational vision for NAEP—Year 2000 and beyond. (draft).* Washington, DC: Author.

over the years. A major legislative charge to NAEP is to measure progress in educational achievement.[2] To do this well, confounding effects due to changes from one assessment to the next in instrumentation or procedures must be minimized. That is, certain aspects of the measurement process must remain stable over time. However, NAEP has also been pressured to remain current by allowing the introduction of new curriculum concepts and by permitting the use of new measurement technologies. This has often resulted in new assessment *Frameworks* and corresponding changes in instrumentation and methodology. A solution to the dilemma of measuring trends while maintaining currency has been to institute a multi-component assessment system in which one set of assessments was used to measure trends over the long term while another was designed to be as current as possible. While this system has been successful in meeting both program goals, it has added complexity to sample design and assessment administration, required the analysis of two sets of data, and raised the risk of reporting confusion (when, for example, NAEP reports two sets of unrelated mathematics results).

Due to the range of National Assessment purposes, the program currently has five major assessment components, each of which is designed to accomplish certain goals: (1) assessments for long-term trend; (2) main cross-sectional assessments; (3) bridge studies; (4) special studies and probes; and (5) state assessments. While these components do, at times, share instrumentation (e.g., state NAEP uses a subset of the national instruments), they usually require different samples, different administration procedures, different analyses, and different reports.

In summary, the current NAEP has been developed to meet a number of goals. Specifically, it has been built to provide:

- a system that is flexible enough to allow the introduction of new assessment methodologies reflecting the latest thinking and most recent research into main-line assessments and reports

[2] Improving America's Schools Act of 1994 (P.L. 103-382).

- a system that permits assessment content to change and that allows changes in the populations to be assessed in the operational assessment, while attempting to maintain trend measurement

- an array of assessments in which measurement of long-term trend is conducted in a way that does not constrain overall project change or creativity

- a system that allows changes in instrumentation, sample design, and administration procedures very late in the cycle while allowing measurement of trend data

- an analytic design that allows assessments to be used operationally with limited field testing

- an assessment structure that allows aggregate coverage of broad content areas while limiting respondent burden to, for the most part, one hour

- a scoring and analysis plan that allows for wide use of performance testing

- an overall system that produces reliable statistics

- a rich database that supports both primary and secondary reporting and analysis

The assessment meets these priorities through a complex design, sophisticated sampling and administration procedures, state-of-the art scoring procedures, and complex and sophisticated data analysis and psychometrics.

Competing priorities, that the current assessment was not designed to satisfy, are:

- Simplicity of design

- Speed of reporting

- Low cost

A Streamlined NAEP

There is room for increased efficiency within the general constraints of the current NAEP system. A number of systemic changes could, independently or in

combination, reduce cost, time, and complexity while maintaining the central features and flexibility of the current NAEP. These changes include the following:

- **A more efficient use of released items** would have a dramatic impact on item development, scoring, and analysis. Under the current system, roughly one-quarter of the blocks of items presented in any assessment are released to the public. Item release serves the goals of helping the public understand what is being assessed and providing NAEP items for use by secondary researchers, teachers, and policymakers. However, new items must be developed to replace those released, and this entails costs for developing and field testing the new items, costs for developing scoring rubrics for the constructed-response items, and costs for analysis of the new items. A reduction in the number of items to be released could, by itself, lead to substantial savings. An alternative to releasing items actually used in the assessment is to construct and release items that appear parallel to the assessment items (these are called item variants). These items would give the public an idea of the characteristics of the assessment. However, since these item variants would not have been administered, item-level performance data would not be available. Another alternative that might be worth pursuing in "low-stakes" assessment circumstances is to simply reuse the released items, assuming that teachers will not teach to the items, given the low stakes nature of the assessment. For a National Assessment, without a state-level component, and without a link to a National Test, this might be a reasonable assumption.

- **Holding the *Framework* fixed for several cycles** would result in significant savings. Under this model, sufficient items would be developed and field tested for the first assessment to allow a release of exemplar items while maintaining a sufficient pool of secure items representative of the assessment *Framework*. All subsequent assessments would be based on the same set of secure items, with the consequent advantage that no further item development and field testing would be required. Assessment results would be in terms of aggregate scores from the secure items, while the characteristics of the assessment would be demonstrated by the exemplar items from the first assessment. Assuming that some sort of scaling is applied to the assessment data, item-level performance on the exemplar items could be predicted from the scaling model. This approach is completely consistent with the notion of a comprehensive assessment followed by a series of standard assessments.

- **More efficient field testing** would be possible if the assessment system were made more stable by holding the *Framework*, goals, specifications, and target populations constant. In such a situation, elaborate field tests are not as necessary. For example, new items can be embedded into the operational

version of NAEP when the specifications have been locked in place and the only need is to freshen the item pool. A full-scale field test would only be conducted when the assessment is changed in a meaningful way.

- **More efficient use of special studies** would simplify the assessment and would reduce costs. For example, while popular, the science hands-on experiments were quite expensive to administer, particularly to the 100,000 students in the state assessment. Rather than administering these tasks to all students in the assessment, treating the hands-on experiments as a special study to be administered to, say, 500 students in each state would have still allowed for state-by-state comparisons while reducing the number of kits needed by a factor of five. If state-level comparisons on these tasks are not needed, quite acceptable and useful national-level results for a variety of subgroups could be obtained from a sample of 5,000 students, a twentieth of the sample used for the 1996 assessment. (These special studies are examples of modules, to be discussed in the next section.)

- **More efficient use of performance tasks.** While NAEP has been viewed more favorably and seriously by curriculum experts because of the increasing use of performance tasks such as hands-on experiments and extended constructed-response exercises, such testing comes at considerable cost. Since performance tasks tend to take considerably more examinee response time than multiple-choice items, increasing the number of performance tasks in an assessment reduces the number of items that can be presented to each student. Furthermore, since these tasks require more effort on the part of the student, less motivated students tend to skip such items, leading to much higher omission rates and much lower quality of information relative to multiple-choice questions. Additionally, the student responses to the performance tasks must be professionally scored, raising resultant issues of scorer reliability. Also, a growing body of literature suggests that the generalizability of performance tasks may be disappointingly low. And, performance tasks are sometimes expensive to administer. Finally, there are situations in which performance measures provide little statistical information beyond that acquired through multiple-choice questions. Consequently, there are costs and benefits associated with the use of performance tasks. Certainly, if performance tasks are to be used, there should be an indication that additional information is being provided by these tasks beyond that provided by multiple-choice questions When such additional information is not provided, there is a real cost savings in reducing the number of performance tasks.

- **Better use of data available from other sources.** It may be possible to acquire valuable contextual information with little or no burden on the schools, teachers, and students. For example, NCES staff were able to work with the Department of Agriculture to obtain information from schools on students' eligibility for the national free and reduced-price lunch program. Other

federal government instruments, such as the School and Staffing Survey, gather information that might be linked to NAEP data. If such linking could be performed on a regular basis while protecting respondent confidentiality, it would greatly increase both the information available to NAEP and the accuracy of such information, and it would do so at minimal cost. Issues to consider are the quality of the data, unit of reporting, and age of the data.

- **Reduce the number of distinct assessment sessions required.** An assessment session within a school typically consists of 25 to 30 students, all of whom can be assessed following the same procedures. Thus, among other characteristics, the timings of the sections of the assessment booklets must be the same. For example, the 1996 assessment required different sessions for mathematics, whose booklets consisted of three 15-minute cognitive blocks, and science, whose booklets consisted of three 30-minute cognitive blocks, thus restricting the flexibility of the sampling and adding apparent burden to the schools to the potential detriment of school cooperation.

- **Reduce the sample size.** This has an obvious impact on the cost of the assessment. The obvious drawback is the reduction in the precision of the assessment results.

- **Reduce the frequency of background questionnaire collection.** For example, by eliminating the collection of teacher questionnaire data for certain assessments, the time and expense required to link these data to student cognitive data would be avoided. This is consistent with the notion of the standard assessment having a reduced collection of background information. The obvious apparent cost is the loss of information about the relationships between teacher practices and student performance. As we argue in Chapter 5, however, the link between teacher practices and student performance achieved by the current NAEP does not provide much useful information.

- **Reduce the size and complexity of reports for the standard assessments** to simple, short reports, with limited subgroup reporting, accompanied by a release of data on the NAEP Web site.

A Modular NAEP

NAEP's current practice is, for the most part, to give all types of cognitive items to all students. All students sampled for the reading assessment receive about the same percentage of extended constructed-response questions as is included in the assessment

as a whole. All students assessed for science in 1996 completed hands-on tasks. Such practices do send a message championed by curriculum specialists, and such exercises do fundamentally affect what the assessment measures.

However, such procedures come at a cost that is multiplied with greater assessment use. For instance, there were few, if any, economies of scale in administering the science assessment at the state level. In addition, the inclusion of large numbers of extended constructed-response items or hands-on tasks in core assessment scales means that subsequent trend uses will continue to be complex and expensive. Finally, the use of multiple-choice and constructed-response questions in integrated sections may have unintended effects on student response and omission patterns.

For these reasons, a modular design may be more efficient for use in a NAEP where cost, timeliness, and state-level testing are priorities. A modular NAEP might be built around a core assessment that consists of multiple-choice and constructed-response questions. The core would be designed to allow rapid and inexpensive reporting, form the basis of the basic assessment scale used for state and trend measurement, and facilitate linking of state and international assessments to NAEP. The core might use a Balanced Incomplete Block (BIB) spiral design (as does current NAEP) or might be structured in a different fashion.

Supplementing this core would be special modules that are designed to answer specific research questions or to provide additional data. These modules might include exercises that would prove prohibitively expensive if included in the core assessment given to all sampled students. Alternatively, the modules might consist of special studies. The modules would primarily be used at the national level, but could be used by states if so desired. In either case, the modules could be analyzed off the critical work path. The 1996 mathematics assessment approached such a design. Special estimation, thematic, and advanced blocks were administered at the national level only and were analyzed off the critical path.

Further, the modules could provide the venue for new item and assessment development. By excluding these innovations from the core assessment, the assessment would be made more robust for its chief purpose of reliably and efficiently measuring trends in achievement. At the same time, the modules would allow the innovations needed to reflect current thinking.

As noted by the Board on International Comparative Studies in Education,

> Many significant questions in comparative education are best addressed by small, focused studies, which may draw on a broad range of techniques and provide a deeper, richer sense of what education is, can, and should be. Thus, in addition to large-scale surveys, there is a need for a wide range of other cross-national research, such as ethnographic studies, case studies, small-scale focused quantitative and qualitative studies, and historical studies that would allow us to understand what it means to be educated in diverse settings throughout the world.[3]

The advantages described above for international comparisons are also applicable for the National Assessment. The modular design is consistent with the suggestion that NAEP be organized around a series of comprehensive and standard assessments. In this vision, the comprehensive assessments would consist of a wide variety of instrumentation, including special studies. The standard assessments could consist of only the core containing both multiple-choice and constructed-response items. Alternatively, the assessment of a broad domain such as science could consist of three core modules: life science, physical science, and earth science. Special studies could be added to this as needed.

As suggested in the above quote, not all modules have to be presented to nationally representative samples. For example, a module might consist of one-on-one intensive interviews in which a small number of students are asked about their approach to answering certain items. As another example, a small controlled study could be conducted to establish the relationship between student performance and certain characteristics of the student, teacher, or school: this study could include pre-

[3] Dorothy M. Gilford (Ed.). (1993). *A collaborative agenda for improving international comparative studies in education.* Board on International Comparative Studies in Education, Commission on Behavioral and Social Sciences and Education, National Research Council, p. 22. Washington, DC: National Academy Press.

and post-treatment measurement of student performance.

In examining a modular design, several issues must be considered. First, the changes in the construct measured by the core scale have to be investigated and discussed with subject-area educators. For example, there are areas in which constructed-response testing yields limited information over more traditional item types. If these areas can be identified, changes that have only a limited impact on construct validity can be recommended. More broadly, it is conceivable that the core would be designed to measure only a portion of the full framework.

To consider an example, the current mathematics assessment includes a mixture of multiple-choice, short-constructed response, and extended-constructed response items. All students receive a mix of the three item types, and all three item types are included in the mathematics scale. However, the core assessment could be defined to consist only of multiple-choice and short-constructed response items. This core would be presented at both the national and state levels. The extended-constructed response items would be in a separate module presented to a national sample of students only. The downside of such an approach is that the core scale would not include a measure of extended problem solving. The upside is that the administration of the assessment would be a good deal easier and less expensive, and there would still be a good deal of information from the module about how students perform with extended-response questions. In fact, since the module would be presented only to a national sample of students, there may be the opportunity to do more in-depth studies of student performance with such tasks than would be possible if all students in the national and state samples received these items. This approach would, in fact, provide more information about extended problem solving than an approach where extended problem-solving performance is forced onto the same scale as performance on the other item types. Finally, if particular states desired to obtain information about their students' extended problem-solving skills, those states could opt to participate in an administration of the module.

The implications of such modular designs for analysis and reporting must be examined. It is necessary to determine how modules can be reported along with the core scale and how the analysis of such modules can be integrated with the other parts of the assessment. How this would be done, and if it needs to be done, depends on the needs, goals, and available resources of the assessment. For example, a special study in the 1992 reading assessment used an integrated reading performance measure to determine students' oral reading fluency. Since the goal was to link oral reading fluency to reading ability as determined by the pencil-and-paper based assessment, it was necessary to test students under both conditions, and link their results. This was more complicated than the NAEP Reader study, where students were permitted to choose from a number of passages to read for their reading assessment; no linking was required in this study, thereby permitting the use of separate samples of students.

Finally, it is important to note that a modular program may only be acceptable to subject-area educators if they can be assured that the modules will be a regular, predictable part of NAEP.

A Parallel-Forms NAEP

Depending on its conceptualization, a parallel-forms NAEP could be a radical shift from the current instrumentation of NAEP. In current NAEP instrumentation, the full item pool is divided into subsets called blocks, and each assessed student is presented a booklet consisting of one or more of these blocks of items. Because NAEP uses various psychometric models to analyze the resulting data, and because student-level results are not legally permitted, it is not necessary that the distinct booklets used in the assessment be statistically parallel or that each booklet cover all the content domain specified by the assessment *Framework*. Further, the current method of item administration implicitly holds that the content that needs to be covered to adequately assess a given subject area is larger than one would want to present to any given student.

In contrast, a parallel-forms NAEP would consist of test forms that are statistically parallel and are, to the extent possible, equally representative of a content domain. Such forms might be administered in conjunction with a larger assessment during their first use, and might be used as stand-alone units later. Depending on their length, these forms would likely be adequate for the purposes of ordering students along broad dimensions, but, if they do not contain sufficient items, would be unlikely to provide much in-depth information about overall attainment in the subject area.

The parallel-forms NAEP subsumes as a special case the short-form NAEP with the chief difference being the length of the instrument. That is, we allow for the possibility of a form longer than that presented to a student in the current assessment. In a recent draft policy statement, the National Assessment Governing Board states:

> The purpose of the short form NAEP is to develop smaller modules of the NAEP assessments which have certain characteristics, including, but not limited to, the following: such short form surveys would (1) be faster to administer, score, and report than the full NAEP assessment; (2) allow replications of comprehensive/standard assessments in off-years; (3) provide linked estimates of NAEP scores to users; (4) provide linked estimates of achievement levels to users; (5) allow embedding of NAEP modules into states' systems of assessment.
>
> The purpose of the short form NAEP is *not* to provide estimates of individual student performance. It will be designed specifically to provide robust estimates of *group performance*, insofar as the technology will allow.[4]

The parallel-forms NAEP would serve the same purposes as the above described short form, while, if necessary, being longer than a typical NAEP test form. Depending on its length, parallel-forms NAEP is essentially Variant #1 or #2 in the Design/Feasibility Team description of market basket reporting.[5] The parallel-forms NAEP could easily be a component of a modular NAEP, either as the core or as a module. Following the purposes of the NAGB policy draft, the parallel-form is a

---

[4] National Assessment Governing Board. (1997, May). *Developing a short form of the National Assessment of Educational Progress* Draft policy statement. Washington, DC: Author.

[5] Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design/Feasibility Team: Report to the National Assessment Governing Board*, p. 6-28. Washington, DC: National Assessment Governing Board.

module that will be linked to overall NAEP. However, parallel-forms NAEP could also be used as the core assessment, although this use will likely require longer forms than that implied by a short form.

Parallel Forms As a Module

The NAGB policy suggests that the short form be linked to the full NAEP assessment. This corresponds to a modular NAEP where the core is an extensive assessment, perhaps based on a BIB design. Linking short forms to a main NAEP reporting scale would likely require statistical linking procedures such as projection. For reasons given below, the quality of the linkage might increase as the length of the instrument increases. Thus, the link between a parallel-forms NAEP and the main NAEP reporting scale might be stronger for a longer form. Issues of fatigue and motivation begin to come into play as the test length increases, however (see Chapter 9 for a discussion).

Alternatively, the forms could simply be treated as alternate booklets and included into the standard item response theory (IRT) and conditioning procedures. Such forms, already calibrated to the main NAEP scale, would facilitate the linking of states' assessments to the NAEP scale, through the parallel-forms item parameters. This was the approach used to link data from the 1994 NAEP mathematics assessment and the North Carolina End of Grade mathematics test.[6]

Parallel Forms As the Core

Using a set of parallel forms as the core assessment could put severe limits on the coverage of the *Framework* if the forms are required to be of the same length as the current NAEP booklets. The key issue is the number of items in a given form required to adequately cover the *Framework*. As the length of the form decreases, the tendency of the parallel forms to be largely measuring the core content of the *Framework* increases.

---

[6] Williams, V.S.L., Billeaud, K., Davis, L.A., Thissen, D., & Sanford, E. (in press). Projecting to the NAEP scale: Results from the North Carolina End of Grade Testing Program. *Journal of Educational Measurement.*

(Of course, other knowledge and skills could be measured in modules.) Depending on how many parallel forms there are, and their length, the richness of the item-level data could also suffer, since there could be fewer examples of each type of item (such as items measuring the ability to multiply fractions). Requiring rigidly parallel forms places additional constraints on booklet assembly, which, in turn, influences item development.

The constraints imposed by the length of the form are most apparent when performance tasks are to be included in the assessment. Statistical algorithms are available for constructing a number of parallel short forms from a suitable startup pool of items.[7] However, these work best when there are many items comprising each form, as is the case with multiple-choice items and short constructed-response tasks. Having many items in a form provides considerable degrees of freedom in terms of matching the forms on content, difficulty, and other characteristics.[8] There are usually fewer performance tasks included in any one form because these require more time for completion. The result is that there is less freedom in terms of matching the performance tasks in terms of, for example, their expected score distributions. It is still possible to build parallel forms by first matching up the performance tasks as well as is possible, and then selecting the multiple-choice and short constructed-response items to make the overall expected score distributions as close as is possible. This matching up comes at the cost of assuming that the performance tasks and the multiple-choice items are measuring the same dimensions. It also assumes that the parallelism holds across subgroups of the population.

The low generalizability of performance tasks[9] across students is an additional concern with a series of parallel forms that are to include such tasks. Simply put, a growing body of literature indicates that the performance presented by a student on

---

[7] Stocking, M.L., Swanson, L., & Pearlman, M. (1991). *Automated item selection (AIS) methods in the ETS testing environment.* (Research Report 91-5). Princeton, NJ: Educational Testing Service.

[8] Even so, more detailed specifications than currently provided for NAEP would be needed to allow for parallel-forms construction.

[9] Brennan, R.L., & Johnson, E.G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14,* 9-12.

one performance task may not well predict that student's performance on another task (this is called person-by-task interaction). This means the supposedly parallel forms may only be parallel in the sense of expected performance. What is not being measured is the variability due to the person-by-task interaction.

This variance component, which can easily be more than half of the total variance of a student score on a given task,[10] needs to be included in the measures of the standard errors of any group-level statistics based on the parallel forms. The magnitude of this component can be reduced by lengthening each form to include more performance tasks and/or having several parallel forms used in any assessment. Estimating this component would require having pairs of performance tasks administered to the same students. This could be done during a field test.

The implication of using parallel forms as the NAEP core, particularly when performance tasks are specified in the *Framework*, is that a choice must be made. On the one hand, if the desire is to mirror the *Framework*, so that performance tasks are to be included in the parallel forms, then the cost is low generalizability and high person-by-task interaction—unless there are many parallel forms or unless each form contains a number of performance tasks, that in turn implies a significant resource expenditure. Alternatively, the forms could be restricted to multiple-choice and short constructed-response items, resulting in a high reliability measurement of a restricted portion of the *Framework*.

Other Issues

- **Costs of the necessary field work to develop rigidly parallel forms.** These would include issues and costs related to a design that requires an initial pilot to screen items (of around the size of a current NAEP field test), followed by a large-scale field test (of the general size and complexity of a current main assessment). It needs to be stressed that much of the needed analysis would be performed on the field test data. Ignoring this earlier analysis, the activities required after the operational assessment appear much simpler. However, the complexity of the full process, beginning with the initial design of the instruments, would be at least as high as with current NAEP.

[10] Ibid.

- **The score to be used for the parallel forms.** Another question concerns the way in which the multipoint constructed-response items are to be counted into the score. (Chapter 9 has a discussion on this topic.)

- **Effects of releasing example parallel forms.** The benefit of a release is that the public can see exactly what the test instrument is. The disadvantage of a release is the need either to continually develop other forms that maintain parallelism or to reuse the released forms. The former implies additional expense; the latter leaves results open to artifactual increases in performance due to knowledge of the test items.

- **Stability of meaning of the parallel forms over many assessment cycles.** There are examples from the long-term trend assessment of items that have become easier or harder over time because of changes in societal and educational emphasis. Similarly, forms that are parallel at the time of construction may not remain parallel if specific knowledge and/or teaching practices change. Such items will not operate in the same way as when the parallel forms were developed. The result is that the forms are no longer parallel. (This is an example of what is called item parameter drift, which is discussed further in Chapter 9.)

- **Consistency of parallelism across population subgroups.** The various forms are parallel in terms of expected performance across the entire population. While NAEP items are evaluated to assure that they do not differentially discriminate between students from different subgroups with matching overall abilities (this is called differential item functioning), the demands of rigid parallelism for the alternate forms make it essential for the test characteristic functions for the alternate forms to be equivalent across subgroups. It is entirely possible to construct forms which are parallel for the population as a whole, but not for certain important subgroups.

Recommendations for the Overall Design

The relative merits of the three potential designs (Streamlined NAEP, Modular NAEP, and Parallel-Forms NAEP) are best considered in the context of the kind of data on cognitive skills and contextual information that is desired.

However, there are four main findings in this chapter.

1) Stability of assessment frameworks and instruments will result in savings under any model.

2) In the short and intermediate term, substantial savings and efficiencies may be realized from a streamlined version of the current NAEP. In particular, special studies, limitation of item release reductions in performance assessment, reconceptualization of field testing and reconsideration of background questionnaire analysis procedures may introduce real efficiencies.

3) In the longer term, modular testing designs likely make the most sense from economic and measurement perspectives (especially in programs designed for use at the state level). These approaches will reduce cost and reporting time, while maintaining system flexibility and allowing NAEP as a whole to continue to cover broad context domains.

4) While short-forms may be useful for some linking or reporting purposes, it is not possible to replace the current NAEP matrix sample design with a system of parallel forms. Use of parallel-forms would severely limit the content coverage of NAEP and greatly increase test development and field test burden, and would likely necessitate increases in student testing time.

This page intentionally left blank.

CHAPTER 4
MEASURING COGNITIVE SKILLS

EXECUTIVE SUMMARY

This chapter contains an examination of cognitive testing in NAEP. Specifically, we discuss the appropriate mix of multiple-choice and performance measures in NAEP instruments from both evidentiary and business perspectives, possible uses of computer based testing (CBT) in NAEP, and possible efficiencies that might be introduced. The following arguments and recommendations are included in this chapter:

- The types of evidence that can effectively be gleaned from large-scale testing should be considered when selecting item-types for NAEP. Items that yield information that can only be interpreted if much other data is present should not be included in the core components of assessments used to produce scales.

- Modular testing designs offer the best hope of introducing efficiencies in cognitive measurement, while at the same time allowing for the coverage of broad content frameworks and for the evolution of assessments.

- Initial forays into CBT should focus on skills not amenable to traditional measurement, and should be used to study feasibility and equity issues.

- Because of motivation and fatigue factors, student testing time should not be expanded beyond current limits. NAEP should also maintain its focus on the coverage of broad content areas. To meet both these goals, NAEP must continue to rely on instruments using matrix-sample designs.

This page intentionally left blank.

CHAPTER 4
MEASURING COGNITIVE SKILLS

- Stephen Lazer / Robert J. Mislevy / Kim R. Whittington / William C. Ward -

Introduction

Cognitive measurement is central to the National Assessment of Educational Progress (NAEP). Determining what students know and can do is the central goal of the program. Without valid cognitive testing, all other aspects of the program are rendered meaningless.

The purpose of this chapter is to examine the key questions facing NAEP in the area of cognitive testing and to make recommendations for this component of the NAEP program. We will also endeavor to place cognitive testing within the context of the unified NAEP system.

This chapter has seven distinct sections.

- The first section includes a brief review of **cognitive testing in NAEP today.**

- In the second section, we discuss the **roles of multiple-choice and constructed-response testing** in NAEP, and make recommendations as to the **"appropriate mix"** of item types that NAGB seeks. This section also discusses changes in the structure of the assessment that may make performance assessment better suit program goals. Finally, this section discusses the cost implications of different "mixes" of items.

- The third section of this chapter contains a discussion of **computer-based testing,** and the ways in which such testing can most profitably be implemented in the NAEP program.

- The fourth section examines the current structure of NAEP cognitive testing to identify any places in which current practices have an adverse effect on the system.

- In the fifth section we discuss efficiencies in cognitive measurement that will profit any model of NAEP.

- The sixth section includes an examination of cognitive testing in a streamlined version of the current NAEP, in a modular NAEP, and in a parallel-forms NAEP.

- The final section summarizes concrete recommendations.

Measuring Cognitive Skills in the Current NAEP

It is perhaps in the area of cognitive testing that NAEP has realized some of its greatest successes and advances. In recent years, NAEP has pioneered the use of performance testing methodologies in large-scale assessments. Specifically, consistent with the increased interest in performance tasks, NAEP has expanded its use of these sorts of exercises to the point where such items make up over half the assessment time on most NAEP surveys. Every student in the 1996 NAEP science assessment completed a hands-on experiment as part of their testing experience. These developments have ensured that NAEP scales measure skills and outcomes not limited to those amenable to multiple-choice items. Such innovation has doubtless helped increase acceptance of NAEP by the educational communities in the various subject areas, and has rendered NAEP a respected and much-copied program. It has forced substantial psychometric innovation and has created significant opportunities for interesting and educationally meaningful reporting.

However, the ever-expanding use of constructed-response and performance testing in NAEP has also had negative impacts. The dollar cost of scoring and analysis has become substantial, especially as NAEP has moved into state-by-state testing. In fact, an Educational Testing Service project conducted for Peat-Marwick indicated that constructed-response testing was one of the major cost drivers in the NAEP program. In some subjects performance testing may have introduced or worsened motivation problems on the assessments. Constructed-response questions have also increased the complexity of trend measurement, one of the core goals of NAEP.

Thus, in recent years, NAEP has both borne the costs and reaped the benefits of including performance testing in its large-scale assessments. Any NAEP redesign must

take stock of these costs and benefits, and explicitly examine the role of performance testing in the national assessment. The questions must not be limited to "all or nothing" scenarios about whether or not one should use performance testing, but must rather include considerations of the uses of such testing that will best meet the needs of the program, and will best fit into a globally optimal program plan.

The situation is further complicated by the fact that technology may be providing the means to test outcomes previously beyond the reach of large-scale assessment. Computer-based testing will likely revolutionize cognitive measurement, and as NAEP moves toward the future we must examine ways to bring such technologies into use.

NAEP today stands at a crossroads. Demands for wider and more frequent use of NAEP instruments call on us to reexamine the uses of performance measurement in the national assessment and to determine the ways in which constructed-response and performance exercises can be used most appropriately and efficiently. Yet the program must endeavor both to stay at the forefront of measurement and to make effective use of new technologies as they become available.

The cognitive skills to be measured within subject areas (e.g., reading) are determined through a legislatively mandated consensus process managed by the National Assessment Governing Board (NAGB). These measurement objectives take the form of *Frameworks* delineating the important content and process areas to be assessed. In general, the *Frameworks* are updated to reflect the most current thinking in the field and to reflect changes in curriculum and instruction. Even when the objectives of the assessment remain the same, the *Frameworks* may be updated to reflect advances in content and instruction, as exemplified by the 1996 mathematics assessment.

Spurred on by these ambitious *Frameworks*, NAEP has pushed back the borders of large-scale testing. NAEP assessments make extensive use of constructed-response exercises; in most subject areas well over half of student time is spent on such items. Constructed-response exercises are not segregated from multiple-choice items. Students may be asked to answer multiple-choice items and constructed-response questions about a given piece of stimulus. Hands-on performance testing has been used at the national and state levels in science. The reading assessment is based on authentic,

naturally occurring passages. BIB spiraling has allowed these item types to be used in assessments where student testing time is restricted. And NAEP partial-credit analysis models have provided for the creation of integrated scales from "mixed" instruments (that is, instruments with a variety of item types). Table 4-1 indicates the number of exercises and percentage of assessment time devoted to multiple-choice, short constructed-response, and extended constructed-response questions in a number of main NAEP assessments,[1] so that the large proportions of NAEP instruments devoted to performance testing are clear.

These advances have come at a price. In 1996, the NAEP program scored almost 10 million student responses. Such scoring has both cost and schedule implications. Open-ended questions introduce complexity into trend measurement. Constructed-response testing often provides much information at the upper end of the scale, which few students reach. In all cases, open-ended questions take longer to answer than multiple-choice items; thus, wide inclusion of the former reduces the number of questions that any single student is presented. In such a situation, analysts are forced to rely more on statistical models and less on direct measurement of student proficiency. These costs and complexities will need to be examined as part of any NAEP redesign.

In addition to specifications of the content and process dimensions to be assessed, the *Frameworks* also provide preliminary achievement level descriptions. Furthermore, the *Frameworks* specify the proportions of various types of assessment items—such as multiple- choice, short constructed-response, extended constructed-response, and performance tasks—as well as other types of assessment activities such as group assessment.

[1] The figures for assessment time are based on the assumption that each multiple-choice question takes students one minute to answer, while short and extended constructed-response questions take two and five minutes, respectively.

Table 4-1: Numbers and Percentages of Items and Distribution of Assessment Time in NAEP Instruments

	Number of Items				Percent of Items			Percent of Time		
	mc	scr	ecr	total	mc	scr	ecr	mc	scr	ecr
1992 Reading										
Grade 4	42	35	8	85	49%	41%	9%	28%	46%	26%
Grade 8	57	53	13	123	46%	43%	11%	25%	46%	29%
Grade 12	63	54	16	133	47%	41%	12%	25%	43%	32%
1994 Reading										
Grade 4	39	37	8	84	46%	44%	10%	25%	48%	26%
Grade 8	41	55	13	109	38%	50%	12%	19%	51%	30%
Grade 12	44	62	13	119	37%	52%	11%	19%	53%	28%
1992 Mathematics										
Grade 4	99	54	5	158	63%	34%	3%	43%	47%	11%
Grade 8	118	59	6	183	64%	32%	3%	44%	44%	11%
Grade 12	115	58	6	179	64%	32%	3%	44%	44%	11%
1996 Mathematics										
Grade 4	80	55	9	144	56%	38%	6%	34%	47%	19%
Grade 8	93	62	7	162	57%	38%	4%	37%	49%	14%
Grade 12	91	68	7	166	55%	41%	4%	35%	52%	13%
1994 Geography										
Grade 4	59	23	8	90	66%	26%	9%	41%	32%	28%
Grade 8	84	32	9	125	67%	26%	7%	44%	33%	23%
Grade 12	85	25	13	123	69%	20%	11%	43%	25%	33%
1994 History										
Grade 4	62	26	6	94	66%	28%	6%	43%	36%	21%
Grade 8	101	35	12	148	68%	24%	8%	44%	30%	26%
Grade 12	104	33	19	156	67%	21%	12%	39%	25%	36%
1996 Science										
Grade 4	51	73	16	140	36%	52%	11%	18%	53%	29%
Grade 8	74	100	20	194	38%	52%	10%	20%	53%	27%
Grade 12	70	88	30	188	37%	47%	16%	18%	44%	38%

NOTE: Main BIB only; exclude theme blocks and estimation blocks
mc = multiple-choice
scr = short constructed-response
ecr = extended constructed-response

NAEP's experience with these sorts of measures has been largely successful. NAEP has developed cognitive exercises that measure the range of outcomes described in forward-looking assessment *Frameworks*. NAEP psychometricians have built analysis

techniques that have enabled performance exercises to be included along with more traditional measures in estimates of population abilities. NAEP has achieved a position in the forefront of large-scale testing programs.

Yet new demands on NAEP have changed the ways in which one must view the role of performance testing in the national assessment. Calls for expansion of the state program, more frequent assessments, faster reporting, reduced costs, and simpler analyses require a new look at performance testing. Specifically, they must lead to a reexamination of the percentages of open-ended exercises in NAEP assessments, of the appropriate and efficient ways to place such exercises into the assessment system, and of the best ways to analyze and report the results of such testing. It is to these concerns that we now turn.

The "Appropriate Mix" of Multiple-Choice and Constructed-Response Items

The NAGB *Policy Statement on the Redesign of NAEP* indicates that NAEP assessments should be composed of an appropriate mix of multiple-choice and constructed-response items. The statement, however, does not indicate the precise nature of that mix. Nevertheless, investigations in this area are at the core of any efforts to find efficiencies in cognitive testing in particular and NAEP in general. In 1996, constructed-response scoring for mathematics and science cost over $4 million for the operational testing alone. Field test scoring costs added roughly $1.2 million. Reducing the number of science items by 50 percent would alone have saved NAEP $700,000 to $800,000 in scoring costs during the operational 1996 science assessment. In 1994, the use of open-ended items in the reading assessment added substantial complexity to the measurement of trends. In 1994 and 1996, scoring added eight to ten weeks to the analysis and reporting schedule and also added substantial statistical work because of the many constructed-response items.

It does not follow that NAEP should simply abandon performance testing in favor of multiple-choice measures, however. Constructed-response questions were

included in NAEP for a number of good reasons, most of which still apply. To determine the "appropriate mix" of multiple-choice and open-ended items, several questions must first be answered. These include:

- What are the purposes for including constructed-response questions in NAEP? Are these exercises designed to provide evidence of student thinking or achievement, to contribute to overall proficiency scales, or both? Are they intended largely to send the "right instructional message?"

- Can NAEP effectively use certain types of assessment that have had their genesis in the classroom? What sorts of contextual information are needed if we are to make real sense of the results of performance assessment?

- Is it necessary to give all types of exercises to all students or to include them in overall proficiency estimates?

- What costs are associated with performance assessment, and what savings would result from reductions in the proportion of NAEP instruments devoted to performance items?

The problem in considering these points is that there has been little systematic thinking about the purposes of multiple-choice, constructed-response, and performance exercises in NAEP. Specifically, there has been little research regarding the types of evidence best produced by different forms of testing, nor how those types of evidence fit into the NAEP schema. Neither has there been much careful thought given to the costs of such testing. The fact is that multiple-choice and performance exercises both have distinct advantages and disadvantages for programs such as NAEP. Exclusive reliance on one form of testing would likely hamstring the program. Using only multiple-choice items would at least harm the program's face validity. Relying solely on open-ended testing would introduce serious measurement and cost inefficiencies into the program. In a simple sense, both types of testing have costs and benefits that can be clearly described. We will now examine these advantages and disadvantages.

Advantages and Disadvantages of Multiple-Choice Items

Multiple-choice testing has the advantages of facilitating broad content-area coverage and being relatively inexpensive. Such testing is also well understood by the assessment community. Far from being limited to recall of knowledge, multiple-choice testing can be used to measure a broader range of skills than is commonly understood. In the specific context of NAEP, students do not tend to omit such items, and use of multiple-choice makes it more likely that tests can be targeted to difficulty levels appropriate to assessed populations.

However, such items have disadvantages as well. They give little or no evidence of student thinking. They provide no examples of student work that can be used in either reports or secondary analyses. They simply cannot measure the full range of educational outcomes (for example, writing is not effectively measured through multiple-choice testing). Multiple-choice testing may emphasize factual recall or the disaggregation of skills that are, in their real-world applications, integrated. Finally, multiple-choice items often have limited content validity in the eyes of subject-area curriculum experts, who worry that such testing may send a message that does harm to American education.

Advantages and Disadvantages of Constructed-Response Items

These limitations moved the designers of NAEP *Frameworks* to call for greater use of constructed-response items and other performance testing. And such assessment methodologies have several obvious advantages for NAEP. Because of NAEP's unique position, many educators have argued that the use of performance testing can send a powerful instructional message to America's schools. The use of substantial amounts of constructed-response testing has enabled NAEP to build instruments that have far greater legitimacy in the eyes of content-area experts than did older national assessment surveys. In other words, they have allowed NAEP instruments to seem to measure skills most valued by educators. And this is not mere appearance: in a real sense, they have allowed for the measurement of skills not possible in more traditional instruments. Open-ended testing has also provided examples of student work to enrich

NAEP reports, and has allowed for the sorts of secondary analyses most important to many educators (see, for example, the 1993 report Can Students Do Mathematical Problem Solving?[2]). Such assessment has helped to solidify NAEP's reputation as a leading innovator in the field of large-scale testing.

Finally, performance testing has had two advantages in measurement terms. Open-ended exercises have tended to provide good discrimination at upper parts of the scale. In the past, NAEP had been criticized for being weak in this area. Also, performance testing has provided some insight into whether or not students are capable of certain complex behaviors (although the reasons for the students' failures are a good deal harder to glean from NAEP data, a problem that is discussed in more depth below).

But while performance testing has added much to NAEP, it has had a negative impact as well. We have already discussed the cost associated with such testing. Such cost is exaggerated in large applications like state NAEP. Use of open-ended testing often comes at the expense of content coverage. In NAEP, this leads to one of two problems: Either assessments as a whole do a worse job of covering a content domain, or assessments expand to allow for both performance testing and domain coverage. The latter development necessitates concomitant increases in sample sizes and in development and scoring budgets. Finally, use of open-ended testing, while forwarding certain of NAEP's goals, actually introduces complexities into other of its core goals, including the measurement of trend.[3]

So we have a seeming conundrum: There are advantages and disadvantages to the widespread use of constructed-response and performance assessment in NAEP, but these factors do not, in and of themselves, suggest an appropriate mix of items. In addition, the situation is more complex than the above discussion suggests: There is

---

[2] Dossey, J.A., Mullis, I.V.S., & Jones, C.O., (1993). *Can students do mathematical problem solving?* (Report No.23-FR01). Washington, DC: National Center for Education Statistics..

[3] Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design/Feasibility Team: Report to the National Assessment Governing Board.* (pg. 6-28). Washington, DC: National Assessment Governing Board.

almost certainly not one "solution" that fits all NAEP subjects. Answers in mathematics are likely to be different than those in writing, because the best way of measuring knowledge and skills will necessarily vary from subject to subject.

What is important is the development of a perspective that will allow assessment designers to evaluate how best to use different types of exercises in NAEP. To begin to develop such a model, we must first conduct two different analyses. First, we must examine the costs and benefits of the use of different types of testing in NAEP from an evidentiary-reasoning perspective. Solid judgments can only be made if we have a solid understanding of the types of inferences that we may and may not draw from different exercises. Certain forms of testing, while seemingly attractive, may be close to useless within certain assessment contexts. Second, we must understand the operational, schedule, and cost implications of different models, so that we can accurately evaluate the potential of different models to introduce cost efficiencies.

Performance Tasks, Content Validity, and NAEP—An Evidentiary Perspective

Factors that Determine the Evidentiary Value of Assessment Data

An evidentiary-reasoning perspective[4] is often helpful for examining assessment issues. It encompasses traditional measurement concerns of reliability, generalizability, and validity, but extends easily to less familiar issues and examples.[5] Such a perspective is especially important in NAEP, where performance tasks have sometimes been used without a firm understanding of the inferences one intends to draw from the results of such testing. In addition, the context of NAEP places severe limits on how the evidence from such testing may be used. Only through understanding the types of evidence produced by performance testing and the intricate relationships between the

[4] Schum, D.A. (1994). *The Evidential Foundations of Probabilistic Reasoning*. New York: John Wiley and Sons.

[5] Mislevy, R.J. (1996). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

context of testing and the inferences one may draw can one move to find how performance testing can best be used in NAEP.

Specifically, consider three factors that determine the evidentiary value of data evoked by a given task—that is, the observable behavior by a student induced by the task, ranging from a simple right/wrong answer to a videotape and a verbal protocol made as students work and talk their way through a complex open-ended task.

The identification of the target inference. Data do not convey evidence other than in relation to some inference (or hypothesis, or conjecture, or projection)[6] so that the same data can provide much evidence about some inferences but little or none about others.

What else one knows. For example, finding a carpet fiber at the scene of a crime has no relevance, or evidentiary value, to the hypothesis of Tom being guilty if an observer doesn't know that Tom owns a car with that kind of floor covering; but if the observer does know this, the same data can have considerable value. The value for the "Tom is the perpetrator" hypothesis drops if one learns his car was stolen before the crime was committed (although finding this out may inspire one to generate new conjectures, for which the same data may have high evidentiary value).

The knowledge and state of mind of people involved in producing data and evaluating data, with regard to conventions under which data are produced, the hypotheses that these data will be used to test, and the way in which they will be interpreted. This is especially important in assessment. Unless a student knows what is valued and how it will be evaluated, the chances of that student doing his/her best will drop considerably. When performances must be evaluated and these evaluations constitute the data, their evidentiary value drops to the degree that raters' notions of what to attend to and how to evaluate it are unclear or variable.

---

[6] Schum, D.A., op. cit.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational Measurement* (3rd ed.). New York: American Council on Education/Macmillian.

How do these factors come into play in different assessment contexts? These factors have differing impacts depending on the nature of an assessment task and the specifics of the testing context. Assessment in a classroom situation in which teachers know quite a lot about students may yield evidence that is impossible to obtain in a large-scale setting. Conversely, large-scale testing may allow for gathering evidence on the proficiency of broad populations on wide content domains. Such evidence cannot be gathered at the classroom level without making insupportable assumptions about the representativeness of a given classroom.

In the sections below, we will examine the different ways in which these "rules of evidence" inform the discussion of the structure of NAEP assessments. In a recent paper, Baxter and Glaser[7] have suggested an interesting categorization of assessment tasks. Specifically, they pose two continuums—process openness and content richness—into which tasks must be categorized. On one axis, tasks may be viewed as "process-constrained" (such as a multiple-choice question) or "process-open" (for example, an essay question that uses very little scaffolding). There is no dividing line here; most tasks will fall at some point between the extremes of the continuum. Similarly, tasks may be categorized as being "content-rich" or "content-lean" depending on the extent to which test takers are asked to bring external knowledge to the tasks.

Assessment exercises may fall into any of a number of categories, according to Baxter and Glaser. "Process-constrained, content-rich" tasks may be used to measure factual knowledge. Such exercises may be either multiple-choice or open-ended. "Process-constrained, content-lean" tasks may be used to test basic skills in a matter comparatively unaffected by content. "Process-open, content-lean" exercises are often designed to assess reasoning skills in a way uncolored by content knowledge or the lack thereof. Finally, "process-open, content-rich" tasks allow for the assessment of the ability of test takers to integrate content knowledge and skills.

---

[7] Baxter, G.P., & Glaser, R. (in press). *An approach to analyzing the cognitive complexity of science performance assessment.* Los Angeles: UCLA Graduate School of Education, Center for Research on Evaluation, Standards, and Student Testing.

There are a number of evidentiary reasons to use constructed-response or performance exercises. On the one hand, one may wish to assess content knowledge without the subtle hints often dropped in multiple-choice settings. Further, "knowing" and presenting an answer is surely a somewhat different skill than "choosing" the best answer from a selection. Constructed-response testing may also be used to measure simple reasoning or presentation skills not effectively tapped by multiple-choice questions. Such measurement can often be done through short constructed-response items that pose comparatively few problems for large-scale assessments, although the benefits of such items may or may not outweigh the monetary costs.

On the other hand, the greatest attraction of use of extended constructed-response and performance exercises is that they seem to provide measurement and evidence of students' reasoning skills. This is especially true of tasks that are "process-open." Thus, "process-open" tasks relate to the work being done by many cognitive scientists in a way that more constrained tasks do not. However, the ability of NAEP to profitably use such tasks in its core assessment instruments is open to serious question.

Participants' Understanding of Conventions and "Rules of Evidence"

Tasks that are "process-open" provide the possibility of obtaining evidence about students' planning, problem-formulation, organizational, and inquiry skills. If useful evidence is to accrue, however, there must be a common understanding of what is valued and how performances will be evaluated, among those who design the tasks and the assessment, those who take them, those who evaluate performances, and those who analyze and report the results. Such tasks work best when students have a schema to guide what they produce, even though the task is open to many choices and much freedom. Lots of examples and talked-though evaluations are the only way to do this, AP Studio Art being an excellent exemplar (see Example 1 below). The less open the process, the easier this phase is, and that is one reason that constrained tasks like multiple-choice are so widely used; almost all American students know the rules of this game.

The issue is this: To obtain high evidentiary value from a complex task, it is necessary (but not sufficient) that the student understand what is being sought, that he or she be motivated to perform accordingly, and that the evaluators interpret the performance along those same lines. Following is one example from the Advanced Placement program in which this issue was dealt with successfully, and three (examples 2, 3, and 4) from NAEP, in which problems emerged along these lines.

Example 1: AP Studio Art Portfolio Assessment. Students spend hundreds of hours creating their portfolios, and raters who are art educators and teachers spend hundreds of hours evaluating the work. The program can be viewed as a framework for evidence about skills and knowledge, around which high school teachers build art courses with wide latitude for topics, media, and projects. A common understanding of what is valued and how it is evaluated in the central scoring emerges through teacher workshops, talked-through examples with actual portfolios, and continual discussions about how to cast and apply rating rubrics to diverse submissions. This is, at heart, a social phenomenon, not a measurement phenomenon. Measurement models for ratings are valuable nevertheless to illuminate how raters use evaluative criteria, to characterize uncertainty about students' scores, and to highlight ways to improve the program[8]—but using the models does not *create* the shared understanding of what is valued and how it is evaluated, and this shared understanding does not occur by accident.

Example 2: NAEP Hands-On Science. A small group of students took one of NAEP's hands-on science performance blocks and one regular block with multiple-choice and short open-ended tasks, then retrospectively talked through their work.[9] Their scores were similar on the regular block, but the score distribution for the performance block was bimodal. At first blush, this might suggest that the students were similar with respect to science "book-learning" knowledge but differed as to the

[8] Myford, C.M., & Mislevy, R.J. (1995). *Monitoring and improving a portfolio assessment system.* (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

[9] Yepes-Baria, M. (1995, April). *Task analysis of science performance tasks and items: Identifying relevant attributes.* Paper presented at the annual meeting of the National Council on Measurement in Education San Francisco (ERIC #ED-388-676/TM-023-675).

degree to which they could apply their knowledge. However, the protocols suggested that score differences on the hands-on block had little to do with the students' understanding of the science involved in the open-ended block. Rather students'scores depended on how they managed their time as they worked through the block—either conscientiously, often at the cost of not answering all parts, or apportioning time so as to answer everything, often at the cost of care and comprehension.

Example 3: NAEP Arts. The field trials of the NAEP Arts assessment presented examples of "interesting tasks" that the committees of content area experts expected to produce relevant *data*, but ended up producing little *evidence* because the committees did not communicate targets of inference, or ways of parsing performances for evidence about any particular targets of inference. Students were asked to, say, "dance like a tree in the wind;" their performances were videotaped, and much *data* were gathered. The question then was "How do we score it?" After lively discussion, groups of raters could usually come to some working agreements, but they often had material disagreements as to what to value and how to evaluate it—which could mean the difference between a high rating and a low rating for the same performance, depending on criteria about which the performer had no information.

The data provided can allow for some discussion of what representative samples of students did when presented with a particular set of "process-open" tasks. However, the usefulness of such data as a means of spreading students across a scale and discussing their "arts proficiency" is limited, at least partly because of the idiosyncrasies associated with the small number of tasks that could be offered. Also, the assessors have limited information about the instructional background of the students, and thus getting any evidence about the effectiveness of dance instruction is problematic. In general, this supports the notion that performance assessment provides rich evidence in settings in which one knows quite a lot about the test takers; divorced from such a rich interpretive context, these data may be of less use.

The experiences of NAEP in the arts also point to another issue related to performance or extended constructed-response testing. Some forms of performance

measurement are associated with the establishment of mastery (for example, a solo flying test as a prerequisite of a pilot's license). Others are associated with "expert" populations (such as musicians auditioning for a part). At the very least, they seem to presume a population of some homogeneity, at least of general instructional experiences. In programs like NAEP, we have none of these advantages. Using performance tasks to measure novice populations is likely to provide little usable evidence. One would not give the solo flying tests used to certify pilots to groups who had never flown a plane: it is a costly and expensive way to find evidence for a proposition that is on its face obvious—that is, that people with no training or experience cannot pilot planes.

Example 4: NAEP Extended-Response Tasks. Performance tasks have been used increasingly in NAEP because they provide a different kind of evidence than multiple-choice items about what students know and can do. There is a fairly clear distinction between what might be called "short constructed-response" tasks and "extended constructed-response" tasks. Both require open-ended responses, as opposed to multiple-choice items, but they differ notably in the amount of time and entry required of students. Short-constructed response tasks often involve one or two minutes of student time, and a couple of sentences or a paragraph, for a response. In subjects like mathematics, the tasks may involve solving simple problems. Extended constructed-response tasks generally require more time, exhibit interconnections among aspects of the performance, and can lead to poor performance for a variety of reasons—one of which is lack of the requisite skills or knowledge.[10] We have seen, as an example of motivation problems, large discrepancies in performance tasks among various students who do equally well on multiple-choice questions. In the 1994 geography assessment, as many as 40 percent of the twelfth-grade students simply did not bother with individual extended-response tasks, which required greater effort, as

[10] This is something of a general problem in performance testing. It is often difficult, even in well-defined tasks, to determine whether failures are due to a lack of knowledge, a weakness in skills, or the inability to integrate knowledge and skills. See Messick (1989) for a discussion of this.

opposed to omit rates of 5 to 10 percent for short-constructed response tasks and 1 to 5 percent for multiple-choice items.

"Process-open" tasks allow for the *possibility* of observing students carry out processes that one has not explicitly told them to do. As a result, they inevitably allow for the possibility that the students will not do those things anyway, just because they did not know that was what the developer hoped they would do. These kinds of tasks can thus provide evidence about the existence of a student's skills, knowledge, and strategies, if one happens to see the corresponding behavior; but failure to see it is not necessarily strong evidence of its nonexistence. The weight of evidence to be expected from such tasks will depend partly on how well the students understand what the developer of the tasks is looking for. Tasks that are more complex and rich, therefore, provide more evidentiary value when students have had a chance to become familiar with the conventions and rules of evidence of the situation—generally over a period of time, with talked-through examples, perhaps in the context of instruction (e.g., AP Studio Art, ARTS Propel). The same tasks can be expected to have less evidentiary value, other things being equal, when they are "dropped in from the sky" on assessment day, as in NAEP. In such situations, the processes must be far more constrained; if one wants evidence of a given behavior one must specifically instruct students to engage in such behavior. However, this limits the extent to which the tasks will provide independent evidence of reasoning and complex problem-solving skills.

The Role of "What Else One Knows"

The evidentiary value of a datum with regard to some particular conjecture is defined by the force and direction by which it changes one's beliefs about that conjecture, and that will depend on what else one knows. (Testimony that "someone who looked like John" was at the scene of the payroll robbery would seem to hold evidentiary value for the hypothesis that he committed the crime—but it does not if one already knows with certainty that John either was or was not present at the scene.) Moreover, different observers have different knowledge bases from which to interpret

data. The evidentiary value of an assessment task can take very different values within its inherent potentials as a function of who is considering which inferences, and what else they know.

Perhaps the most robust finding from research on large-scale assessments is that the most powerful predictors of success are "opportunity to learn" variables. Simply stated, students who have never had a chance to work with, say, decimal fractions, usually do not perform as well on tasks that involve them as do students who have worked with them before. What we can learn about a student from such a task will differ if we know he or she has not had a chance to work with decimals, or has worked with them and this task is typical of those used in his or her class, or has worked with them but this task poses a novel (to him or her) connection between decimals and fractions. With such knowledge, we could pose more focused conjectures, and enjoy greater evidentiary weight to address them from a well-chosen task. The same exact task—the same exact observation it evokes—would have less evidentiary weight without the auxiliary information.

This phenomenon is especially pertinent for content-rich tasks. The more ambitious and subtle the connection of a task to particular schemas or concepts is, the more likely it is that particular background experiences determine the possibility of a student's engagement with that phase of the task. In an instructional setting, when a teacher knows a great deal about the student's educational history, he or she can use such a task knowing the kinds of support and grounding the student has had, thereby obtaining high evidentiary value from the resulting data. In a large-scale assessment, when one knows little or nothing about individual students, one enjoys far less evidentiary value because observers must entertain such a wide variety of conjectures to account for poor performance. Well-constructed content-rich tasks can provide invaluable evidence about students' learning when they are interpreted in light of knowledge about those students' curricula and background. The same tasks result in weak evidentiary value ("low generalizability") when they are "dropped in from the sky" to support inferences about "typical performance" in the lot of them.

Content-rich tasks provide the most evidentiary value for the sharply focused conjectures that can be posed in light of much background information, best exemplified by how a teacher uses them in one-on-one tutoring. They provide much evidence, though somewhat less, when teachers use them at the classroom level, since teachers know the relationships of these tasks to the assignments, lessons, and experiences that have transpired in the class. They provide useful evidence, but again less, to once-removed assessors in curriculum- or syllabus-based large-scale assessment, since there is less knowledge about students' experience with the relevant content. They hold least potential for evidentiary value in large-scale assessments such as NAEP. The lack of contextual knowledge limits their value regardless of the conjectures on which they are brought to bear. Nevertheless, they can prove much more useful for some conjectures in this setting than others, as discussed below.

Defining the "Target of Inference"

A datum's "relationship to the target of inference" works in two important ways. First, it plays a role in determining potentials and constraints. If one does not build the need for advanced or rigorous use of particular content into a task's requirements, one will not obtain much evidence for inferences about students' degree and structure of knowledge of any specified area of content. If an assessment developer does not provide the opportunity, or impose the necessity, for organization, exploration, or planning, he or she will not obtain much evidence for inferences about these aspects of students' proficiency.

Suppose a task is constructed so it *is* possible to observe data that bear on the skills or knowledge of interest. The second way that "relationship to the target of inference" comes into play is that the particular form of this intended inference determines in part how much evidence is conveyed by the data from the particular task. The evidentiary value will be greater when the inference is specific to the task or to others very much like it, and lesser when the intended inference is to a domain of tasks that vary in content or processes (i.e., a broadly defined scale). The former

situation means evidence can be obtained for inferences that are detailed, rich, and cognitively interesting, but more closely bound to the particulars of the task. The latter situation leads to the "low generalizability problem" that Shavelson[11] and others associate with performance tasks. This low generalizability often leads to the need for unrealistically large numbers of time-consuming tasks if assessments are to achieve the reliabilities we have become used to with traditional short-answer tests. The evidentiary value of data from a task depends not just on the particulars of that task, but on how it is to be combined with data from other tasks.

The former situation—inferences more closely bound to the particulars of the task—are desirable for use in instruction and in psychological research. For instruction, a teacher wants to learn how the student is thinking, to give feedback or otherwise extend the student's knowledge; the detail and richness of contextualized processes are invaluable for doing this. For research, protocol analyses provide clues for insight into problem-solving and learning processes. Either way, the target inferences concern discovering the *existence* of which of many possible ways of thinking the student is pursuing. *Appropriately targeted* content-rich and/or process-open tasks fit the bill. In the latter situation—inferences from a sample of contexts to a broader domain of contexts—the target inferences concern *averages* or *typical behaviors*; they are often meant to inform people who are removed from the instructional setting, by means of a broad summary of the outcomes of instruction. The low generalizability phenomenon suggests that the time and expense obtained to secure detailed but contextualized data from content-rich and process-open tasks do not constitute weighty evidence about performance in other, different, content-rich and process-open tasks.

This distinction is analogous to one that arises in survey sampling. There is a sense in which one obtains "the same amount of data" by asking one respondent 100 different questions, and 100 different respondents from the same city one question each, about attitudes concerning politics, for example. The first mass of data can yield

[11] Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessment: Political rhetoric and measurement reality. *Educational Researcher, 21* (4), 22-27.

insights into the beliefs and motivations of that individual, and bring considerable evidence to bear on inferences concerning that individual—but not much evidence about the typical opinions or the distributions of opinions in the city. The larger sample of respondents (about each of whom little is learned) provides more evidence for inferences of these latter types and correspondingly less evidence for inferences concerning individuals.

As an aside, it is worth mentioning that studies with multiple targets of inference may find those targets in tension. For example, attempting to delve in depth into the cognitive processes of students may lead one to develop an assessment that is far less than optimal for measuring changes in student achievement over time. For the former purpose, one likely wants performance instruments that are closely tied to current classroom practice, complex judgments from raters, and substantial background information about the instructional experiences of students. To serve the latter purpose, one wants instruments that are not substantially changed between assessments, and are not unduly influenced by changes in the way judges score student performance. This has, in programs like NAEP, led to the expensive maintenance of duplicate systems of assessments.

Issues related to the definition of targets of inference (or the lack of such definitions) that arise in NAEP, concern the interaction between two kinds of distinctions: inferences about particular contexts versus generalizations to broader contexts, and inferences about individuals versus inferences about groups. Let us consider the possible combinations.

Inferences About Individuals and Generalization to Broad Domains. This is the typical context of large-scale educational assessment: The primary inferences concern individual students, regarding generalization to a domain of tasks on which those presented are considered a sample. Process-constrained short-answer tasks, especially machine-scorable ones, evolved to meet the needs of this particular niche. Relatively inexpensively, one can administer to students about whom one knows relatively little, large numbers of these tasks in order to obtain high reliability for

inferences about individuals' overall proficiency in the domain. Broader ranges and greater richness of content generally require greater numbers of tasks to achieve a given level of reliability, although commonality in instructional background among the students being tested effectively renders the range of content less broad. Thus, tasks that are more content-rich tend to appear more often, and provide greater evidentiary value, when they are employed in assessments linked with instruction than when they are "dropped in from the sky." Tasks that are process-open tend to be more time consuming, and therefore provide less evidentiary value, than process-constrained tasks for a given amount of resources if both types of tasks are meant to provide evidence about the same domain of generalization.[12] As noted above, the problem is exacerbated if the students are not familiar with the standards by which their performances on process-rich tasks will be evaluated. In sum, when tasks that are content-rich and/or process-open are meant to provide evidence about individual students' performance in larger domains, they provide considerably greater evidentiary value when both the tasks and the domain of generalization have been thoughtfully matched to the students' experiences.

Inferences About Individuals and Inferences About Particular Contexts. This is the typical context of cognitive and educational psychological research and tasks that are used as an integral part of instruction. The observations of the particulars of solutions to tasks that have been carefully targeted to both the students and the instructional or research objective can provide the most valuable evidence for the specific and carefully targeted inferences that typify these settings. For example, a tutor seeking to extend a student's knowledge will select a task that has both some familiar aspects and other carefully chosen novel elements—a choice that clearly demands considerable knowledge about the student. The same one-hour process-open and content-rich task that provides sharp evidence and learning opportunities for the

[12] Wainer, H., Lukhele, R., & Thissen, D. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31,* 234-250.

Wainer, H., & Thissen, D. (1993). Combining multiple choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education,* 6, 103-118.

student to whom it is well targeted, however, will produce no evidence and much frustration to the student wholly unfamiliar with the content or the process expectations.

Inferences About Groups and Inferences About Particular Contexts. This is the context originally envisaged for NAEP: Data were gathered with a large collection of tasks, chosen to reflect curricular and educational aims rather than to provide scores for individual students. The primary inferences concerned patterns of response (including descriptive information as well as overall quality or correctness) to these tasks, one by one, from samples of students with various demographic and instructional backgrounds. NAEP was, for its time, a pioneer using content-rich and process-open tasks in large-scale assessment, and subject area specialists obtained a wealth of evidence for inferences of most interest to them—inferences concerning span of skill and knowledge, curriculum coverage, and relationships between instruction and performance on particular tasks. These same data, however, proved cumbersome and ineffectual for rather different inferences that were foremost in the minds of other stakeholders in NAEP, namely policymakers and the public at large. Simple summary indices on "how kids were doing overall" were more to the point for them—inferences more in line with generalizations to domains of tasks, to which we now turn.

Inferences About Groups and Generalization to Broad Domains. NAEP evolved through a series of steps toward first defining, then successively refining, machinery to provide group-level indices of overall proficiency and change in proficiency over time. From the original item-by-item results, NAEP incorporated average percent-correct for subsets of items in the 1970s, distributions of domain scale-scores in the 1980s, and proportions of students above designated points along domain scales in the 1990s. The latter two developments were *defined* in terms of student-level performances in broad domains of items, such as reading or science, but *reported* only at the level of groups. Given the breadth of content definition—e.g., science tasks covering, as an inclusive picture, all the content and process that might be appropriate for fourth graders all across the nation—and the equally broad ranges of knowledge

and curriculum that exist at any grade level, the resulting task pool was, not surprisingly, ill-suited to obtaining accurate measures of domain proficiency for individuals to aggregate as evidence for inferences about groups. Innovative statistical techniques made it possible (with considerable effort) to support group-level inferences directly. However, the shift in inferences about domain proficiency introduced a profound change in the relationship between tasks and inferences: Content-rich and process-open tasks, formerly conveying valuable information for inferences close to the content and processes they directly evidenced, were now serving as expensive but relatively uninformative sources of evidence about broad content/process domains.[13] Statistical problems caused by lower student motivation can reach beyond the affected tasks themselves. The inclusion of these tasks in aggregate scales may, to an extent, corrupt the data obtained from more process-constrained and content-lean tasks. When overall proficiency in a broadly defined domain is the target of inference, content-rich/process-open tasks must be included if they are called for in the domain definition, but they prove costly and inefficient vehicles for gathering evidence about that proficiency.

NAEP's Dilemma

There are strong reasons to include in NAEP tasks that are process-open, content-rich, or both. In terms of information flowing *from* the student population, and about what *is* happening in schools, cognitive research tells us that only such tasks provide direct evidence about the skills and knowledge that mark the competencies we want them to develop. In terms of information flowing *to* educators and the general population, and about what we *want* to happen in schools, NAEP's unique prominence for focusing national attention argues for their inclusion regardless of their evidentiary value. Yet most facets of the NAEP context thwart the success of such tasks. Spanning broad ranges of content and process, presented to students representing the full

[13] Such tasks do continue to provide the former type of evidence, inasmuch as they are still reported at the individual task level or are put in focused reports about student performance.

spectrum of instruction and background across a large and heterogeneous nation, content-rich and process-open tasks provide meager evidence, and introduce sources of misinformation about students' overall proficiency in NAEP content-area scales.

How can we resolve this dilemma? Despite current interest in overall proficiency, and the allied proportions of students at or above designated levels thereof, narrowing the NAEP content domains away from the most content-rich and process-open tasks seems undesirable in light of NAEP's historic mission of telling us about what students are accomplishing under a broad conception of the content domains. Approaches that utilize such tasks more effectively include more targeted administration and modular testing combined with separate reporting tracks for key scales and for performance components.

Targeted administration requires adaptivity in selecting which tasks to present to sampled students, in light of their responses to tasks that are more constrained as to process and lean as to content, so that they receive tasks that are more likely to be meaningful to them. This approach applies the principle of using what else is known about the student. To obtain consistent estimates for groups, however, this targeting would have to be probabilistic; decreased expectations of meaningfulness would lead to decreased chances of being presented an item, so each student might receive a majority of tasks selected adaptively for him or her, but a few selected at random. In this way, higher proportions of content-rich and/or process-open tasks could be employed, but with greater evidentiary value because of their more careful targeting.

The second possibility—modular testing and separate reporting tracks—was suggested by the Design/Feasibility Team as an option for streamlining overall proficiency reporting, while maintaining and extending reporting based on broader varieties of tasks. The idea would be to divide the task domain into two parts: one set of tasks which are more efficient for drop-in-from-the-sky administration (multiple-choice and relatively short constructed-response tasks, say), and another set of more ambitious extended constructed-response tasks that require more in-depth engagement. The former would constitute the domain on which overall proficiency indices would be

constructed and reported—an incomplete domain, to be sure, but one suited to the character of current NAEP assessment procedures. The latter tasks would be administered to smaller samples of students, perhaps in conjunction with extended time and engagement (e.g., including interviews and protocols), and possibly more complete collection of background information. Portions of the core assessment component of NAEP might also be administered to these students, so that the relationships among the skills best evidenced by both kinds of tasks would be observed. Separate reports would be issued periodically for the more open and richer tasks. This approach has several major advantages. And it is important to realize that it does not represent a reduction in the emphasis or importance placed on performance or extended tasks: rather, it gives NAEP the possibility of building assessment contexts and obtaining background information that will make the tasks far more valuable as sources of educational evidence than they are in the current NAEP.

Such an approach represents a fundamental change in how NAEP approaches reporting scales and content areas. In the past, the program's tendency has been to insist that reporting scales are based on as thorough a "content and skills portion" of a domain as is possible. When in doubt, it was routinely presumed that all items should contribute to key reporting scales. At first blush, the propensity to include the broadest possible range of exercises in key reporting scales seems entirely appropriate: The main scale that gets most public attention and is used for group and state comparisons should represent the field as broadly as possible. And NAEP frameworks have defined domains in ways not amenable to measurement only through traditional means.

However, decisions aimed at ensuring that core scales represent as broad a domain as possible have, as we have seen, had major implications that have reverberated throughout the NAEP system. While NAEP has found ways to estimate group proficiency without instruments that are individually reliable, a given NAEP assessment is used, as a whole, to produce a scale score that summarizes what students can do in a given domain. This means that, in a sense, the assessment must be reliable as a whole. In this context, the decision to include complex performance and extended-

answer tasks in overall reporting scales has had immediate implications. Complex performance tasks are time consuming and do not tend to cover broad expanses of content knowledge. This means that the aggregate testing time needed to cover a domain at a given grade will tend to expand markedly in an assessment that plans to use performance testing as part of its reporting scale. This proclivity is particularly clear in "content-heavy" disciplines such as science.

In addition, once one assumes that the results of complex performance tasks should contribute to reporting scales, one may introduce pressures that increase the need for expanded assessments. If one believes that the central scale score for science should reflect a vision of the field in which laboratory-type work plays a prominent part, one does not want a single, possibly idiosyncratic task to carry the entire load of representing the domain.[14] This meant that in the 1996 NAEP science assessment one could not use one or two experiments as examples of the sorts of laboratory work students might do; rather, four independent experiments per grade were developed to ensure content coverage.

In the end, these pressures had a substantial impact. The 1996 science assessment at grade 12 was composed of seven and one-half hours of testing (450 minutes). Even though individual student testing burden was raised to a NAEP high of 90 minutes, individual students took only 20 percent of the aggregate assessment. This leads to other complexities, including the extensive reliance on model assumptions in conducting analyses.

Finally, the seemingly simple decision to include performance and extended constructed-response questions extensively in reporting scales means that these tasks must be used in all environments in which scale scores will be used for comparison. Specifically, this has meant, in NAEP, that performance and extended-response tasks became components of state NAEP, which lead to almost geometric increases in cost.

[14] If one intends to report data at the task level, the pressure for domain coverage may be less. However, idiosyncratic tasks are likely to be a problem at either level of reporting. And this is something of a general issue in performance testing, where tasks may be either too constrained (often for purposes of standardization) to reflect "real" thinking in the field, or "process-open" to the point that groups of students fail to understand the directions implicit or explicit in the task.

Under a modular approach, one would assume that some sorts of exercises are of greater evidentiary value when contributing to a scale, while the most appropriate use of other items lies in focused reports and analyses and explicitly not in being scaled with all other observations. Under this approach, we believe, lies the greatest hope for a NAEP that can both produce fast and inexpensive national comparisons and at the same time use its unique position to delve into the learning practices of students across the country.

The approach that yields the most appropriate data should only be adopted if it is operationally and financially feasible as well. Fortunately, modularity also has real advantages in terms of schedule and cost. It is to these advantages that we now turn.

The Mix of Item Types, Modularity, Cost, and Schedule

Under the current NAEP structure, it is assumed, to a great extent, that all participants will receive all types of items in the assessment and that the reporting scale should, to the extent possible, reflect all outcomes described in the assessment *Framework*. Thus, in the 1996 science assessment, all students received hands-on tasks. Because these tasks were being used to contribute to a domain score, multiple tasks were developed and used. In addition, theme blocks, that is blocks devoted wholly to a series of questions on and an in-depth examination of a single topic, while not given to every student, were included in the main scale. These blocks were more heavily constructed-response than the more general science blocks, and their inclusion increased the scoring effort.

The implications of these decisions from a schedule, cost, and complexity viewpoint are fairly obvious. Analytic problems related to these exercise types (including local interdependence and dimensionality issues) must be handled on the critical analysis path. Constructed-response scoring volume (almost five million responses) slowed the production of final reports. Scoring, even using two shifts, took over 12 weeks. In addition, new item types and the scoring of partial-credit items necessitated statistical intervention during the scaling process.

The cost factors are even more striking. Science kits for experiments cost roughly $6.00 per kit. Given standard overages (that is, amounts purchased to cover for breakage, absent students, and other contingencies) kits alone for a national sample cost roughly $100,000 per grade. While this expense may seem acceptable in a national sample, buying kits for the state sample cost roughly $1,000,000. And remember, once the decision was made to include the hands-on tasks in the core scale, they became a necessary component of the state assessment.

None of this is to argue that either the hands-on tasks or the extended-answer questions do not yield valuable information—in fact, they do. But the most valuable information they yield is not, as we have seen, necessarily in their ability to contribute to a scale. It may be that within the context of these sorts of exercises the most useful information is that closest to the exercises themselves—that is, information on what students were able to do on that task. Clearly, we can get this task-level information from the current NAEP, but at major expense (and at an expense that increases if assessments are administered at the state level). And we do not get much of the contextual information that would make the item-level data more meaningful (and such contextual data would be easier to gather for a small sample involved in a module than for the entire sample). So at core the question is, how much is it worth to have these tasks contributing to the reporting scale?

While expense is and should be a factor in these decisions, there has been little systematic thinking about the cost implications of different models. But clearly, cost must be one of the components considered in a systemic attempt to meet NAGB's goals. We have therefore chosen to analyze the cost implications of modular NAEP versus the current NAEP within the area of cognitive testing.

Our analyses have shown that cost reductions related to modular approaches are substantial. However, since these analyses contain confidential budget figures they are included in a separate secure document. The document shows that modular testing not only makes sense from an evidentiary point of view, it makes solid budgetary sense as

well. We therefore recommend that modular versions of cognitive assessment be adopted wherever feasible.

Using New Technologies in Testing

The dual perspectives we have developed—evidentiary and fiscal—provide an appropriate lens through which to evaluate the various ways in which new computer-based testing technologies might be introduced into the NAEP system. In addition, we must view the use of computer-based testing (CBT) in NAEP in terms of equity and feasibility.

Consistent with the discussion of performance testing above, computer-based testing might be used to draw inferences about **broad domains of knowledge** or about **specific skills and contexts**. In the former case, one would examine CBT options to locate ways to use computers to produce scale scores. In the latter, one might view the computer as a way to assess educational outcomes either prohibitively expensive or technically impossible to measure in a pencil-and-paper environment. One also might, in the second model, view CBT as a tool to be used to save money in cases where such savings are possible (for example, in scoring certain types of tasks).

So there are two possible models of CBT for NAEP. One might be designed to generate domain scale scores and "replace" existing instruments. The second might be designed to assess skills not amenable to current measurement methodologies and "augment" the current assessments. These are not mutually exclusive options—for example, one might use innovative measures in an assessment designed to produce a scale score—but they will represent ends of a continuum in the discussion that follows. It is to a consideration of these CBT models that we now turn.

Computerized Adaptive Testing Designed to Produce NAEP Scale Scores

The first model—in which CBT becomes a way to generate scale scores that represent broad domains of knowledge and are NAEP's main reporting statistics—pushes one to consider computerized adaptive testing (CAT). On the surface, CAT

seems to hold great promise for NAEP. By tailoring tests to the ability levels of individual examinees, CAT allows, in some situations, reliable estimation of individual scale scores from instruments that use very few items. This would, in principle, allow NAEP to continue to meet its goal of limiting individual examinee burden while at the same time basing group scores on reliable individual scores—a change that would greatly simplify NAEP analysis. It would also greatly simplify measurement at the edges of the population, which are often regions in which achievement levels are set.

However, early experience with CAT indicates that these goals would be difficult to meet. In fact, upon closer examination the possible savings of testing time and reductions in complexity are likely to prove illusory. And there are other costs to CAT that must be considered as well.

Test Length. One of the major promises of CAT was its potential to reduce testing burden by estimating individual scores through use of a small number of exercises. One can find examples of extraordinary shortening of a test through the use of adaptive testing—of individual tests reduced to three or four items, for example. However, realizable savings in tests that meet professional standards are great but not nearly so dramatic. Reductions in test length are limited not only by the need to assure that an individual score will be reliable, but by two additional factors: the need to assure that each test administered meets appropriate content specifications (even NAEP is likely to have some effective content minimum), and, for the sake of test security, the need to limit the rate of use of items.

An example of the impact of these constraints can be drawn from a comparison of the computerized GRE General Test with the paper version. Test content and level are of course far distant from what would be involved in computerized assessment for NAEP, but the comparison is almost surely of the right order of magnitude. The computerized version of these tests is required to meet all the many constraints on the balance of item formats and contents that the paper test must meet. In addition, sophisticated methods are used to control the frequency of use of individual items (a

modification of the Sympson and Hetter[15] methodology at present; in the future likely a still more demanding method developed by Stocking and Lewis[16] that conditions exposure control on ability level). With these constraints, the greatest reduction in test length is achieved for the GRE Verbal measure—a reduction from 76 items requiring 60 minutes in paper testing, to 30 items requiring 30 minutes in the computerized test. Thus, a 50 percent savings in time is achieved for this test section. However, other parts of the test afford smaller savings. For GRE Quantitative, the reduction is from 60 items requiring 60 minutes, to 28 items requiring 45 minutes; and for the third measure, Analytic, the test goes from 50 items in 60 minutes to 35 items in 60 minutes—no savings in time at all, though a less speeded and therefore potentially more valid measurement is obtained.

Test Tutorials. A further limitation on savings through computerized testing is that examinees must learn to use the computer to take a test. It is possible to set up a computerized test in such a way that very little training or practice is required to use the system; ETS did so successfully, more than 10 years ago, with a computer-based battery of tests of basic skills used in placing students at the time of their entry into college.[17] Students spent no more than several minutes in learning how to take the test; the test worked well with, and was viewed very positively by, students who frequently had little computer experience and minimal academic skills. However, a very important proviso: the tests in this battery used only simple multiple-choice questions, so that examinees needed to understand the function of only three keys in order to navigate through the test. We would not so constrain a test we were developing today. We would use a greater variety of item formats, in order to take advantage of the flexibility offered by computerization. In addition, we would offer other features designed to enhance the examinee's comfort and ease in taking the test—providing an

---

[15] Sympson, M.L., & Hetter, R.D. (1995, October). Controlling item-exposure rates in computer adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association.* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

[16] Stocking, M.L., & Lewis, C. (1995). *Controlling item exposure conditional on ability in computerized adaptive testing* (ETS Research Report RR-95-24). Princeton, NJ: Educational Testing Service.

[17] Ward, W.C. (1988). The College Board Computerized Placement Tests: An application of computerized adaptive testing. *Machine-Mediated Learning,* 2, 271-282.

indication of the time remaining to complete a test section, for example, and also providing the examinee a means by which to hide the clock if seeing the time proved distracting for that individual.

As a consequence of such improvements, taking the test by computer is not quite as simple as in that early example, and examinees must spend time being trained and then practicing. It is reasonable to project that, for a moderately complex computerized test, examinees would require a tutorial that would average around 20 minutes per person. Moreover, fairness to those who have little computer experience probably dictates that the tutorial be untimed and individuals be allowed to spend as long as they need working through it; but under such conditions, at least a small percent of individuals would take as long as twice the average time.

It follows that, if an individual's involvement in testing were to consist of a single session that, in paper, might have been one hour in length, there may well be no average savings in time associated with adaptive administration. The time required for a tutorial before beginning the test could easily consume all the time saved through the adaptive assessment. Total time involvement would also be more variable, introducing some scheduling issues, and for some individuals would be greater than that for paper testing.

Item Pools. The item pool from which an adaptive test is drawn is often the equivalent of as many as six or eight paper test forms in length. There is no simple rule to determine how many items are needed; factors influencing this include the psychometric quality and appropriateness of the individual items available (e.g., whether items are high in discriminating able from less able individuals, whether the range of difficulties is appropriate to the range of individuals to be tested), the number and complexity of content and format constraints that must be balanced within the test, and the importance of enhancing test security by preventing overuse of the more frequently selected items.

The test security concern is intimately tied into the purpose and use of the tests. So long as the use is seen as low stakes, without individual scores or consequences for

administrative units over which performance might be aggregated, tight control over the frequency of use of items may not be important (although the development of national tests linked to NAEP may make this more of an issue than it has been in the past). If higher stakes uses or derivatives of the tests are contemplated, however, test security becomes a paramount issue. Here the concern is not specifically with adaptive testing, but rather with continuous testing. The fact that testing must take place over an extended period of time introduces the possibility that one examinee could share information with another who has not yet been tested. In such a case, adequate security may demand a multifold increase in the number of items available as well as a schedule for periodically (and frequently) introducing new test pools. Items can be recycled from one pool to another, so long as care is taken to limit the overlap of any two pools in the number of items they have in common. These are procedures that are routinely carried out in ETS delivery of secure high stakes tests; they do not pose difficulty but add some cost and complexity to the enterprise.

Limitations on Items in the Pools. CAT assessments are based on the assumption that algorithms can select items to present to individuals. To meet this assumption the items in an adaptive stream must all be machine scorable. While great advances have been made in automated scoring of some types of performance tasks, scoring systems cannot yet cope with all the types of exercises used in NAEP assessments. The limitations of these scoring systems are not limited to extended-answer questions: the ability of systems to effectively rate natural-language responses is still limited. So frameworks would need to be revised to make use of CAT. And it is not clear that curriculum specialists will be willing to live with any limitations imposed by machine scorability.

Operational Pretesting. CAT assessments presume that one has full knowledge of the statistical qualities of all exercises before they are placed in an item pool. This means that operational pretesting of a sort not previously conducted in NAEP would be necessary. Such pretesting—which has been advocated by NAGB for other reasons—would require representative samples of students so that item parameters

could be reliably calculated. Such field tests would have to be large enough to allow any items in need of revision to be deleted from the operational assessment pool. In practice, this will add substantial expense to the program. It will also call for lead times far beyond those normally available in NAEP development cycles.

Context and Position Effects. NAEP has, in the past, gone to great lengths to balance its assessment designs to minimize, to the extent possible, fatigue and position effects. In a CAT, where the system is selecting items, such balancing may not be possible. In operational CAT programs where individual scores are reported, such as GRE, this does not pose much of a problem: the goal is to place an individual on a scale, and context effects are likely to "wash out." However, in NAEP the issue may prove more serious. One of the goals of NAEP has been to report on what students know and can do in a domain. Therefore, to the extent that our views of what students know and can do are colored by context or fatigue effects, they will be incorrect. It may be possible to control some of these effects in a CAT, but that would involve further research.

Feasibility and Equity. Operational feasibility and equity will be issues in any use of computer-based testing (CBT), whether that use be CATs designed to replace existing assessments or computerized modules designed to augment them.[18] However, these issues will be especially salient in the former case.

If CAT assessments were used to replace existing instruments, these assessments would necessarily need to be administered in the schools.[19] And while CATs reduce the number of items any student receives, the use of such tests would not, in and of itself, exert any downward pressure on student sample sizes. Related to this, it seems imprudent to rely on equipment in schools for computerized administration. There is great variability in the quality and quantity of computers available in schools; some

[18] Concrete recommendations for the study of the feasibility and equity issues associated with CBT will be made at the end of this chapter.

[19] This might not be the case with special pilot modules involving small numbers of students. In such small-scale settings, it may be possible to bring students to testing centers in which appropriate equipment is available. This approach has been taken by the Johns Hopkins talent search program. However, this delivery system makes little sense as a long-term vehicle for general delivery of the assessment.

schools do not have computers capable of running assessment delivery software. In addition, even among schools whose equipment met some minimum standard, platform variability would likely introduce large degrees of differences between local testing situations. This, in turn, would lead to problems of standardization. In addition, the ability of NAEP administrators to gain access to school computers may be limited. In the long term, reliance on equipment in schools may become a viable option; in the short term it is not.

Given this conundrum—the need to test in schools and the nonavailability of computers in schools—administration of a NAEP CAT would seem to involve bringing equipment to schools. Such administration would introduce real issues of cost. While this cost might be limited by selection of a platform with somewhat limited capacities (which would lead to other problems), it is likely to be high. This cost would go up geometrically in any case in which a state assessment is offered. In such a situation, NAEP would have to consider fundamentally changing its data collection philosophy, possibly to one in which data are gathered over an entire school year to maximize the efficient use of computer equipment.

We should mention here that even the costs of a modular use of CBT are likely to be nontrivial. However, given the notion that all NAEP students may not, in such a setting, need to be tested on computer, costs are not as likely to prove prohibitive as in the CAT case.

As with feasibility, equity will be an issue in any use of CBT. Equity problems are likely to be more severe in a CAT designed to produce domain scale scores than in a CBT module designed to measure particular outcomes.

Students have differential access to computer technology, and it is not immediately clear what impact a change to CBT would have on the comparisons of population subgroups that are such a visible feature of NAEP. It is not clear that one could generalize on what has been learned about CBT and equity in programs such as the Graduate Management Admissions Test (GMAT), the Graduate Record Examination (GRE), and the Praxis Series Teachers Examinations: those programs

involve self-selected populations all having high levels of education, while NAEP involves heterogeneous national populations. Because of this, it seems wise to examine these issues in limited settings—such as modular or special studies—before deciding whether or not CAT is appropriate for NAEP in the intermediate term. Additionally, in such limited settings it is likely to be possible to gather contextual information needed to meaningfully interpret the CBT performance of different groups.

Before leaving the discussion of CAT, we should mention that there may well be alternatives to fully adaptive assessments that make use of adaptive technology. For example, adaptive instruments might be used as locator tests. Once students finished these tests they could be given a performance or constructed-response module selected to be appropriate to their level of ability. However, while options such as this may solve some of the item-type limitations inherent in the need for machine scorability, they still involve the expense and logistical problems inherent in a general or universal use of adaptive testing in NAEP.

CAT may, in the long term, be a viable step for NAEP. Yet, uncertainties, limitations of item-types, and expense do not likely make it a viable short term option. It seems wise to examine the other end of our continuum, and look at modular uses of CBT designed to expand the range of outcomes measured by NAEP.

Computer-Based Testing Designed to Assess Skills Not Amenable to Pencil-and-Paper Testing or to Introduce Efficiencies

As was mentioned above, there have been two main impulses behind the development of CBT: the desire for adaptive measures and the desire to expand the range of outcomes that could be measured. These are not necessarily contradictory impulses: measures of new skills can be adaptive and amenable to machine scorability. However, in the case of NAEP, if one accepts that the replacement of pencil-and-paper NAEP instruments with CATs is, in the short term, an unreachable goal, then the second impulse may well prove more promising as an area for initial work.

The measurement of new domains is a real possibility for CBT. Interactive CBT modules covering extended mathematics problem solving, or other such areas, can be developed. These might make use of adaptivity or have program features that give students hints, and would allow measurement in a way not now possible.

Related to these sorts of tasks, some measures that NAEP has developed for special assessments might be done in a more interesting fashion on computer. For example, John Frederiksen[20] has developed a series of simulated laboratory experiments which allow for measurement of the types of outcomes that NAEP measured through the hands-on science tasks. Because his experiments are simulated, students may be given the freedom to control more experimental parameters than is possible in a hands-on setting. In addition, this is one situation in which the cost of CBT may compare favorably with the cost of purchasing laboratory kits for participating students.

Computers also allow for the presentation of stimulus that is impossible in pencil-and-paper settings. Multimedia presentations could be used as part of history assessments. And unlike use of videotape, students taking CBTs could stop and replay video clips. On a geography or science assessment, students might be presented with atmospheric models and asked questions about these.

Related to this, CBTs can be used to measure the research skills of students in a way impossible in standard testing situations. Assessments are often criticized for relying too much on memorization; in a CBT module it would be possible to give students access to encyclopedias, atlases, dictionaries, or other reference materials.

Finally, computers allow test developers to give students tasks that call for tools not available to pencil-and-paper test takers. For example, spreadsheets or calculators can be made available.

---

[20] Frederiksen, J. R, & White, B. Y. (1997, May). *A computer enhanced curriculum for teaching and assessing scientific Inquiry* (Preliminary proposal to Instructional Materials Development Program, National Science Foundation).

In summary, CBT does seem to have real possibilities for the expansion of the range of outcomes that can be measured by NAEP. However, it is important to understand that cost will be an issue in this area as well. Specialized modules are expensive to develop, and delivery systems must be built. Administrative costs will be high.

Yet making a decision to focus on modular uses of CBT, at least in the next few years, has benefits that outweigh these costs. NAEP would be shortsighted to do nothing to prepare for a future that is surely coming: it is essential that the National Assessment gain experience with CBT. But there are so many questions and so much development work that the wisest approach would be to make a staged move to CBT to allow time to consider the potential benefits and pitfalls of computerization. We strongly recommend that NAEP follow this measured approach, and learn as much as possible from the advances and errors made in other programs.

Designing a CBT Delivery System for NAEP

As discussed above, one of the key problems in considering CBT implementation is that NAEP administrators cannot rely on computer equipment that is in the schools. Therefore, we must work under the assumption that NAEP must supply equipment to participating schools. There are several plans that might be implemented, and the costs and benefits of all such plans must be evaluated before work on delivery systems begins.

The management of computer-based testing in NAEP will require careful planning. Questions such as what machines will be used, where they will be located, who will provide technical support, and who will monitor testing all must be answered in a way that considers quality and standardization of the administration, test security and security/integrity of the equipment, and cost.

It seems reasonable to propose that all computer-based testing in NAEP should be set up and monitored by (or under the direction of) a central contractor, and that standard, centrally controlled and maintained equipment should routinely be brought

to the school for use in testing. However, the precise nature of the equipment to be used will profoundly affect the types of instruments that can be offered, and the costs of the program. For example, there are some relatively inexpensive technologies that might be used to deliver assessments in NAEP. An example is Brainchild, a $200 (retail) hand-held computer for which a range of supplementary instructional packages are available. Developers might contract with manufacturers and arrange to build assessments that run within their standard format. From the perspective of providing some delivery of assessments by computer, this might prove the least expensive solution. However, it would restrict assessment severely in terms of the complexity of material that could be presented and the sophistication of the test model that could be used. Thus, it is likely to be a desirable solution only if computerized assessment were seen as a goal in its own right but one that must be achieved as inexpensively as is possible.

Other equipment options are more in the spirit of using the new possibilities afforded by CBT, but are also more expensive. One might bring into the schools fully capable laptops loaded with test items and test models that are as complex and innovative as we care to make them. A $2000 (retail) machine could provide displays, speed, and memory capability that are more than adequate for any material we would want to employ, eliminating any machine limitations on what we could present. (A machine in this price range might have just adequate display capabilities today; but it is safe to assume that whis will not be a limitation in several years.) The major drawbacks are the expense of providing such machines, along with the likelihood that theft and breakage would be nontrivial problems. This solution is possible if one envisions limited uses of CBT in NAEP; it makes little sense if one wants to replace existing NAEP instruments with CATs.

Finally, more radical delivery solutions may be needed. One might consider developing mobile testing centers. Such centers might be busses with fixed workstations which would be driven to testing sites; groups of 8 to10 students could then be tested within the bus parked at the school. Setup and space problems would be

eliminated, and security and breakage problems greatly reduced. However, the costs of purchase and design are likely to be prohibitive.

At some point, but likely not in the initial years of the next NAEP agreement, reliance on equipment already available in the schools will become reasonable. When that point is reached, the use of the Internet as the means of inexpensive, machine-independent delivery of assessments to the school building will be desirable. Then many of the short-term delivery issues that would currently face NAEP might have gone away.

Recommendations on CBT and NAEP

In line with the above discussions, we believe that NAEP should, within the next two years, follow a three-part strategy regarding CBT. We recommend that as a first step NAEP develop an explicit set of criteria for evaluating the potential uses of computerized testing in NAEP and then use these to identify a small number of promising opportunities. Criteria would include such considerations as enhancing measurement beyond what would be feasible with paper-and-pencil, making manageable demands on students (i.e., requiring only minimal tutorials before the examinee could use the computer for the task), providing tasks that could be completed in a single session, having reasonable exercise development costs, and so on. Mathematics tasks requiring some kind of extended problem solving, science tasks involving simulated laboratory experiments, and social science or history tasks requiring creation of a report drawing on source documents are possibilities that seem promising.

Following the establishment of the evaluative criteria, we recommend that NAEP commit to use one CBT pilot module with either the science or mathematics assessment by the year 2000. This model should be experimental, and would likely involve the use of response protocols with students. The results of this study would give NAEP valuable information on equity and feasibility issues, and would help point the direction for future efforts. We strongly believe that the initial uses of CBT in NAEP

should be to measure outcomes not easily accessed through pencil-and-paper testing, and that the program's short-term goal should not be to use CATs to generate the domain scale scores used in primary NAEP reporting.

Finally, we recommend that, as the development efforts are beginning, NAEP immediately begin studies of how CBTs might be administered within the constraints of this survey. These studies should consider cost, robustness, and the constraints imposed be different delivery platforms.

Cognitive Instrumentation: Areas in Which Current Practices Affect the System

Current cognitive testing practices influence the remainder of the NAEP system in a number of ways. Perhaps the greatest of these is the varied ramifications of extensive use of performance and extended constructed-response measurement. Because factors related to such testing—including evidentiary, schedule, analytic, and cost issues—have been discussed in depth in this chapter, we will not repeat the discussion here. Suffice it to say that these items add great cost, introduce complexity into analysis and trend measurement that fall on the critical reporting path, and are often not administered in a context that maximizes their usefulness.

There are other ways in which cognitive testing practices affect the remainder of the NAEP system. These include limits on student testing time, use of a BIB-spiral design, and use of field testing. These aspects are discussed in the following sections. Any redesign of the NAEP system should at least acknowledge these factors and determine whether or not change is possible or desirable.

Limitation of Student Testing Time

NAEP has long operated under the assumption that the time needed for cognitive assessment should not exceed one hour. While this limit has been exceeded in certain assessments (science and arts), on the whole NAEP has struggled to avoid expansion of the testing window.

The reasons why NAEP has striven to limit testing time are severalfold. NAEP is a voluntary program, and school and student participation will likely be affected if testing burden is increased. Past assessments have also provided ample evidence of fatigue effects;[21] asking students—especially young students—to write or perform tasks for extended periods will not provide an accurate representation of what they know and can do. In addition, extending the testing period is likely to intensify motivational problems that, to some extent or other, are part of NAEP.

Therefore, the reasons to limit testing time are good, and testing time should, in all likelihood, not be substantially expanded within the context of the current program. Yet, reviewers should be aware of the implications of limited testing windows. Specifically, no students can take the entire assessment, and no students take tests that are designed to be parallel or individually reliable.

NAEP analysts have found ways to work within these constraints, but the limitation of testing time has led to complex administration and analysis techniques. These techniques do add time and complexity to the system.

There is also a way in which use of performance testing intensifies problems caused by the limitation of student testing time. Constructed-response testing drives down the number of exercises any student receives. It also leads to increases in the aggregate size of an assessment necessary to cover a given content domain. These phenomena in turn lead analysts to rely more on measurement models and background variables in the calculation of proficiency estimates, and less on direct observation of student performance.

In some, testing time does have impact throughout the system, but the cost of increasing testing time likely outweighs the benefits.

We should mention in closing that modularity should help to limit testing time. In the core components of an assessment from which domain scales would be generated, individuals would take more questions. This would reduce the burden on

[21] See, for example, the *1988 NAEP Technical Report*, in which position effects in writing assessment were evaluated.

model assumptions, but would probably not eliminate the need for sophisticated analysis techniques.

Use of BIB Spiraling

NAEP assessment design is influenced by short individual student testing time, the length of aggregate assessments, the need to balance the position of blocks, and the need to pair each block with several other blocks in the assessment. In the 1994 history assessment, for example, each block was paired with every other block and appeared an equal number of times in both the first and second position in a booklet. This meant that the eight grade 12 blocks were spiraled in 32 discrete booklets.

This spiraling plan adds cost to the production of the assessment, and requires that careful planning precede analysis. It also involves substantial setup and programming before constructed-response scoring can take place. However, during the initial year of an assessment, such spiraling is necessary if NAEP is to draw appropriate inferences about student performance.

One area of potential streamlining centers around the question of whether or not the entire BIB must be used in subsequent administrations of an assessment. Researching the impact of not using the entire BIB is beyond the scope of this redesign project; however, we recommend that reducing the size of the BIB be investigated as part of ongoing NAEP research efforts.

Uses of Field Testing

The core purpose of the NAEP field tests is to try out new assessment items in order to determine whether or not they "work" statistically. Ultimately, the goal is to emerge with operational assessments that achieve optimum content coverage while also meeting basic statistical conditions and standards.

Accordingly, professional test developers review a detailed set of statistics for each item administered in the field test, including biserial and polyserial correlations, difficulty levels, omission rates, scoring reliabilities, and speededness rates. When the

statistics indicate that an item is flawed, the test developers use their professional judgement to determine whether or not the problem can be corrected. It is important to emphasize that item revisions are made only when statistics indicate that such changes are necessary.

The practice of revising items after field testing does have some associated costs. For example, program analysts must make statistical adjustments to certain items (such as collapsing score categories), which extends the analysis schedule. Yet the practice of revising items is wholly consistent with the goals of field testing and is economically efficient.

If the goals of NAEP field testing were either to establish "reportable" statistics on items or develop individually scorable tests that meet rigid statistical specifications, it would not be advisable to revise items after testing. In addition, if the primary aim of field testing was to allow for "pre-operational analysis" (including achievement-level setting), item revision would also be harmful. But because the NAEP field testing is not intended to meet these other purposes, reasoned revision of items is acceptable.

It is possible to conceive of a NAEP in which "operational pretests" are designed to allow for assembly of parallel forms or for "pre-analysis" of data. In the case of a parallel-forms model of NAEP (that is, a model in which NAEP is a series of "true-score" tests rather than a group-score assessment), such pretesting would be necessary. In addition, NAGB has suggested that this form of testing be built into the NAEP system to allow for achievement-level setting and faster reporting. However, if pretests designed to support strong statistical inferences were programmed into NAEP, several other aspects of the program would have to change. First, far longer work schedules would be needed. Currently, assessment frameworks are available as little as 20 months before an operational assessment. Thus there is not sufficient time to conduct both a field test to try out items *and* an operational pretest. If the practice of administering field tests at the same time of year as the operational assessments were continued, the addition of an operational field test would add a year to the development cycle.

Meeting these statistical goals would also necessitate field testing a far larger number of items in a more elaborate and expensive fashion. It would specifically require a field test with a BIB design, and operational scoring and analysis. In addition, if item developers were unable to revise exercises, they would need to pretest a large enough number to allow for failure of items. This number would increase markedly if the goal of field testing was the assembly of parallel forms. Currently, NAEP field tests are limited in scope: between 150 percent and 200 percent of the exercises needed in an operational assessment are field tested.[22] NAEP does not field test more items for a variety of reasons. First, program managers have been wary of increasing the expense of an already costly program. Both item development and field testing cost money. Second, the schedule issues referred to above provide practical limitations on the number of items that can be developed.

Not only are NAEP field tests limited in scope; they also often have an "experimental" component. NAGB frameworks have tended to call on NAEP to stay on the cutting edge of measurement. Item types such as open-ended exercises in history at grade 4 (a subject not specifically taught at that level), performance tasks in science and the arts, and map-creation questions in geography have exemplified this pressure to innovate. It is unlikely that these item-types could have been used in a large-scale setting without the ability to revise field-test items.

Clearly, *adding* an operational pretest to the existing development scheme would allow for revision of items after initial field testing *and* make it possible to preset achievement levels and precalibrate item statistics. However, such an approach would add substantial expense to the NAEP program. Whether or not the operational pretests called for by NAGB are financially feasible will necessarily be a policy rather than a technical decision.

[22] By comparison, the SAT program, whose goal is the assembly of parallel forms, pretests seven times as many items as it needs in any given year.

Local Optimizations That Will Profit Any Model of NAEP

Many "local" optimizations of cognitive testing that will profit any NAEP model have been discussed in other parts of this chapter and report, so is unnecessary to discuss them extensively here. However, we will remind the reader of several such optimizations—specifically those which save time or money. There are negative impacts as well, however, and these are specified. Modular designs are not discussed in this section because they are one of the three main models discussed both in Chapter 3 and in the recommendation section that concludes this chapter.

Reduce the Percentage of Items Released. Reducing item release will result in lower costs for assessment development, scoring, and analysis. This will profit any model for NAEP. The cost is that the public will have fewer items to use for linking and other purposes.

Reduce the Proportion of Core Assessments Devoted to Performance Testing. By core assessments, we mean any component of testing used to produce scale scores for central reporting and group and state comparisons. This optimization will save substantial time and money. It will also limit the use of performance measures in components of assessments in which we have access to little of the corollary information needed to give meaning to the results of these items.

There are several negative ramifications of a reduction in the percentage of open-ended items in core assessment components. Scales will incompletely reflect broad domains, and be more influenced by multiple-choice and short-answer results. Such a change might also be viewed negatively by content specialists. Also, if such an optimization is not used in conjunction with a modular approach, it may lead to a reduction in the volume of interesting student responses that can be featured in NAEP reports.

For these reasons, we strongly believe that the best strategy for NAEP is to modularize cognitive testing. This will allow reduction of performance exercises in the high-cost, critical-path, core components of the assessment, while maintaining their

high-profile use in the overall assessment system. It will also allow for the design of assessment contexts and the gathering of supporting information that will make the results of performance and extended-response exercises more meaningful.

Cognitive Testing Under Different NAEP Models

Cognitive testing would be organized differently under the three different NAEP testing models. A summary of these organizations is given below.

Cognitive Testing in a Streamlined NAEP

Under a streamlined NAEP, the cost and complexity of field testing could be reduced if the release of items was limited. Streamlined field testing models could be implemented. Operational pretests might be desired, but would likely not be necessary. A streamlined NAEP would likely make less use than the current NAEP of performance exercises and extended constructed-response testing. Sample sizes might be reduced , but with the possible loss of reporting variables and the power to make meaningful comparisons. The number of distinct assessment sessions might be reduced.

However, there would be negative implications for cognitive measurement as well. There would be no easy or standard way to introduce new content into the assessment. Standard assessments would provide a good deal less information than comprehensive assessments. And in this model, the choice between having comprehensive scales and reducing testing costs remains stark: If they are not included in the scales, performance items are likely to be excluded from the standard assessment. There would be no systematic attempt to place performance studies in situations in which meaningful supporting information could not be gathered. Finally, cost savings would be limited by the fact that national and state components would likely still need to include some large number of performance exercises.

Cognitive Testing in a Modular NAEP

A modular NAEP provides an easy mechanism for controlling the release of items: items in core components might not be released after the initial comprehensive assessment. This would lead to substantial reduction in field testing cost and complexity. Because modules might not all be on critical reporting paths, field testing more limited than that currently used in NAEP might be employed. Operational pretests might be desired, but would likely not be necessary. The core components of a modular NAEP would make less use than the current NAEP of performance exercises and extended constructed-response testing. These would be embedded in special modules given at either the national level only or to limited local samples. In most cases these would be excluded from state assessments, resulting in great cost savings.

Sample sizes might be reduced (with possible loss of reporting variables and the power to make meaningful comparisons) for core components of the assessment. However, the cost reduction in the core assessment might make this unnecessary. Performance and constructed-response modules would be given to smaller samples; these samples might be large enough to be nationally representative and to allow subgroup comparisons, or not, depending on the purpose and nature of the module.

The main negative associated with modularity is that reporting scales would no longer reflect the entire range of the framework. However, modularity would allow NAEP to do a more thorough and complete job of administering, analyzing, and reporting the performance and extended constructed-response portions of the assessment.

Cognitive Testing in a Parallel-Forms NAEP

Of all the NAEP models discussed in this document, the model in which NAEP is composed of parallel, fixed forms and in which true-score type equating is used to produce estimates would have the greatest impact on cognitive testing. This model would likely lead to substantial changes in the way NAEP conducts cognitive measurement.

To assemble parallel forms, both field testing (to evaluate the functioning of items) and operational pretesting (to evaluate the parallelism of forms) would be required. The latter is, to some degree, optional under the other two assessment models. While both a streamlined NAEP and a modular NAEP would allow some use of short and extended constructed-response testing, the search for parallel forms would probably lead these forms of testing to be sharply constrained. This is because parallel statistical characteristics would be most easily achieved through instruments relying largely on multiple-choice and short answer questions.

Perhaps the most important implication for cognitive testing in NAEP would be some necessary limitation of content coverage. In general, because the goal of the program has been to report what students know and can do, it has seemed appropriate to cover as broad a range of content and skills as is possible. This has allowed NAEP to report how students are doing in certain subcomponents of disciplines.

If forms are to be "parallel," however, then parallelism must be manifest with respect to statistics and content. Clearly, there will always be some minor variation in content among forms even though developers try to hold such variation to a minimum. Since the length of any test will be limited, the amount of aggregate content coverage possible in an assessment may also be constrained. This would lead, in the view of Robert Linn, an external advisor to this project, to a major diminution in the importance and acceptance of NAEP.

In addition, individual forms of limited length might not be able to effectively produce subscale data, unless sophisticated analysis techniques were used. To combat these problems, pressure may arise to increase cognitive testing time per student. However, as we have seen, such an increase would lead to various other serious problems.

There are, of course, positive sides to the parallel-forms approach as well. Analysis and reporting might be eased and streamlined. However, we believe that the negatives far outweigh the positives.

Before leaving this topic, we should mention that the problems listed above are most severe if one attempts to "replace" NAEP with a system of parallel tests. Use of market-basket reporting metrics, or short linking forms, will not cause problems as substantial.

Recommendations for Cognitive Testing

The recommendations on cognitive testing are scattered throughout this chapter. For summary purposes, let us restate them here.

Use Modular Testing Designs. Such designs will allow NAEP to meet goals that have been difficult to achieve in the past. On the one hand, modular testing will allow NAEP to save money and to produce core reports in shorter periods of time than are possible under the current system. Savings from modular approaches will be especially great in programs administered at the state level. In addition, modular designs will enable NAEP to garner more meaningful data from performance and extended constructed-response testing, by allowing the program to set appropriate testing situations and gather contextual data in an economically and operationally feasible context.

We strongly recommend that modular approaches be considered as new frameworks are written. That means that framework developers should explicitly consider which aspects of a domain are amenable to mass measurement and inclusion in reporting scores, and which can only be measured in ways that make them appropriate for modules. We further recommend that NCES consider rebuilding certain high-cost current assessments—such as the science assessment—to fit modular structures.

Use A Staged, Modular Approach To Computer-Based Testing (CBT). In the short term, we recommend against replacing existing core NAEP instruments with computer-adaptive tests. We rather suggest the following three-part strategy:

- Develop a set of criteria to evaluate potential uses of CBT, and conduct a study of possible CBT delivery systems.

- In the short term, focus on CBT applications that will allow NAEP to measure outcomes not possible in a pencil-and-paper setting or that introduce efficiencies.

- Include one operational CBT module in NAEP by the year 2000, and use this module to study equity and feasibility issues.

Limit Item Release. This will reduce the need for expensive field testing and instrument development.

Gather Data in Ways and Contexts Appropriate to the Inferences One Wishes to Draw. Item types should not be chosen simply because they "send the right message," nor even because they "measure the right skills." Rather, they should be chosen for components of NAEP based at least partly on their appropriateness to those components. Only those exercises that are amenable to large-scale assessments should be included in core components.

On the other hand, some exercises require special testing situations (that is, special motivation or information given to students before testing) if the results are to be interpreted meaningfully. Such meaningful interpretation also often requires that we know more about students than we can learn in a large-scale setting. For these types of exercises—which will likely remain a valuable part of the NAEP landscape—modular approaches will allow stronger, more robust analysis than is now possible. However, we should limit the use of these types of exercises in core components.

NAEP should continue to cover broad content areas and limit student testing time. Because of motivation and fatigue effects, student testing time should not be expanded beyond current limits. NAEP should also maintain its focus on the coverage of broad content areas. To meet both these goals, NAEP must continue to rely on instruments using matrix-sample designs.

CHAPTER 5
MEASURING CONTEXTUAL INFORMATION

EXECUTIVE SUMMARY

This chapter contains an examination of contextual data collection in NAEP. Specifically, it examines validity issues, ways to reduce respondent burden, and means to improve analysis. The following arguments and recommendations are included in this chapter:

- To reduce administrative complexity and costs, and to speed analyses, NAEP should not ordinarily link teacher questionnaire responses to student proficiency scores. Instead, samples of teachers should be used to generate data on instructional characteristics nationwide.

- Various studies indicate high degrees of uncertainty associated with student responses to certain background questions. Responses to these questions should be reported with appropriate caveats, or not reported if the data warrant exclusion.

- The size of questionnaires used with general assessment components should be limited, and should focus on items that yield valid data. Information that can most effectively be gathered through methods not amenable to mass data collection should be acquired through special studies conducted on limited samples.

- Whenever possible, data from alternative sources should be used in NAEP analyses.

This page intentionally left blank.

CHAPTER 5
MEASURING CONTEXTUAL INFORMATION

- Gita Z. Wilder -

Introduction

Taken alone, scores on achievement tests may be of limited use to policy makers, educators, and researchers. Aggregate results raise far more questions than answers unless they are accompanied by contextual and explanatory information. Most readers wish to know if overall increases or declines in achievement occur equally in all subgroups of the population, or whether school or instructional policies have changed along with student achievement. For these reasons, NAEP has included background questionnaires along with its cognitive instruments. The goal of NAEP background questionnaires is to provide context for student achievement results.

Although background questionnaires provide a rich context for the interpretation of NAEP results, there are ways in which these questionnaires and related analyses might be improved.

- The first section of this chapter looks briefly at the **measurement of contextual information in NAEP today.** In the second section **issues in measuring contextual information** are examined. Specifically addressed are questions about the **validity and quality** of NAEP contextual data. A special sub-category of validity and quality relates to the difficulties NAEP has faced in gathering data on socioeconomic status (SES) and home and community educational emphasis.

- The second section discusses ways to minimize the **burden** imposed upon participants and discusses the ways in which NAEP could profit from **better use of data that may be available from other sources.**

- The third section looks at measuring contextual information under different NAEP models—streamlined NAEP, modular NAEP, and parallel-forms NAEP.

- Finally, recommendations are made concerning the improvement of contextual information collection.

Measurement of Contextual Information in NAEP Today

Four families of background questionnaires are currently a part of NAEP. Three **student questionnaires**—general demographic and background questions, subject-specific instructional questions, and questions designed to measure student motivation—are included in the assessment booklets. **Teacher questionnaires** measure teacher characteristics, experience, training, and classroom practices. Teachers of all sampled students in main NAEP at grades four and eight are currently asked to complete questionnaires; the sample of students, rather than the universe of teachers, therefore remains the population to which teacher responses can be applied.[1] Because of NAEP's mandate, teacher responses are linked with the assessment and questionnaire responses of individual students. This linking process is complex and time-consuming.

School characteristics and policies questionnaires ask about issues such as resource availability, staff morale, and student-body demographics. The **SD/LEP student questionnaire** is completed by the school staff member most familiar with a given student who has a disability or disabilities (SD) or who is classified as limited English proficient (LEP). One such questionnaire is completed for each sampled student in either of these categories, whether they are included in or excluded from the assessment. The questionnaire is designed to gather data about the demographic characteristics and instructional experiences of special needs students.

Finally, NAEP also gathers some contextual data from sources other than these questionnaires. Schools provide assessment administrators with information on which students are receiving Title I services, which students are eligible for the national free

[1] For example, NAEP data collection and analysis techniques might support a sentence such as, "50 percent of students had teachers with college degrees in mathematics." On the other hand, current NAEP practices would not allow for the statement, "50 percent of teachers had college degrees in mathematics." Samples of students and schools are, of course, representative of the universe of students and schools.

and reduced-price lunch program, and which students have disabilities or are classified as limited English proficient. In addition, NAEP gathers some contextual information from sampling frames, such as the Common Core of Data.[2]

Issues in Measuring Contextual Information

As noted earlier, NAEP faces several important issues in the collection of contextual information. Briefly stated, the validity and quality issues have to do mainly with the student questionnaires and specifically with the fact that students do not always understand or know the answers to the questions that they are asked. These problems are particularly troublesome in the case of items that are surrogates for socioeconomic status (SES) information.

Respondent burden is an ongoing challenge for NAEP. As the requirements of the cognitive assessment become more complex, total testing time is a major concern from the perspectives of both the school and the validity of the assessment itself. In addition, context questions, especially those that ask about the home environment and family practices, have come under increasing scrutiny for their potential for invading the privacy of students and families. At the very least, eliminating items that are of questionable validity would reduce respondent burden. Finding alternative sources for selected information might enhance the validity of the information gathered, and also reduce the burden for students.

The teacher questionnaires, particularly those administered to teachers with sampled students from multiple classes, are lengthy and complicated. Typically, teachers are asked to complete separate sections of the questionnaire for each class in which there is a sampled student. Past analyses of teacher responses have shown that there is little new information beyond the initial classroom section of the questionnaire

[2] This includes data on school type and type of geographic location.

about classroom practices in second or third or fourth classes taught by a single teacher. Moreover, many of the questions in both the teacher and school questionnaires are not routinely used in reporting. Re-evaluating the need for specific information will enable NAEP to retain the questions that are essential for understanding the assessment data. Important information not required for reporting might then be collected in less burdensome ways.

Related to both the validity and burden concerns is the possibility, alluded to above, of employing alternative sources for at least some of the school- and community-level variables that have been collected through the school questionnaire. Such alternate data might be linked to the assessment data at the school level. Potential sources for some data that are currently collected from school personnel include such federal activities as the School and Staffing Survey and the Department of Agriculture data on the national free and reduced-price lunch program; indicators tracked by state departments of education; and special studies linking NAEP data to extant indicators.

The sections that follow explore many of these ideas separately, however, issues regarding the collection of context information are interrelated in practice. The descriptions rely on information from "cognitive laboratory studies" conducted with students and, sometimes, their parents; and on appraisals of the uses of contextual information in past NAEPs.

The assignment of contextual data collection to particular data sources is very much related to the overall structure of the assessment. Contextual data are gathered to support the purposes of the overall assessment, and decisions about their nature and source have implications for linking and reporting. For this reason, any recommendations about the collection of contextual data will be weighed in light of their contribution to the overall redesign.

Validity and Quality

Student Questionnaires. Small-scale interview ("cognitive laboratory") studies recently conducted by researchers[3] provide some insights into the validity and quality of fourth and eighth grade student responses to selected student background questions currently included in NAEP. While a full rendition of the findings of these studies is beyond the scope of this document, it is useful to consider some of the main findings that emerged.

Background Questions. Many of the questions asked of students—even those questions that seem straightforward on the surface—are not well understood by them. Two of the more serious threats to the validity and quality of NAEP data are located in students' reports of race-ethnicity and parental education. For example, fully one-third of the fourth graders interviewed in both studies selected options that did not seem, to the interviewers, to be accurate reflections of their racial-ethnic background. When interviewed, even some of the students whose answers appeared to be "correct" expressed confusion about how to describe themselves. There was a tendency to use the "Other" category when the listed options did not include ones that matched the student's self-description ("Caucasian," for example, or "Polish"). For students of mixed parentage, the "correct" choice was not always apparent. It is likely that, as racial mixtures become more common within the U.S. population, the need for a category that reflects more than one race will become more acute. Finally, for a relatively large number of students, the meaning of the term "Hispanic" appeared unclear; some—even those who did know the meaning of Hispanic—seemed to have difficulty with the questions that referred to their Hispanic origin.

Because the response to the race/ethnicity question is required as the basis for a key set of reporting categories, it seems important that the information gathered be as accurate as possible. Currently, NAEP collects information via test administrators from

---

[3] Campbell et al. (1997). *Preliminary findings: Cognitive laboratory study.* Princeton, NJ: Educational Testing Service.

Levine, R., Allen, J., Dubois, P., Huberman, M. & Belli, R. (1997). *Preliminary findings: Cognitive survey laboratory investigations of 4th and 8th grade students' responses to background items.* Palo Alto, CA: American Institutes for Research.

school records as well as from the students themselves.[4] Tables 5-1 and 5-2 show that the agreement levels are high but not perfect, and that group estimates derived from the two sources could yield disparate results.

Table 5-1: Self-Reported Race/Ethnicity vs. School-Reported Race-Ethnicity 1996 Mathematics Assessment: Grade 4

		Percentage Self-Reported						
% School Reported		White	Black	Hispanic	Asian/PI	American Indian	Un-classified	Total
White	N	4178.	47.	309.	44.	104.	0.	4691.
	Row %	89.3%	1.0%	6.6%	0.9%	2.2%	0.0%	100.0%
	Col. %	97.1%	3.4%	24.2%	15.22%	56.5%	0.0%	62.9%
Black	N	21.	1309.	167.	10.	33.	0.	1540.
	Row %	1.4%	85.0%	10.8%	0.6%	2.1%	0.0%	100.0%
	Col. %	0.5%	94.6%	13.1%	3.4%	17.9%	0.0%	20.7%
Hispanic	N	53.	12.	731.	3.	5.	0.	804.
	Row %	6.6%	1.5%	90.9%	0.4%	0.6%	0.0%	100.0%
	Col. %	1.2%	0.9%	57.2%	1.0%	2.7%	0.0%	10.8%
Asian/PI	N	24.	9.	52.	227.	7.	0.	319.
	Row %	7.5%	2.8%	16.3%	71.2%	2.2%	0.0%	100.0%
	Col. %	0.6%	0.7%	4.1%	78.3%	3.8%	0.0%	4.3%
American Indian	N	17.	0.	11.	2.	33.	0.	63.
	Row %	27.0%	0.0%	17.5%	3.2%	52.4%	0.0%	100.0%
	Col. %	0.4%	0.0%	0.9%	0.7%	17.9%	0.0%	0.8%
Un-classified	N	8.	7.	7.	4.	2.	7.	35.
	Row %	22.9%	20.0%	20.0%	11.4%	5.7%	20.0%	100.0%
	Col. %	0.2%	0.5%	0.5%	1.4%	1.1%	100.0%	0.5%
Total	N	4310.	1384.	1277.	290.	184.	7.	7452.
	Row %	57.8%	18.6%	17.1%	3.9%	2.5%	0.1%	100.0%
	Col. %	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Missing	N	9.	57.	11.	6.	2.	12.	
	Row %	8.5%	37.0%	10.2%	5.8%	2.0%	11.0%	
	Col. %	0.2%	4.0%	0.9%	2.0%	1.1%	63.2%	

While it is certainly not clear, given the results of the cognitive laboratory studies, which set of responses should be considered "correct," it is clear that the racial-ethnic designation contains error.

[4] Student reports are currently the primary sources for determining membership in the racial/ethnic populations reported.

Table 5-2: Self-Reported Race/Ethnicity vs. School-Reported Race-Ethnicity 1996 Reading Assessment: Grade 4

Percentage Self-Reported								
% School Reported		White	Black	Hispanic	Asian/PI	American Indian	Un-classified	Total
White	N	5824.	35.	211.	46.	91.	0	6297
	Row %	93.8%	0.6%	3.4%	0.7%	1.5%	0.0%	100.0%
	Col. %	98.3%	2.0%	15.8%	8.5%	50.6%	0.0%	63.7%
Black	N	28.	1704.	86.	7.	14.	0.	1839
	Row %	1.5%	92.7%	4.7%	0.4%	0.8%	0.0%	100.0%
	Col. %	0.5%	96.9%	6.4%	1.3%	7.8%	0.0%	18.9%
Hispanic	N	30.	6.	991.	11.	7.	0.	1045
	Row %	2.9%	0.6%	94.8%	1.1%	0.7%	0.0%	100.0%
	Col. %	0.5%	0.3%	74.1%	2.0%	3.9%	0.0%	10.7%
Asian/PI	N	10.	0.	28.	457.	2.	0.	497.
	Row %	2.0%	0.0%	5.6%	92.0%	0.4%	0.0%	100.0%
	Col. %	0.2%	0.0%	2.1%	84.2%	1.1%	0.0%	5.1%
America n Indian	N	25.	4.	11.	7.	64.	0.	111.
	Row %	22.5%	3.6%	9.9%	6.3%	57.7%	0.0%	100.0%
	Col. %	0.4%	0.2%	0.8%	1.3%	35.6%	0.0%	1.1%
classified	N	5.	9.	10.	15.	2.	10.	51.
	Row %	9.8%	17.6%	19.6%	29.4%	3.9%	19.6%	100.0%
	Col. %	0.1%	0.5%	0.7%	2.8%	1.1%	100.0%	0.5%
Total	N	5922.	1758.	1337.	543.	180.	10.	9750
	Row %	60.7%	18.0%	13.7%	5.6%	1.8%	0.1%	100.0%
	Col. %	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Missing	N	90.	81.	53.	23.	2.	25	
	Row %	24.7%	22.8%	16.2%	7.7%	0.7%	8.4%	
	Col. %	1.5%	4.4%	3.8%	4.1%	1.1%	71.4%	

Acknowledging that such discrepancies exist, it may be most useful to keep the question alive for continuing examination and study. Racial-ethnic identification can be examined closely in special, small-scale studies aimed at refining specific student questions, especially if teachers and/or parents are also involved. Continuing to collect the data from multiple sources can contribute to more accurate reporting of achievement data by racial-ethnic groups either through the incorporation of the

"error" in racial-ethnic identification in the standard errors, or by alerting audiences for the reports of the potential variations in achievement results that hold for different estimates of racial-ethnic groups. It may well be that incorporating some estimate of that error into the reporting of data by race and ethnicity would both communicate the complexity of the issue and the need for better definitions of the construct of racial-ethnic background.

NAEP should also investigate the possibilities for collecting the information—or collecting validating information in addition to the self-report data—from other sources. Such data collection might take place in small ancillary studies akin to the cognitive laboratories, outside of the data collection processes involved in the main or core NAEP. In addition, NAEP will want to retain consistency with the deliberations at the Bureau of the Census and other agencies that collect racial-ethnic information, with respect to decisions about reporting subgroups of Hispanic and Asian-American populations, and the treatment of the increasing multiracial population.

Socioeconomic status (SES) has traditionally been conceived as an index in which parents' education figures heavily. However, this item is frequently omitted by students, especially fourth graders. Interviews with students in the cognitive laboratory studies produced a range of responses to the question about parental education from "I don't know" through various degrees of uncertainty about a particular choice. Some students said they had simply guessed; others said that they had inferred the answer from the parent's job or from the existence of a yearbook or diploma at home. (A number of students found the question invasive.) Results of the cognitive laboratory studies suggest that students can provide reliable information about whether or not their parents graduated from high school; however, student-supplied information about educational levels beyond high school tends to be inaccurate when compared with parents' accounts. A re-designed NAEP should consider alternate sources for information about parents' education as well as other proxies for socioeconomic status. While we do not recommend adding a parent questionnaire, it would be useful to examine the feasibility of collecting information through small-scale studies that

attempt to gather more complete and more reliable information about SES. Other possibilities include the use of census data or school-level SES data imputed to individual children.

Other Context Variables. In the study by Levin, et al.[5], about half of student respondents overall omitted the question about school days missed in the previous month; the rate of missing data ranges from 58 percent among fourth graders to 44 percent among twelfth graders. Interviews with students about this—and other questions that involved estimates of time—showed that students have a hard time making estimates over a time period (e.g., hours of television watched on school days, hours spent reading for fun each week), resulting in inaccuracies when their estimates are compared with data from other sources, or omission of the question entirely. Such items are candidates for reconsideration. In some cases, the recommendation may be to eliminate the item; in others, it may be preferable to identify alternate sources of the data; and, in still others, it may be far better to collect the data for a small number of students in small studies via interviews that allow probing.

Instructional Context. Concerns similar to those described above abound in relation to students questionnaires in relation to the content areas assessed. Students do not reliably interpret phrases like "reading on your own" and "writing assignments." Many students do not understand what a "draft" is. Instructional context will be discussed largely in the next section, dealing with the teacher questionnaire. In general, however, these results suggest that it may be wise to re-consider using students as sources of information about instructional practice.

To summarize, elements of the context descriptions supplied by students are of questionable validity. We recommend that the items that reflect these elements be omitted, if possible; and/or that alternative sources of the data be found; and/or that additional data be gathered to validate or assist in the interpretation of the data

[5] Levine, R., Allen, J., Dubois, P., Huberman, M., & Belli, R. (1997). *Preliminary findings: Cognitive survey laboratory investigations of 4th and 8th grade students' responses to background items.* Palo Alto, CA: American Institutes for Research. Note that missing data rates are lower in actual NAEP surveys; these findings reflect only the AIR study.

collected from students. Small-scale research studies can be used to assist in the re-formulation of particularly knotty questions as well as to collect more detailed information from smaller samples. The special case of racial/ethnic background information should be monitored in conjunction with the collection of similar data in other national studies.

Teacher Questionnaires. In the current NAEP, teacher questionnaires provide information about two major categories of background variables: teacher characteristics (e.g., educational background, teaching experience, and professional development) and instructional practices. Typically, teachers complete a section of the questionnaire in which they describe themselves, and another—or several—section(s) that describe(s) instruction in the classroom(s) from which targeted NAEP students are drawn. According to this plan, any given teacher may be required to complete separate questionnaire sections for each of four or five classes. The teachers are then linked to students, and reporting takes the form, "Eighth-grade students have teachers who . . ."

Analysis of the teacher questionnaire data from the most recent NAEP assessments suggests that multiple classroom instruction sections completed by a single teacher add little new information to the database (see Table 5-3). In fact, in most cases, the sections of the questionnaire for different classes tend to repeat the same information. At the very least, the re-design should eliminate the multiple instructional sections.

Table 5-3: Redundancy in Teacher Questionnaires 1994 Reading Assessment: Grade 8

Teachers submitting questionnaires for...							
		1 class only	2 classes	3 classes	4 classes	≥ 5 classes	Total
No. of Forms With.....	N	450	175	127	128	108	987
	%	45.6	17.7	12.8	13.0	10.9	100.0
...identical information for each class		450	68	53	46	40	657 (66.6%)
...one different response in multiple questionnaires		0	58	42	40	39	179 (18.1%)
...more than one different response in multiple questionnaires		0	49	31	32	29	151 (15.3%)

We recommend an entirely different approach to collecting instructional information as context for the achievement findings. Instead of attempting to link teacher and student questionnaires, as is now the case, we propose that all teachers—or a rationally-chosen sample of teachers—at each grade level assessed in each participating school complete instructional questionnaires. The questionnaires would be shorter than the current questionnaires, but would allow for a more reliable estimate of instructional practice nationally. Eliminating the process that allows teacher and student questionnaires to be linked will save time and energy at the school level, reduce the burden for individual teachers, and spare test administrators the need to link students and teachers. There will also be savings later, in the analysis process, in that the cumbersome linking process will be eliminated. Moreover, if all or a sample of teachers at the appropriate grade level complete shorter questionnaires without reference to a particular student, the resulting data will offer more meaningful information about national practice. Such data could be collected to answer specific questions about relationships between NAEP findings and instructional practice, such as the relationship between the generality of portfolio use in a given school, a given state, or for the nation, and levels of writing proficiency.

This builds on a recent study conducted by Raudenbush, et al.,[6] who first established relationships between certain background variables and performance on NAEP and then established how those background variables changed across states. The second analysis did not directly consider performance at all; rather, the link to performance was only through the first analysis. NAEP could thus conduct a small satellite study to evaluate the relationship between a background variable and performance. Then reporting could simply be based on distributions of background characteristics in the population drawn from a survey associated with NAEP or from another survey entirely, such as the School and Staffing Survey, with no direct linkage to performance.

---

[6] Raudenbush, S.W., Fotiu, R.P., Cheong, Y.F., & Ziazi, Z.M. (1996) *Inequality of access to educational opportunity: A national report card for eighth grade math.* East Lansing: Michigan State University.

It may also be possible to improve the quality of instructional data by studying instruction over time in a small number of schools. For example, if a decision is made to include panel schools (see Chapter 6), a subset of the panel schools might serve as sites for careful descriptions of instructional practices and their stability over time.

School Characteristics and Policies Questionnaires. Much of the data collected from school personnel about school characteristics and policies in the current NAEP is not used for reporting purposes. It will be important to re-evaluate the purpose of school-level data for the assessment, and recast the questionnaires in light of the re-evaluation. Data that are essential for reporting and/or explaining achievement should, of course, be retained. At the same time, eliminating questions that have little or no value for reporting purposes will relieve school personnel of considerable burden. As it is, the information that is currently collected requires the respondent to search a number of different sources. It may be possible to collect selected data from extant sources without burdening respondents.

Relieved of the responsibility for selected school-level data, a school-based respondent may be able to provide selected student background data that the school routinely collects from sources that are more reliable than the students themselves. For example, inquiries should be made about the kinds of information that schools maintain that may contribute to a better estimate than is currently available of SES and to a more accurate reflection of the student's racial-ethnic identification.

Burden

Many of the issues involved in reducing the burden of providing contextual information have already been touched on in the discussion of validity and quality. Briefly, attempts to reduce the response burden for various categories of respondents will include both removing some of the burden altogether, by eliminating unnecessary questions; and re-assigning the burden to sources other than the current respondents, typically in an effort to improve the validity and/or quality of the data.

From the perspective of the student questionnaire, NAEP might eliminate some questions to which students do not know the answers or for which they are not reliable sources. NAEP could collect such data from carefully drawn samples of students in small-scale studies that employ more interactive data collection techniques like individual and group interviews. NAEP should also, on the basis of the results of clinical studies, sharpen the wording of questions that have proved difficult for students to understand.

From the perspective of the teacher questionnaire, we recommend eliminating the current practice of requiring multiple—and often redundant—information about the classroom instruction of each student sampled for NAEP. By eliminating this complex process for teachers, NAEP can invite more teachers per school to provide concise descriptions of relevant instructional practices, and offer more useful and expansive information about national patterns of instruction in a given subject at a given grade.

From the perspective of the school questionnaire, we recommend eliminating all but the most important questions (those that are considered essential for reporting). NAEP should explore collecting data from sources other than the school (e.g., state records and data from such other surveys as the School and Staffing Survey (SASS)). And NAEP should explore the use of extant school data and/or information from the census as better estimators of such variables as student SES and racial-ethnic background. NAEP should also explore the possibilities inherent in linking data from other sources to NAEP data for creating more useful reports than have been possible in the past about the relationship of school-level variables and achievement results. Such data sources as SASS, mentioned above, and the Common Core of Data (CCD) may enable the production of helpful reports with minimal burden to school personnel.

Use of Data from Other Sources

It may be possible to acquire valuable contextual information with little or no burden on schools, teachers, and students. For example, NCES staff were able to work

with the Department of Agriculture to obtain information from schools on students' eligibility for the national free and reduced-price lunch program. Other federal government instruments, such as the SASS, gather information that might be linked to NAEP data. State departments of education track a range of school indicators. Individual scholars can contribute findings from their research. For example, ETS researcher Harold Wenglinsky conducted a study[7] testing possible relationships between school district expenditures and student achievement. This study involved linking NAEP data from the 1992 mathematics assessment to the CCD, which contains information on school district expenditures on instruction, district-level administration, school-level administration, and capital outlays.

The preceding sections have alluded to the use of data from sources other than the traditional NAEP respondents (that is, students, teachers, and school staff) to gather the contextual information needed to interpret the achievement results. Such efforts involve shifting the responsibility from one data source to another, in order both to improve the quality of the data and to relieve respondent burden. In this way, teachers and school records may be better sources of the information that has traditionally been collected from students. Data collected by other agencies may be more reliable indicators of certain school-level variables than information provided by harried school staff. Moreover, certain of the information currently collected by NAEP itself might be available from other sources. Of course, before any attempts to use non–NAEP data are undertaken the accuracy and validity of those data must be thoroughly evaluated.

Another possible mechanism for eliminating respondent burden is to NOT collect all background information in every assessment. Questions for which the relationship to achievement is fairly stable (a good example is the one about amount of television viewing) could be omitted in certain years to reduce the total amount of

[7] Wenglinsky, H. (1996). *Modeling the relationship between school district spending and academic achievement: A multivariate analysis of the 1992 National Assessment of Educational Progress and the Common Core of Data.* (ETS-RR-96-37). Princeton, NJ: Educational Testing Service. The study linked NAEP and CCD data and used a "teacher's cost index" to account for regional variations in the cost of education. Applying a structural equation modeling program, LISREL, to the resulting database, the study found support for a model in which instructional spending and district-level administration spending were related to lower teacher-student ratios which, in turn, were related to increased mathematics achievement.

testing time. Certain teacher questions might also be administered less frequently than is now the case.

In redesigning the collection of contextual data for NAEP, NAEP should continue the effort we have started, through small-scale clinical studies with students and possibly parents, to assess the validity of the data now collected. NAEP should also undertake a critical examination of external sources of particular data, to assess their usefulness for the NAEP reporting process.

Interactions Between Measuring Contextual Information and Other Program Areas

The collection of contextual information is an integral part of the overall NAEP effort. Any change in the overall design will require a related strategy for collecting contextual information. Data that are needed for student-level reporting will be collected from or about students; data that describe instructional practices will need to be related in some conceptual manner to the student data; and school-level data will be justified only in the context of its overall value to the cognitive results. The interdependence of the separately described elements of NAEP has been pictured graphically in Figure 2-1.

Measuring Contextual Information in Different NAEP Models

In the following section, we explore the implications of three different designs for the collection of contextual information, particularly in view of the previous discussion.

Measuring Contextual Information in a Streamlined NAEP

Almost all of the ideas described in the sections on validity and burden are relevant to the notion of a streamlined NAEP. A streamlined NAEP would involve

eliminating unnecessary and redundant context questions; reassigning some of the questions to other data sources so that individual respondents, particularly students and teachers, are less burdened; collecting selected contextual information less frequently than has been the case; and making use of special studies to answer context questions that have traditionally been assigned to the main assessment. De-coupling the teacher questionnaires from individual student questionnaires would go a long way toward streamlining data collection as well as the preparation of data analysis and the analysis itself, and would also make reporting less complicated. The linkage between the teacher and student data sets could then be established in one or more special studies involving smaller samples. If there is serious resistance to the elimination of the linking altogether, reducing the teacher contribution to a single set of instructional questions rather than multiple sets reflecting separate classes would simplify the process somewhat.

The goal in all of this would be to reduce testing and analysis time, thereby reducing costs, while retaining the central features of the assessment.

Measuring Contextual Information in a Modular NAEP

A modular NAEP would consist of a core assessment designed to yield data that can be analyzed easily and reported rapidly and a series of modules that can answer more focused research or policy questions. From the perspective of contextual information, the core assessment would include only the most central contextual items describing students, teachers, and schools. The modules can take many forms. Some might involve the collection of more complex achievement data (from items that require extended responses, for example) and the contextual information that best supports such data. One such possibility would be small-scale studies that use hands-on tasks, which are more expensive than multiple-choice and constructed-response items to administer and more complicated to score. These could be accompanied by background questions that are keyed to the achievement items and instructional data from teachers and students that can be linked to the individual student achievement results. Another possibility is small-scale studies that provide the linkages between

achievement data and teacher-provided descriptions of instructional variables. The latter might even include classroom observations, to provide rich descriptions of the instructional context. A third kind of module might be one or more cognitive laboratory studies of the sort that have already been productively used for assessing the validity of certain contextual questions, and for refining the wording of questions that are found to be problematic.[8] There is evidence of the value to large-scale surveys of smaller, more focused studies that explore particular questions or findings not included in the core, or that probe correlational data for greater understanding of possible causal connections in the relationships described.

Measuring Contextual Information in a Parallel–Forms NAEP

The collection of contextual data in a parallel-forms version of NAEP would depend very much on which version of parallel-forms NAEP was adopted. How parallel–forms are constructed, how they map onto the *Framework*, and what reporting conventions will be adopted are all decisions that have implications for the number, choice, and distribution of the student background questions as well as for the choice of instructional context variables. If a parallel–forms NAEP is adopted as the core assessment, the student background set will probably be reduced to the most central variables (e.g., gender, race/ethnicity) and instructional variables will be pared to include only the most immediately relevant to the subject assessed. If a parallel–forms NAEP serves as a module, the choice of context items will be made on the basis of the purposes served by the module.

Recommendations for Measuring Contextual Information

There are several recommendations about the collection and analysis of contextual information that result from the investigations undertaken in this chapter. These are described below.

[8] There is some evidence that the feedback from one such study has contributed to a lowered rate of error in data collected since the presentation of initial findings (Levine, personal communication, May 1997).

Teacher questionnaires should not routinely be linked to student achievement scores. To reduce administrative complexity and costs, and to speed analyses, NAEP should not ordinarily link teacher questionnaire responses to student proficiency scores. Instead, samples of teachers should be used to generate data on instructional characteristics nationwide.

Further research into the validity of student responses to background questions should be undertaken. Various studies indicate high degrees of uncertainly associated with student responses to certain background questions. Responses to these questions should be reported with appropriate caveats, or not reported if the data warrant exclusion.

The size of questionnaires used with general assessment components should be limited, and should focus on items that yield valid data. Information that can most effectively be gathered through methods not amenable to mass data collection should be acquired through special studies conducted on limited samples.

Whenever possible, data from alternative sources should be used in NAEP analyses. Such information may increase the validity of background data, would provide information not available from NAEP respondents, and would reduce respondent burden.

CHAPTER 6
SAMPLING

EXECUTIVE SUMMARY

This chapter contains an examination of sample design in NAEP. Specifically, it examines the possible combination of state and national samples, samples for targeted assessments, the use of auxiliary data in sample design, and adjustments to sample weighting procedures. The following arguments and recommendations are included in this chapter:

- Given the adoption of predictable schedules, the integration of state and national samples should prove possible. Such a combination of samples should be pursued.

- Data from non-NAEP sources, including school-level data on other achievement measures, can and should be used to improve sample designs in state assessments. The cost of such improvements would be low. High priority should be given to implementing the use of such data.

- Targeted samples can be effectively designed and should be used where appropriate.

- Changes to sample weighting procedures may lead to improvements, but such changes must be studied carefully and should only be implemented in parallel with continued use of existing procedures.

- Panel sample designs may offer moderate improvements, but should only be considered if the use of auxiliary data of the type described above proves not feasible.

This page intentionally left blank.

CHAPTER 6
SAMPLING

- Keith Foster Rust / Juliet Popper Shaffer -

Introduction

This chapter considers a number of changes to and innovations in sampling procedures that may prove useful in achieving some of the goals of a redesigned NAEP. The issues that we address in this chapter are

- the combination of state and national main assessment samples
- the targeted assessment of particular groups of students, particularly those with certain levels of proficiency
- the oversampling of particular population subgroups
- the use of panels of schools to enhance the reliability of trend reporting
- the use of auxiliary information about schools and students to improve the efficiency of samples
- the broadening of the scope of the assessment to include age-appropriate students in ungraded settings
- the investigation of adjustments to sample-weighting procedures

The Sample Design of the Current NAEP

The design of NAEP samples throughout the 1990s consisted of three distinct components: national main (cross-sectional short-term trend) samples, national long-term trend samples, and state samples. The reasons for the existence of these three distinct components—for which the sample designs and analyses are separate but

related—can be found in the history of their introduction, in their development processes, and in the reporting requirements for each.

Main NAEP Versus Long-Term Trend

In the 1980s, NAEP changed its focus from reporting on populations of students defined by age to reporting by grade. Also at that time and subsequently, the scope, content, and assessment conditions (including the time of the school year when the assessments are conducted) of surveys in key subjects such as mathematics, science, and reading changed markedly from their earlier form. However, the introduction of these changes was accompanied by a strong desire to maintain trend information on the performance of the nation's students in these core subjects. Experience gained from the 1986 assessment demonstrated that such trends could only be securely maintained by repeating the earlier assessments over time following the same assessment conditions.

The programmatic needs to both periodically update the content of the assessments and maintain an extended trend line led to two distinct types of NAEP samples, known as main samples and long-term trend samples. To a very great extent, these two assessment components have in common only the fact that they are conducted during the same school year. Beginning in 1998, even this common feature is to be discontinued. Long-term trend assessments are to be conducted every four years, in odd numbered years, while the main assessments in these core subjects will continue to be conducted in even numbered years.

Main national samples use current assessment administration procedures. These samples involve systematic oversampling and are built around grade-level reporting. State samples use largely the same instrumentation but do not involve oversampling. State-level reporting is also by grade. Long-term trend samples are, on the other hand, designed to satisfy specifications from the earlier years of NAEP. These include age-level reporting and involve no systematic oversampling.

Special Features of National Main NAEP Samples. The national main samples in particular have, in recent years, contained a number of design features aimed at

meeting certain requirements in the analysis and reporting of data. These features, while beneficial, have added cost and complexity to the NAEP system. To allow accurate reporting of racial/ethnic subgroup results, public schools with greater than 10 percent enrollment of Black and Hispanic students are currently oversampled by a factor of two. To fulfill the programmatic need for reporting of reliable results for nonpublic schools, students in such schools are oversampled by a factor of three.

At times, most notably in the 1996 assessment, targeted samples of subgroups of students have been selected for particular parts of the assessment, and certain groups of students have been targeted for oversampling within schools. In the 1996 assessment, these targeted groups included advanced students in mathematics and science and special-needs students. This combination of targeted assessments of well-qualified students, plus oversampling of students with disabilities (SD) and students classified as limited English proficient (LEP), gave rise to very complex procedures for selecting student samples within schools. While the procedures were successfully carried out in the field, the existence of these special sampling procedures and the consequences of the random assignment of schools to different treatment conditions in the exclusion/accommodation experiment meant that sampling, weighting, and analysis were especially complicated.

State Samples. In 1990, state assessments were introduced. From the outset, the state samples were designed to be separate from the national main samples, but with important links between the two. There were four reasons for the decision to have separate state and national samples in 1990. First, there was uncertainty as to how many states would participate in the new program. Second, there was uncertainty about the quality of participation in each individual state, in terms of school and student participation rates, and about the quality of the assessment administration. Third, there was concern that differences in administration conditions and other factors would lead to artifactual differences in achievement between the national and state assessments. Finally, the state assessment covered only a small portion of the

populations and subjects that the National Assessment covered (in 1990 the National Assessment covered three subjects at each of three grades, while state assessments covered only mathematics at grade 8).

Although there has been extensive state participation in NAEP throughout the 1990s, the last three factors above have all been present to varying extents, so that state and national samples have remained distinct.

Issues in NAEP Sampling

The discussion above serves to illustrate that all features of the present NAEP sample design are included for purposes that are still relevant. Over time, however, the array of purposes that the NAEP samples are designed to serve has multiplied. A redesign permits an evaluation of the objectives and priorities of the assessment.

The process of settling on objectives and priorities and determining the sample design is best achieved through an iterative approach. The logical approach would be to postulate objectives and priorities and determine the sample design requirements necessary to meet those objectives. If it becomes clear that two objectives require conflicting designs, or if an objective will require sample resources that clearly cannot be justified, then the objectives can be modified and the sample design revisited. This approach is more likely to produce a sample design that satisfies the global priorities of NAEP than one that poses sample design requirements and then tries to establish whether any of the objectives and priorities could be met under such a design constraint.

For example, a sample design that calls for combined national and state samples will not produce an optimal NAEP if such a design does not permit valid equating of national and state assessment results, does not give any benefits in sampling precision for either component, and results in more cost and burden than an alternative which does not involve a combined sample. However, once the priorities for the state and national assessment programs are established, it would be appropriate to consider the

sample design of each component as a joint exercise in which a combined sample might prove to be the most cost-effective approach.

With this perspective in mind, there are a number of changes to and innovations in sampling procedures that may prove useful in achieving some of the goals of a redesigned NAEP. In some cases, an evaluation of these approaches could be useful in determining a design. Such evaluations permit judgments to be made concerning possible trade-offs among the various priorities and goals for a redesigned NAEP. Of course, the feasibility of a given approach does not, by itself, suggest that the approach should be incorporated into the design. If, however, a suggested approach appeared upon evaluation to be unpromising or problematic, then this might well impact the final determination of priority for any aspect of a redesign that hinged upon the viability of such a sampling approach.

The Combination of State and National Main Assessment Samples

For the reasons discussed earlier, recent NAEP assessments have used distinct samples of schools and students for the national and state assessment components. Since comparable results are reported for states and the nation, it is reasonable to ask whether there may be circumstances under which a combination of these two samples would be both feasible and beneficial.

State and national samples could be combined in two ways. The first would be to combine the samples at the design stage, creating a single sample for a given grade that serves the dual purpose of providing both state and national estimates. In the past, such a combination would have been infeasible and would not have offered any benefits in terms of significant resource savings or burden reduction. However, under a redesign, circumstances may be different, opening real opportunities for an integrated sample design. The second way to combine state and national data would be to design separate samples as in the past, but to use the combined data from both sources to report national and state results.

In the past, the fact that the National Assessment covered more subjects than the state assessments meant that combining national and state samples was impractical. However, NAGB's proposed schedule for NAEP specifies that in even-numbered years the coverage of subjects at grades 4 and 8 will be the same for the state and national programs. The adoption of such a schedule will provide real opportunity for beneficial combinations of state and national samples at both grades.

The outline of such a design is as follows. A sample of schools would be selected for each state, of sufficient size to provide data of the required reliability for that state, should the state choose to participate in the state assessment program for that year. A random subsample of the schools in each state would be designated as the "national" sample. In the national sample, NCES contractor staff would administer assessments in these schools, in all 50 states and the District of Columbia (and would handle recruitment of these schools in those states choosing not to participate in the state assessment program). Participating states would be responsible for recruitment of all schools to be sampled in their state, and for administration in all but the pre-identified national sample schools. All of the data collected within a state would be used for reporting results for the state, and all of the data collected in both the state administered and centrally administered assessments would be used for national results.

An important feature to be included in such a design is the equating of the assessments performed under the state and national assessment models respectively. This equating must be performed using samples that represent equivalent populations. With appropriate weights applied, the schools that are assessed using centrally recruited administrators, restricted to the set of states that take part in the state assessment program, are representative of that aggregate of states. The remaining schools in the state assessment samples will also be representative of that aggregate, provided that they are appropriately weighted. Thus the two components, administered under the two different sets of conditions, can be combined to form national estimates, using one set of weights, and can also each separately represent the

aggregate of participating states, using a different set of weights. The two forms of administration can be equated using the two separate samples with this second set of weights, while results from the assessment at both national and state levels can be reported using the first set of weights.

One important benefit of this approach would be a significant reduction in the numbers of schools and students that would be assessed by the NCES administration contractor, compared to a comparable separate design. There would also be some reduction in the numbers of schools that each state would need to assess, compared to a comparable separate design. Further, this approach would promote synergism, instead of the current competition, in recruiting schools for the national and state components (since all sampled schools in a participating state would be contributing to both efforts).

Finally, the combined sample would lessen the potential for "Lake Wobegon effects" in the comparison of state and regional data, since regional estimates would be obtained using state data from the participating states within the region (plus additional data from the national sample schools in the nonparticipating states).

A number of issues would need to be investigated in detail, however, before such a plan could proceed. Among the questions that must be addressed are the following:

- What sample size will be needed for the national subsample, to be administered centrally?
- How can the national and state sample stratification plans be made compatible?
- What will be the costs associated with having a centrally administered sample that is not clustered geographically—one that includes schools in all 50 states and the District of Columbia?
- What should the national grade 12 sample design look like under this scenario?

- What kinds of national estimates will this kind of design permit NAEP to make reliably that are not available now? (For example, will results for subgroups of larger minority groups such as Hispanic and Asian American students be available, and are these desirable?)

- Will this approach extend effectively to nonpublic school samples, since under the present arrangement many states have little interest or involvement in the recruitment of nonpublic school samples for their state?

To investigate some of these points further, we simulated some of the effects of using an integrated design approach. We then examined the likely consequences of such an approach for the reliability of national estimates.

Three variations are considered in this analysis. In the first, we consider a design using the current approach. That is, a geographically clustered national sample of schools is drawn, in addition to the unclustered samples for each state. The second design follows the integrated plan described above. The subsample of schools designated for central administration consists of 500 schools per grade, with a minimum of two per state. This overall sample size is comparable to national samples at present. The third option is similar to the second, except that the centrally administered sample consists of just 250 schools per grade, with a minimum of two schools per state. Thus this "national" component is considerably smaller than current national samples, and also tends to over-represent small states somewhat.

We evaluated each of these designs under the scenario that national results would be derived from the full set of data, both from schools where the assessment was administered centrally, and from those where it was administered by state or school personnel. Under this approach, the reliability of the national results, and the relativities under the three schemes, will clearly be critically dependent on the number of states that participate in NAEP. We considered three scenarios for this.

The first scenario was the level of participation in 1996. That is, a very large proportion of states participated (44 in total); Illinois and Ohio were the only sizable

states not included. The second scenario was to include those states that have participated in all state assessments since the beginning of the program in 1990. This comprises 33 states, including the large states of California, Texas, New York, and Florida. It also includes most states in the south. The third scenario was to include only those states that have always participated in the state program and have had very high levels of school participation. We viewed this as a reasonable representation of "committed" states that are likely to participate in the state program even if the general level of participation declines markedly. In this case, 15 states are included, and among the larger states, only Florida.

The results for public schools for grade 8 are shown in Table 6-1. The columns of the table represent the three different design options. The rows represent the three different scenarios for state-level participation. The cells in the table show the design effects in each case, and the effective sample size of schools. The design effects reflect the impact of the disproportionate distribution of the total sample across states, when considering national estimates.

Table 6-1: School Sample Size, Design Effects, and Effective Sample Sizes Under Various Redesign Scenarios, Grade 8

State Response Scenario		Sample Design Scenario			Number of Participating States
		1	2	3	
1	School sample size	4,829	4,493	4,449	44
	Design effect	3.6517	2.8559	4.4185	
	Effective sample size	1,322	1,573	1,007	
2	School sample size	3,709	3,432	3,360	33
	Design effect	4.0294	2.9009	4.7795	
	Effective sample size	920	1,183	703	
3	School sample size	1,800	1,768	1,570	15
	Design effect	4.0771	2.9349	5.3365	
	Effective sample size	441	602	294	

The table shows that, when state participation is at 1996 levels, the three designs give similar sample sizes. The design involving separate state and national samples,

combined for analysis purposes, requires more schools since each sample is designed on a stand-alone basis. Design 3, involving just 250 schools administered centrally, involves the least cost to the federal government, at least in the case that states continue to incur the costs of administration of the state samples. However, when considering Design 3, one can see that the effective sample size of schools is quite a lot lower (as the design effect is quite a lot higher) than for the other two designs. This is because the sample available to represent those states (especially larger states such as Illinois and Ohio) that do not participate in the state program under this scenario is relatively small, thus deflating the effective national sample size. When one considers the scenario of relatively few states (15) participating in the state assessments, one sees that the unclustered sample of 500 schools designated for central administration has a substantially smaller design effect, and hence greater effective sample size, than either of the other two approaches. Thus it can be seen that this approach, although possibly incurring higher administration cost to the federal government than the option with only 250 schools designated for central administration, is considerably more robust to the vagaries of the outcome of the state participation process.

The results for grade 4 are very similar to those for grade 8, as can be seen in Table 6-2 below.

Table 6-2: School Sample Size, Design Effects, and Effective Sample Sizes Under Various Redesign Scenarios, Grade 4

State Response Scenario		Sample Design Scenario			Number of Participating States
		1	2	3	
1	School sample size	5,279	4,886	4,856	44
	Design effect	3.2757	3.0748	4.3251	
	Effective sample size	1,612	1,589	1,123	
2	School sample size	4,033	3,709	3,647	33
	Design effect	4.0436	3.3117	5.1943	
	Effective sample size	997	1,120	702	
3	School sample size	2,054	1,995	1,794	15
	Design effect	3.9291	3.2598	6.0642	
	Effective sample size	523	612	296	

Further analyses of this kind, considering the effects on major reporting subgroups such as race/ethnicity and region, would help to fill in the picture. The results above, however, suggest that a design in which a subsample of the schools selected for the state samples are designated for central administration will be very effective in meeting reporting needs for state and national samples. If the size of this designated subsample is comparable to national school sample sizes in the past (Design 2 in the above tables), with a minimum of two schools selected per state, then the design can be sure of providing sound national estimates, no matter what the ultimate pattern of state participation in a given assessment turns out to be.

The benefits of this design, then, compared to current and past NAEP practice, are as follows:

- A reduction of about 10 percent in the total number of schools participating in NAEP, assuming current levels of state participation. This reduction would come from those schools in which the assessments are administered by states, so that the greatest reduction would occur in the largest states. In California, the state burden might be halved, for example.

- Synergism in the recruitment process for state and national assessment components. Instead of a school contributing either to state or national NAEP, schools would contribute to both, so that state efforts to enhance participation would benefit the national samples also.

- The data sets available for producing national results would be much larger (and hence more reliable) than they are currently. Especially if state participation remains high, the national data will be greatly enhanced for the analysis of small and clustered subgroups, such as those for race/ethnicity and for students with disabilities or of limited English proficiency.

The above discussion has concentrated on the issue of the design for public school samples for grades 4 and 8. Obviously, there would also need to be samples for grade 12, and samples of private schools. The approach to private school sample design at grades 4 and 8 could parallel that for public schools. The resulting national data sets

would be likely to provide much richer private school data, if it is possible to raise private school participation in the state assessment program to higher levels than have been achieved historically.

For grade 12, it would seem that continuing the past practice of having national samples which are centrally administered and are somewhat clustered geographically will continue to meet the needs of NAEP.

The Targeted Assessment of Specific Groups of Students, Particularly Those in Certain Proficiency Categories

Because the sampling procedures used for advanced targeted assessments at grades 8 and 12 in 1996 were successful, it is possible that this approach could be used more widely in NAEP. This approach might be particularly useful for enhancing the ability of NAEP to report on students who are at or near the advanced achievement level in each subject. This group constitutes only a small percentage of the population, and the knowledge and skills of these students may not be effectively measured by current "one size fits all" NAEP assessment instruments or samples.

Targeted-sampling procedures could also be used to draw samples of SD and/or LEP students, who could then be assessed with special versions of the NAEP instruments that are better suited to measuring their skills than the standard approach. Finally, targeting can be used to test students in subjects that are not widely taught. Such an approach will be used in the 1997 Theatre probe.

Questions that need to be addressed before the use of targeted assessments can go forward include the following.

- How well can schools identify the students to be targeted, prior to sampling?
- Is it feasible to have schools review the status of students with respect to a targeted assessment after sampling, and then to redirect the students as appropriate? Such an approach was used successfully for the grade 12 national component of the Third International Mathematics and Science

Study (TIMSS) in 1995, but required considerable contractor and school resources.

- If targeting is used for assessments that will ultimately be part of an overall national scale, what will be the consequences of such targeting for sampling errors and design effects at the overall level?

The Oversampling of Particular Population Subgroups

National NAEP samples involve oversampling of nonpublic schools and of schools in which over 10 percent of the population is composed of Black and Hispanic students. In 1996, oversampling at the student level was also used to increase the samples of SD/LEP students in the national samples. State samples have not involved any oversampling of student subgroups.

A redesigned NAEP might have different or additional reporting requirements that might require oversampling of different subgroups. Such oversampling might be accomplished at the school level or by oversampling certain types of students within schools. Some of the additional categories of students for which oversampling might be required include Asian American and American Indian students at the national level; Black, Hispanic, Asian American, and American Indian students at the state level (with the relevant groups varying across states); and private school students at the state level.

Some of the issues that require investigation include:

- What are the required levels of precision for each group that might be oversampled?
- How effectively can oversampling be accomplished by oversampling at the school level?
- What effect will oversampling at the student level have on design effects?
- What effect will oversampling of certain groups have on overall levels of sampling errors, and on sampling errors for those groups not oversampled?

- How well can schools identify students who are in categories to be oversampled?

- Would there be any problems of perception and acceptability within schools of a design that called for the oversampling of certain student groups within the schools?

The Use of Panels of Schools to Enhance the Reliability of Trend Reporting

In many national surveys that are repeated over time, sample units are retained for a number of administrations of the survey. Where it is applied, this process often greatly reduces sampling error in trend estimates and, through a procedure known as composite estimation, it also improves the precision of current estimates. For NAEP, the measurement of trends in achievement over time has always been a priority. Historically, NAEP has not used any form of sample retention over time.

The reasons for this are twofold. First, it has been considered important not to overburden individual schools with repeated requests for participation every two years. However, in the state assessment program, this is happening in small states anyway because such a high proportion of schools must be included in the sample for each assessment. Second, there is concern that, should individual schools come to be viewed as "NAEP schools," they may no longer be representative of the country or state.

A redesign presents the opportunity to consider whether an overlap (or panel) design is appropriate for NAEP. Note that NAEP would differ from other surveys that use this approach in that the ultimate sampling unit, the student, would not be retained in the sample. Indeed, due to the length of time that will pass before a given assessment is repeated, the student population at a given grade will have completely turned over. Enough is known about the variance components of NAEP estimates to show that, because the student sample would not be the same, dramatic gains in sampling error will not accrue from retaining the same schools in the sample. However, modest gains

might result. It is an empirical question just how great these would be, and whether they would outweigh the disadvantages of such an approach.

Issues that would need to be addressed before introducing a sample design that had built in overlap of school samples over time include the following:

- How much reduction in sampling error for trend estimates would result?
- How difficult is it to match schools from one sample with a school sampling frame several years later?
- How serious might the potential bias be that would arise from keeping the same schools in the sample, relative to the sampling error gains?
- Are there any incentives to participate that would be effective when keeping the same schools in the sample over time, that would not otherwise be appropriate (keeping in mind that a given assessment will be repeated every four years, or less frequently in some cases)?

Two studies have been carried out to investigate consistency of school means over time and school clustering effects. The aim of the studies was to assess the possibility that a panel design would improve the estimation of population values.

Study 1. In small states, achieving sufficient accuracy of estimates has necessitated sampling most of the state's schools every time the state is assessed. A first study investigating consistency of school ranks over time was carried out using data from 10 states with substantial numbers of schools assessed at least twice. The correlations of school means were calculated for eighth grade mathematics in three time periods: 1990, 1992, and 1996. Correlations were obtained among normits, averages of plausible values, and percent of students above each of the achievement levels. These were based on unweighted data, and on non-SD/LEP students (except for the normits).

The median of the 3 correlations (1990-1992, 1990-1996, 1992-1996) among the school means on the average plausible values are given for each state in the

first column of Table 6-3. The values in the second column are the multiple correlations, predicting 1996 means from 1990 and 1992 means, for the seven states with data in all 3 years.

Table 6-3: Consistency of School Means Over Time

State	Median Correlation	Multiple Correlation
Delaware	.74	.76
District of Columbia	.91	.92
Hawaii	.69	.76
Maine	.46*	---
Montana	.47**	---
Nebraska	.75	.78
New Hampshire	.40	.46
Rhode Island	.81	.81
Utah	.58*	---
Wyoming	.54	.75
Median across states	.64	.76

* Data for 1992-1996 only
** Data for 1990-1996 only

These values are impressively high, considering that in some states certain schools had very small numbers of students taking the assessment. For example, the minimum number of students assessed in a school in Montana was 2; in Rhode Island, 4; in Hawaii, 6; and in New Hampshire and Wyoming, 9. There is a slight tendency for states with smaller minimum school sizes to have smaller correlations.

The most useful school data for design and/or estimation purposes would probably be mean values on recent state assessments, since those would be given to virtually all students in the school. However, the values above are high enough, at least in some states, to be useful in cases where state assessment data are unavailable, or possibly useful in combination with available state data. Furthermore, correlations

might be higher if all students sampled in a school were included, regardless of the subjects in which they were assessed. This possibility should be investigated. One possible use of these data is for ordering schools in choosing systematic samples, perhaps allowing for use of a smaller number of schools, or a smaller number of students per school.

Study 2. In addition to the correlation study in small states, a nested analysis of variance was carried out in large states to assess the degree of clustering within schools, using 1996 state mathematics data. The design for each state used schools as the highest variable, students within schools as the next variable, and plausible values within students as a third nested variable. Components of variance for schools and students are given in Table 6-4, as are the intraschool correlation coefficients (i.e., the school component of variance divided by the school + student components of variance).

Table 6-4: Components of Variance for Students and Schools

State	School Variance	Student	Intraschool
Alabama	345	718	.32
California	340	740	.31
Colorado	200	672	.23
Connecticut	363	721	.33
Florida	231	868	.21
Kentucky	143	664	.18
Michigan	366	732	.33
Minnesota	145	746	.16
New York	385	724	.35
Texas	226	710	.24

The values break rather sharply into two groups of five each, based on the intraschool correlation values:

- The states with relatively low values were Colorado, Florida, Kentucky, Minnesota, and Texas.

- The states with relatively high values were Alabama, California, Connecticut, Michigan, and New York.

These values could be useful in determining relative numbers of schools and students within schools required for a given degree of accuracy of sample estimates.

However, further analyses and more refined studies beyond these two studies are needed before panels of schools are employed as a part of the sample design. One feature of the state sample designs not considered here is the effect of stratification. As this removes a component of between-school variance in every state, and removes all of the between-school variance in smaller states where all schools are included with certainty, there will be some mitigation of the potential benefits from utilizing the correlation over time at the school level.

Nevertheless, the results of these two studies are impressive enough to indicate that using past NAEP data, or state assessment data, to stratify schools is likely to give important reductions in sampling error for many states, resulting in more reliable estimates, smaller samples, or both. This is the topic of the next section.

The Use of Auxiliary Information About Schools and Students To Improve the Efficiency of Samples

Currently NAEP stratifies schools using characteristics that are somewhat related to achievement, but does not attempt to stratify students within schools. Very substantial gains in reducing sampling error could be realized if it were possible to stratify schools using school-level achievement data from related assessments, and/or to use such data in the design of samples of students within schools.

The type of school-level data that are most likely to prove useful are school mean scores on state assessments. These data would not need to exist in every state, or even to be consistent across states, to be useful for stratification purposes. Indeed, even having the data missing for a portion of the schools in a given state would not be a serious drawback. With a demand to reduce state school sample sizes, further investigation of this approach seems warranted as a part of the redesign effort.

The American Institutes for Research have in recent years undertaken several evaluations of NAEP, for which they have obtained school-level assessment results for state assessment programs. Their research efforts indicate that these data are available at the school level for many states, and that state assessments in the same subject as a given NAEP assessment generally correlate highly at the school level. For reading assessments, for example, correlations in the range of 0.5 to 0.8 were typical.

Furthermore, these data are widely available at the school level in many states. For a sizable number of states, extensive information about schools, including school-level performance on standardized tests, are actually available through the Internet. Clearly, in these cases there are no issues of confidentiality, sensitivity, or ease of access in using these data in the design of NAEP samples.

With the high levels of correlation that are evident for these school-level results, it seems that significant sampling precision gains could be realized from using these data in the design of NAEP school samples. With the kinds of correlations shown in recent studies, it seems very likely that the contribution of between-school variance could be reduced from that currently experienced. That is, it seems very likely that school mean achievement on a state test is likely to be a superior stratification variable to at least some of those currently used: type of location, enrollment of minority students, and 1989 median household income of local ZIP code area (1999 income will not be available until at least 2002, and perhaps never).

The major difficulty in using these school achievement data in designing NAEP samples is likely to be that formats will be very inconsistent across states. Thus,

matching these data to schools on NAEP sampling frames may often prove difficult. However, the likely benefits seem to be so great that this challenge must be faced. The major step that will be needed is to devote more time and resources to developing NAEP frames. There are also important implications for the timeliness of the availability of Common Core of Data (CCD) files to begin the NAEP frame building process.

An alternative use for such data is in the NAEP weighting process, rather than in stratification. Although sampling precision benefits are accessible through this approach, we believe that there are several reasons why this is much less desirable than using the data for stratification. First, using the auxiliary school-level achievement data in weighting will likely add significant time to the weighting and analysis process. With use in stratification, the extra time needed can be added at the front of the process, rather than delaying the results. Second, in cases where a few schools do not have comparable state-level results available—which is likely to be a common situation—it is much more straightforward to deal with this process at the stratification stage than at the estimation stage.

Third, we believe that using data of this kind for stratification rather than analysis (weighting) will have greater face validity to states and other users of the data. Most people can appreciate that it is reasonable to ensure that the sample has good representation across the spectrum of achievement on state-wide tests. But people may be concerned about using these data in a way that affects the relationship between the student responses actually collected in NAEP, and the NAEP assessment results produced, which is what happens if the external data are used in weighting.

Finally, if there are a few schools with exceptionally high or low achievement on the state test, through stratification we can be sure that these are represented in the sample (with appropriate weighting). But if the information is only used at the weighting stage, nothing can be done if perchance none of these schools happens to be selected in the sample.

An alternative use for data from state and local assessments would be in the design of the student samples within schools. If students could be stratified or sorted for systematic selection on the basis of individual results on a related assessment, again very significant reductions in sampling variance might be achievable.

Although this approach might be technically worthwhile, we believe that any initiatives to use student-level data in this way should proceed very cautiously. Unlike the case with school-level data, concerns about student confidentiality and privacy will be paramount in this situation. Circumstances may vary greatly from school to school, as to both the availability of suitable data and the sensitivity to requests to use it in drawing student samples. It would not be wise to proceed to adopt this approach wholesale, on the basis of just a few successful case studies, for example.

An additional issue is timeliness. Whereas using school-level data for the same grade from an earlier cohort would be quite satisfactory, at the student level this is of no use. Thus one would need to be sure of getting timely data about students for their current grade, at the time of student sampling in January, or else use data from an earlier year. This information will be missing for some students and will be of less value all around than an assessment from the grade being assessed by NAEP. Unfortunately, then, the alternative of using student-level auxiliary assessment data in weighting has all of the negative features associated with using school-level data for weighting (timeliness, face validity, etc.), as well as the negatives associated with using student-level data for sampling (confidentiality, school sensitivity, etc.). Thus, we would propose that this approach not be pursued initially.

We should note that our general conclusions about the use of auxiliary data from state assessments are supported by our advisory panel. These conclusions are:

- School-level data on achievement are likely to be readily available in many states.

- Such data are likely to offer substantial benefits in enhanced sample design and reduced sampling error.

- It would be better to use these data for sample design than in analysis and weighting.

- Any endeavors to use student-level achievement data in the design of NAEP samples within schools should proceed very cautiously.

The Broadening of the Scope of the Assessment to Include Age - Appropriate Students in Ungraded Settings

NAEP currently targets students defined by grade membership in its cross-sectional national assessments and state assessments. But there is interest in enhancing the inclusiveness of NAEP until it includes as broad a spectrum of the student population as can be accommodated. To this end, NAEP investigated the use of accommodations for SD/LEP students in the 1995 field test and 1996 assessment. A logical extension of these efforts at inclusiveness would be to add to the surveyed population those students in ungraded programs who are of appropriate age. Thus at the middle-school level, national and state results might pertain to students who are in grade 8, plus ungraded students who are, say, 13 years of age.

If ungraded students are to be included in NAEP assessments, issues to be addressed are:

- What age definition should be used to define eligibility for ungraded students?

- Will special accommodations be required to assess ungraded students?

- Does the Common Core of Data file which is used to create NAEP school sampling frames include information as to which schools have ungraded students of the relevant age?

- What proportion of the NAEP population would be made up of ungraded students?

- How would issues of short-term trend be handled, given that ungraded students have not been included in NAEP populations in recent years?

The Investigation of Adjustments to Sample-Weighting Procedures

Westat recently investigated the possibility of using a procedure (called raking) for adjusting national NAEP sampling weights as an alternative to the poststratification procedures that have been used over the past several years.[1] Although this research suggests that raking using national population data offers little advantage over poststratification, the possibility remains that alternative weighting procedures using auxiliary data at the state or national level might increase the reliability of state or national estimates, or both.

Some of the same kinds of data discussed above as possible enhancements in the design of samples might be utilized in weighting instead. This might be beneficial if, for example, the relevant data could not be obtained in time for sample design and selection, but could be obtained in time for sample weighting. Alternatively, different variables might be used in weighting than in sampling. In addition to data on academic achievement at the student, school, and state levels, school characteristics associated with achievement also might be used effectively.

Issues that would need to be addressed before such data could be used in weighting include the following:

- What potential auxiliary variables are available on a wide scale that could be used in estimation?

- What gains in sampling error could be achieved from using these data, both for national and state level results, and also for population subgroups?

[1] Wallace, L., & Rust, K. (1996, August). *A comparison of raking and poststratification using NAEP data.* Paper presented at the annual meeting of the American Statistical Association, Chicago.

- What form of estimation would be most effective: poststratification, raking, generalized regression?

- What would be the implications of using these auxiliary data, and associated estimation procedures, for the reporting schedule for national and state NAEP?

This last question indicates that it will be important to investigate any modifications to NAEP weighting "off line". The NAEP reporting schedule will not permit a scientific investigation of alternatives in time for any of them to be used for the assessment for which they are being investigated. Thus the 1996 evaluation of raking, discussed above, was carried out using 1994 NAEP assessment data. Had the approach proven more beneficial, it could have been incorporated in the 1996 weighting procedure, but certainly could not have been implemented in time for use in reporting 1994 results.

It is also the case that, if new procedures are developed, the first time that they are implemented in production it will be necessary to produce two sets of weights: the new set and a set calculated using previous procedures. In this way there will be both a check and a back-up of the new procedure. If the assessment shows any unusual results, it will immediately be possible to ascertain whether or to what extent these are a function of the change in procedure.

Interactions Between Sampling and the Other Program Areas

Sample design impacts analysis and reporting, since these are dependent on the data gathered. It also impacts data collection, since the design of the sample determines, in part, how the administration of the assessment will proceed. Sampling affects database complexity as well. The *Frameworks*, instrumentation, and reporting requirements all influence sample design. For example, an assessment in which each student takes half of all items will necessarily have a smaller national sample than an assessment in which students take, for example, 10 percent of the entire pool. Assessments that allow different subjects to be spiraled together allow for more

efficient sampling designs than do those that require subjects to be administered in separate sessions.

Recommendations for Sampling

State and national samples should be integrated for common grades and subjects. This should be achieved by identifying a subset of state sample schools in all 50 states and the District of Columbia to be administered centrally, regardless of whether the state participates in the assessment. The sample size for central administration should be about the same size as national samples have been in the past.

This approach will lead to much more useful data at the national and regional levels. It will enhance participation in centrally administered schools. It will have little impact on cost. The approach is robust to the level of state participation in NAEP.

Targeted assessments can be effectively designed from a sampling perspective, at least for schools with central administration. Thus sampling considerations should not be an impediment to the use of targeted assessments in national samples.

Retaining schools in NAEP samples over time has the potential to significantly improve the precision of trend estimates (and, through the use of composite estimation, of cross-sectional estimates also). However, there are issues of burden and operational efficiency. A consideration of whether to retain schools in the sample over time should be made in conjunction with considering the use of other options to reduce the school component of variance (which will have benefits for both cross-sectional and trend estimates).

The use of auxiliary school level achievement data in sample design should be a high priority. This has the potential to significantly improve the precision of cross-sectional estimates (and hence, correspondingly, trend estimates). The use of such data in weighting and analysis should be given lower priority, and requires an evaluation of

the impact on schedule, and of user perception. The use of student level auxiliary data for use in sampling should be investigated, but very cautiously, with full recognition of the privacy issues involved for students and schools.

Any significant changes to NAEP sample weighting procedures must be made in parallel with continued use of existing procedures. This is necessary both to ensure that new procedures do not negatively impact timeliness, and to thoroughly evaluate the impact of the new procedures on results.

CHAPTER 7
DATA COLLECTION

EXECUTIVE SUMMARY

This chapter contains an examination of data collection in NAEP. Specifically, it discusses the ways in which data is currently collected in NAEP, ongoing issues regarding complexity and respondent burden, and questions related to computer-based testing. The argument views data collection in a supporting role; in other words, it considers the implications of other assessment instructions for data collection rather than attempting to first design a data collection system and then attempting to fit other survey components to that system. The following arguments and recommendations are included in this chapter:

- The implications for effective data collection must be considered before any assessment framework or design is adopted. However, data collection should not be the primary factor considered when designing assessments.

- The use of computer-based testing methodologies should be preceded by careful studies of operational feasibility.

- NAEP should reexamine administrator training models and look for cost and burden efficiencies. Currently, state administrators must attend in-person training. This adds cost for the federal government and for participating states. It also makes NAEP more burdensome for local participants. NAEP should consider eliminating this requirement, and should investigate the use of distance-learning methods for administrator training.

This page intentionally left blank.

CHAPTER 7
DATA COLLECTION

- Nancy W. Caldwell -

Introduction

This chapter considers a variety of issues that affect data collection. A major concern is how to reduce both **actual and perceived burden.** Another is the impact of **innovative assessment procedures** on data collection. **Non-standard administration** is an area that, while improving inclusiveness, adds complexity to data collection. Since data collection is affected by all other functional areas, decisions about other aspects of assessment design will have implications for data collection.

Data Collection in the Current NAEP

Until the early 1980s, administration of NAEP assessments was relatively simple. Only a small number of session types were conducted; small numbers of students were assessed in each school; ancillary materials were limited; and the sessions were directed by a paced audiotape that presented the directions, as well as all questions and answer options, and controlled the timing of each section. This meant that the administrator's main responsibilities were to distribute materials and maintain standard conditions in the classroom. In addition, the number of background questionnaires was limited.

Since that time, the assessments have become much more complex. Balanced Incomplete Block (BIB) spiraling permits many different booklets to be used in the same session, expanding the item pool but also requiring that the administrator play a more active role in the sessions and that students read most of the assessment to themselves.

NAEP *Frameworks* today require constructed-response and performance items. Because of the time requirements of these types of items, the total length of the assessment is longer than in the past. This increases the burden on schools and students. The increased time burden and greater use of authentic materials also make the administrator's job more complex. As the administrator is required to be more actively involved in the assessment, there is a greater chance of administrator differences affecting the outcome. Therefore, training and monitoring of administrative staff have become more important.

Additionally, the desire to capture more contextual information about the students, their teachers, and their schools has led to the use of a variety of background questionnaires. Coordinating the use of these questionnaires has led to increased complexity of administration as well as to increased burden.

Issues in Data Collection

Reducing Costs

Within the current NAEP design there are areas where data collection costs could be reduced, but at a reduction in oversight and control. Fairly sizable savings could be realized, for example, by eliminating the requirement that all state assessment administrators attend in-person training sessions. This requirement not only adds cost to the system, but engenders some ill will among the state staff responsible for the assessments. This is particularly true in states where the same schools are in the sample from year to year and often the same people serve as assessment administrators. To conduct training sessions each year costs about $20,000 per state, or $1 million total, for personnel costs, travel, room charges, etc. This does not include the in-kind contribution of the states and districts for the time of their staff to attend the sessions and for the substitute teachers to cover for them.

There are several scenarios that could be considered, especially in years when the assessment is a repeat of a previous administration or does not involve a lot of

ancillary materials. The first would be to provide a home-study refresher for persons who had been assessment administrators in previous years. This would not eliminate the training costs, but could reduce them by up to 25 percent. Another approach would be to pilot a more detailed home-study package with accompanying video which all assessment administrators would be sent, instead of attending a training session. This would save more costs, but would introduce more variability into the assessment procedures.

Minimizing Burden

There are two ways to approach the challenge of minimizing burden while maximizing information. They are: 1) to reduce the actual burden on the respondent by shifting the burden to some other part of the system, and 2) to reduce the perceived burden by increasing the benefit of participation. Our discussion of approaches to revising some of the other functional areas of NAEP relates directly to what can be done in the area of data collection. In particular, **sampling** procedures intended to minimize school sample sizes and methods of **measuring contextual data** address ways to reduce actual burden, or at least get more information while maintaining existing levels of burden. **Reporting** techniques to better inform educators, parents, and leaders, and innovative means of **measuring cognitive skills** also offer ways to reduce perceived burden. We will address each of these and their relationship to data collection in the sections which follow.

Interactions Between Data Collection and Other Program Areas

Sampling Procedures and Data Collection

One of the proposals under consideration regarding the sample design is to integrate the state and main NAEP samples. The presumption is that the overall total

sample would therefore be smaller than under the current design where each state is a separate sample. A reduced total sample reduces the burden in the system as a whole. Precisely where the burden is reduced depends on how the integrated sample is operationalized. For example, if the decision is that NAEP staff will conduct all assessment activities in all schools in the integrated part of the sample, then the burden will be greatly reduced on states and schools, but increased on the NAEP contractors, and therefore the federal government.

On the other hand, if schools providing data for both the state and national analyses were the responsibility of the states, the burden would be reduced at the federal level. There are some operational concerns with this approach, however. What happens if a key "national" state decides not to participate in the state assessment, or even worse, drops out at the last minute? How will this gap in the national data be filled? Similarly, if states are given options to select a subset of the subjects and grades being offered by state NAEP, then the benefits of integrating the samples are less apparent. There would still have to be nationally representative samples in all the subjects and grades. This might lead to a combination of national and state administrations in the national sample, which would have implications for equating.

Another aspect of the sample design under consideration that can have operational implications is oversampling of particular population subgroups. Whether to identify subgroups to receive a targeted assessment (such as students studying advanced subjects) or to oversample subgroups for reporting purposes, oversampling within schools adds complexity for schools and field staff. Schools must identify the subgroups of interest. Field staff must sample them at a different rate than other students. Depending on how easily identifiable the group is in school records, creating lists for sampling purposes may be quick or time-consuming. Schools may or may not allow NAEP staff to have access to the appropriate records, shifting the burden to school staff. Or, schools may not be willing to devote staff to such a task, leaving the burden with NAEP field staff.

The use of panels of schools to enhance reliability would increase the burden of

participation on the cooperating schools and their teachers. It probably would increase the burden on NAEP staff in securing initial cooperation. In projects such as the National Education Longitudinal Study (NELS), where schools are in the study consecutive years because students are being followed over time, the initial school response rates are under 70 percent and it takes a long time to build the sample by including substitutes. In the out-years, however, since the schools have agreed to participate, there would be reduced burden on NAEP staff.

Using auxiliary information about schools and students to improve the efficiency of samples would have few operational implications unless the information to be used had to be collected by field staff directly from schools. The level of burden would be directly related to the extent that the information is readily available.

Assessing age-appropriate students in ungraded settings offers some challenges operationally. Students may be in ungraded settings for a variety of reasons. Just because they are age-appropriate does not necessarily mean that they have been exposed to the curriculum appropriate to the grade being assessed. Schools may resist using assessment materials clearly labeled with a grade for students who are not receiving instruction at that grade.

Measuring Contextual Information and Data Collection

Our approach to improving the collection of contextual information has at its core a commitment to reducing burden on the schools, teachers, and students selected for NAEP. This can be done in a variety of ways including: collecting information from only a subsample of the population, using computer or telephone technology rather than printed questionnaires, and carefully pruning the questionnaires so that only information that has a role in reporting is collected. These options are described more fully in other sections.

It appears unlikely that telephone polling procedures can be fruitfully used to collect data from teachers, although such techniques might work for schools. The

reason for this is the perception that teachers would prefer the printed questionnaires because these can be filled out in odd moments. Finding time when they can come to the phone to answer questions during the school day may be problematic for teachers, and calling the teachers at home is not an attractive option. This may be a subject for a focus group of teachers.

Making better use of data available from other sources is another approach we discuss for improving the measurement of contextual information. Again, depending on the source, there may or may not be operational implications. If the source is another federal survey, such as the School and Staffing Survey, there would be no implications for data collection. If, however, the source is something like the school records on the school lunch program, there may be considerable burden on the field staff. This information is confidential and closely guarded by school staff. In some schools, only the director of the cafeteria knows which students are eligible. In these schools, then, the NAEP field staff must locate this person and have him or her review the list of sampled students to identify these students. This adds time and cost.

One of the options under consideration is for NAEP to collect teacher data from all teachers in a school, not just teachers of sampled students, and to eliminate the linking of student and teacher data. In many schools, NAEP is collecting some data from all or almost all teachers currently. Given that NAEP selects a cross-sectional sample of grade-eligible students, it is likely that each teacher at that grade will have at least one student in the sample and therefore will be asked to complete a questionnaire. However, since the questionnaire currently focuses on how the teacher teaches those classes containing sampled students, information is lost on how they teach other classes. For some teachers, the picture received by NAEP is very incomplete. If the questionnaires addressed the more general questions of instructional approaches and were given to all teachers, then much more information would be collected. Operationally, the number of questionnaires that would have to be distributed and collected might increase, but not substantially.

Measuring Cognitive Skills and Data Collection

One approach to measuring cognitive skills that has implications for data collection relates to the use of performance tasks. Performance tasks generally require more time and skill to administer and therefore affect the training of field staff and the amount of time required to conduct all assessment activities in sampled schools.

Another approach we discuss is a modular design where subsets of students take different assessment modules. To the extent that a modular design is more efficient—requiring only a small number of students—the overall sample sizes are reduced and therefore data collection costs are lowered. However, with this efficiency may come some complications, and some costs. For example, more session types will have to be conducted, increasing the sampling and administration complexity.

Computerized testing has come under increased scrutiny as more and more students become computer-literate. The computer engages students in ways that a test administrator cannot. Students can work at their own pace, the computer can make sure that they are on-task, and most "paperwork" is taken care of by the computer. There are some disadvantages operationally. Most disadvantages revolve around the assumption that schools cannot be expected to have available at the time of the assessment the required number of computers that are in working order and that are compatible with the assessment. Therefore, a major disadvantage involves the logistics of transporting, setting up, and protecting the security of large numbers of computers. Some schools require that all NAEP assessments be conducted at the same time (although in different locations). In large high schools, 150 to 200 students may be in the sample, so that in those schools, the NAEP field staff would have to bring in 150 to 200 laptop computers. Multiply this by 50 field supervisors, and the number of computers needed becomes very high. Another challenge is that time will be needed to train the students on the particular computers that are being provided.

Reporting and Data Collection

Chapter 10 suggests that NAEP reporting be targeted to address the interests and concerns of educators, parents, and policymakers. All of the techniques discussed work on increasing the perceived benefit of NAEP, thereby reducing the perceived burden. One of the major concerns of schools and teachers when they are contacted about participating in NAEP is "What is in it for my school, my students, my teachers?" Although NAEP cannot provide individual, school-level, or district-level data, it can provide information on instructional practices and policies. The reports that NAEP has produced have been very important tools in securing cooperation in the past. The more they are designed to address the concerns of these key groups, the greater the perceived value of NAEP and participation in it.

Data Collection Under Different NAEP Designs

Data Collection in a Streamlined NAEP

A streamlined NAEP would impact data collection in several ways. To the extent that special studies are eliminated and the assessment is made simpler, the amount of national NAEP field staff time required for each school will be reduced. Cost savings are not likely to be great, however. For example, with a sample of 2,000 schools, eliminating the need for one exercise administrator at each school would save approximately $75, for a total savings of $150,000. Depending on how streamlined the assessment becomes, these savings could increase somewhat but not greatly. There would be comparable savings for the state assessment, but these are in-kind costs (and savings) for the states.

If NAEP were significantly streamlined—for example, if the questionnaires were significantly reduced, eliminated, or simplified—then the in-school time requirements of field staff would be reduced. In addition, the more the assessment is just an assessment, without all of the current accompanying questionnaires and procedures, the more reasonable it becomes to consider reducing training requirements for the state

assessment. In the current training program, the majority of the time is spent on forms, questionnaires, and paperwork. The assumption made is that the school staff attending the session know how to conduct testing programs; they just need to know the particular paperwork and reporting requirements of this assessment. The need for in-person training is reduced as the complexity of the assessment is reduced as well.

Data Collection in a Modular NAEP

The discussion above regarding a streamlined NAEP applies to a modular NAEP as well, particularly if the "core" is simple to administer and uncluttered with questionnaires. If the modules are implemented in the national assessment and not in the state assessment, the cost savings would accrue in the state assessment. To the extent that modules require special procedures, such as identifying and sampling subpopulations, they might increase some data collection costs while reducing others. For example, the advanced mathematics assessments required that students taking certain subjects be identified and sampled. In some schools this information had to be obtained by the field staff from school records, increasing the amount of time they were in the schools preparing for the assessments. By contrast, the fact that there were a limited number of sessions reduced the amount of time overall that field staff were in schools conducting assessments.

Data Collection in a Parallel-Forms NAEP

Our understanding of a parallel-forms NAEP is that there would have to be significant piloting of a very large number of items, followed by field-testing of each of the proposed parallel forms with a nationally representative sample of students and schools. This would add significantly to data collection costs, at least during the developmental cycle. The level of additional costs would depend on the number of students and schools required to conduct the pilot and field tests. The long-term trend assessment provides some guidance in this regard. In the typical school in the long-

term trend assessment, NAEP conducts two to three sessions, each involving 25 to 30 students, in about 300 schools per grade, at a cost of about $1,000 per school for all activities. The advantage of the parallel-forms NAEP, once it is developed, is the simplicity of administration. If parallel forms are used as the only assessment vehicle in the state and/or national assessments, the advantages and cost savings described under the modular and streamlined NAEP apply. If the parallel forms are just one module or component of a complex design, there may or may not be any cost savings.

Recommendations for Data Collection

There are three recommendations about data collection that issue from the investigations undertaken in this chapter. These are described below.

Data collection should not be the primary factor in design considerations; however no design that is untenable from a data collection perspective should be adopted. Different assessment designs will lead to different data collection systems. However, one would not wish to first design a data collection system and let that govern what one could measure or report. Yet the converse is not true; data collection must not be ignored in system design. Some measurement goals might prove prohibitively expensive or not feasible to implement in the field, and these should be avoided.

NAEP should reexamine administrator training models and look for cost and burden efficiencies. Currently, state administrators must attend in-person training. This adds cost for the federal government and for participating states. It also makes NAEP more burdensome for local participants. NAEP should consider eliminating this requirement, and should investigate the use of distance-learning methods for administrator training.

The use of computer-based testing methodologies should be preceded by careful studies of operational feasibility. Computer-based testing will likely change

NAEP data collection in a fundamental manner. Such effects must be studied carefully before planned implementation.

This page intentionally left blank.

CHAPTER 8
SCORING

EXECUTIVE SUMMARY

This chapter contains an examination of constructed-response scoring in NAEP. It begins with an examination of the current NAEP image-processing scoring system, and examines ways in which the system can be improved or expanded to remote sites. The chapter then examines expert systems that enable computers to score student responses. Finally, the chapter discusses ways to improve rater reliability, and methods for statistically adjusting for rater unreliability. The following arguments and recommendations are included in this chapter:

- Existing image-processing scoring tools should be expanded to include enhanced training and monitoring capabilities.

- A feasibility and cost/benefit study of remote–site Internet–based scoring should be conducted. Such a study should consider fiscal, technical, quality, and security factors.

- Automated scoring programs should be evaluated for possible use in components of NAEP. Research into such programs should involve identification of exercises amenable to computerized scoring, development and testing of machine-scorable performance items, and investigation of the effects of separating machine and human scorable exercises into different sections of cognitive instruments.

- NAEP should research methods of statistical adjustment that would compensate for any effects introduced by rater unreliability.

This page intentionally left blank.

CHAPTER 8
SCORING

- Christine Y. O'Sullivan -

Introduction

Over the past two decades there has been increasing interest within the educational assessment community in the use of items that allow students to construct their own responses as opposed to choosing from a selection of answers. Recent NAEP assessment *Frameworks* have endorsed this interest. For example, the 1996 NAEP science assessment at grade 8 contained 74 multiple-choice questions, 100 short constructed-response questions, and 20 extended constructed-response questions, whereas the 1990 NAEP science assessment at grade 8 contained 95 multiple-choice questions and 17 short constructed-response questions. The importance of constructed-response items is noted in a report produced for the National Assessment Governing Board (NAGB) by the Design/Feasibility Team[1] and endorsed by the National Research Council's evaluation of the NAEP redesign.[2] Given the desire to maintain a mix of constructed-response and multiple-choice items in NAEP assessments, this chapter considers scoring issues that relate to cost efficiency, timeliness, and reliability within the current NAEP model and relates the findings to the proposed models.

First, the existing **computer-based scoring system** is examined to determine ways to reduce costs and scoring time while retaining high rater reliability. The feasibility of modifying the current scoring model to include such attributes as scoring via a secure Internet site is also considered. Second, **automated scoring systems** that are currently being developed and/or evaluated are investigated to determine which

[1] Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design/Feasibility Team: Report to the National Assessment Governing Board.* Washington DC: National Assessment Governing Board.

[2] National Research Council. (1996). *Evaluation of "Redesigning the National Assessment of Educational Progress".* Washington, DC: National Academy Press.

types of performance items can be scored accurately by machine. Third, although **rater reliability** has improved since the advent of electronic scoring, rater errors do occur, and statistical models are proposed to identify aberrant raters and detect types of rater errors during the scoring process. In addition, statistical means that can be used to calibrate the scores of aberrant raters during analysis are indicated. Fourth, given the integrated nature of NAEP and the probability that change in one area will inevitably lead to change in another area, ways in which different scoring methodologies affect the overall project in terms of cost and time are also examined. Fifth, where appropriate, the impact of scoring changes with respect to the proposed new models—streamlined NAEP, modular NAEP, and parallel-forms NAEP—are discussed. Finally, recommendations are made that will ultimately move NAEP into the 21st century by encouraging technological methodologies that will aid in streamlining NAEP and reduce costs and cycle time while maintaining high rater reliability.

Computer-Based (Image) Scoring

Since 1994, all constructed-response questions administered in NAEP have been scored using an image system. The answers are scanned, and actual images of student responses are transferred directly to data files. The scanned images are then routed electronically to rater workstations. During the image scoring phase, the rater accesses the images and enters the scores directly into the database. This allows item-by-item scoring rather than the traditional scoring by block of items, thus improving accuracy and speed of scoring. The technology allows data and images to be effectively managed and securely maintained in a digitized data form rather than a paper format. Improvements to the image system have been accruing since its introduction in 1994.

Despite these innovations, rater training is still carried out using techniques similar to those used in the earlier paper-and-pencil scoring. Training packets of sample responses are assembled by hand and the trainer and raters work from these paper versions. There is also no mechanism at present that allows responses already scored (validity papers) to be dropped back into the queue to establish whether raters are still scoring accurately. Mechanisms do exist, however, to enable raters to score a set

of predetermined papers (calibration papers). These are usually administered after long breaks, again as a check for rater errors such as scoring drift.

Costs

Overall, the benefits of image scoring outweigh those of pencil-and-paper scoring. However, the costs of image scoring remain high, due mostly to the leasing of expensive equipment. The process itself has also led to some unwanted outcomes. For example, the scanning process is very sensitive. This means that student responses are picked up even when they are erased; thus responses may be scored that would not have been in paper-and-pencil scoring. In addition, some NAEP items are difficult to score on-line. Most of these are items that were developed prior to the introduction of the image scoring system. For example, templates (overlays) are used to score questions involving graphing. Since these templates are fixed on the screen, each student response has to be in exactly the same place on the screen so that it can match the template. If response and template do not match precisely, items may have to be scored by hand. Lastly, data may be missed because the area that is scanned (clipped) is predetermined, and students do not always write their responses in the designated areas.

Benefits

The image system has led to an array of improvements. Training can take place on one item at a time, and since items are not scored one after another in a student booklet, unwanted rater errors such as "halo effects" are eliminated. Responses can now be scored according to domain, so that raters can be grouped according to their qualifications. Since reliability data are present almost immediately, problems in the scoring guide and/or among the raters can be diagnosed and corrective action taken.

Remote-Site Electronic Scoring

The high interrater reliabilities attained in scoring recent NAEP assessments have demonstrated the advantages of item-by-item scoring. The current scoring model could be modified so that instead of the current scoring work flow, in which servers deliver work to multiple workstations in a single building, these servers would route the responses to be scored to a secure Internet site. Raters, sitting in their own homes or offices, would log on to the Internet site and, after some significant security log-on procedures, be presented with responses to be scored in a manner very similar to what is done in a scoring center.

In order to train raters at multiple sites and at various times, training documentation must be available on-line and would be delivered via encrypted Internet transmission. Once received, raters would print training materials (for example, rubrics and some prescored papers) in preparation for a real-time video conference/training session. The training session would be recorded to enable raters to refer to the material while scoring, to use again in the event that the item is rescored at a future time (that is, for trend studies), and for review by secondary researchers.

Video conferencing must be an interactive process, and to this end, a variety of personal input devices would be available to both raters and trainers. For example, large blackboard-sized or notebook-sized "dry erase" boards can be connected to workstations for transmission of information written on them. Questions by raters during training and scoring are an integral part of the scoring process, and these devices would allow raters to augment their questions with pictures and equations. Likewise, the trainers would be able to present their materials in a more graphic form for transmission to all raters. Additionally, point-to-point communication between an individual rater and the trainer using these devices would be a possibility, whether it be during the training process or later during actual scoring of an item. Picture-within-a-picture capability would allow simultaneous viewing of electronic images, the rubric, and the trainer.

Before a rater was cleared to score an item, he or she would take a qualification set(s) on-line; the results could be reviewed by the rater and would be also transmitted to the trainer for review and approval to score. Calibration sets and validity sets would be electronically delivered to raters to ensure that scoring proceeds in a manner consistent with training and the rubrics. Authorization to score could be suspended upon review of these prescored papers. Additionally, validity papers could be delivered in a way that is completely transparent to a rater and could be delivered individually or in sets. The frequency with which sets of papers are routed to a rater could be dependent upon interrater reliability with the flexibility to allow the trainer to override or trigger the delivery.

Since the trainer may be located at a very distant site, a rater must have the capability to electronically upload a student response (send a copy to a trainer) for simultaneous viewing and discussion. The trainer could then have the ability to annotate the response, add the response to the training materials, archive the paper as an unusual response, or electronically deliver the annotated paper to all raters scoring that item for review.

Distribution of items to readers can be accomplished in a work flow model similar to what is now done. However, the use of other types of text files (such as uuencoded files) of students' responses instead of the much larger image files, would save considerable disk storage space, file transmission cost, and ensure more reliable transmission of the data over a nationwide network. Additionally, the servers at the NAEP scoring Internet site would have the capability to randomly select individual papers for second scoring, choose which raters receive the papers, and send them one at a time instead of in packets.

Costs

The startup costs of establishing remote-site electronic scoring would initially be enormous due to setting up an infrastructure. In addition, raters would have to be in place with equipment that meets certain specifications, and factors such as training and security would have to be resolved.

Benefits

Under a remote-site electronic scoring system, scoring would be done in many locations by raters and monitored from a scoring Internet site. Thus, travel and relocation costs incurred by the scoring coordinators and trainers would be eliminated. The use of video conferencing would replace the necessity for face-to-face meetings and allow scoring coordinators to talk with raters, trainers, and each other from their own bases of operation. Likewise, the central scoring Internet site would have much lower overhead costs than a large, fully staffed scoring facility housing hundreds of raters. The electronic transmission of student responses to the scoring Internet site servers would eliminate the cost of mailing large numbers of documents to the scoring contractor's facilities and would allow the scoring coordinator direct access to files of student responses for construction of training sets. Once files were arranged in training sets, these sets could be electronically delivered to the appropriate, qualified raters without the need for producing costly photocopied sets. After completing scoring of a particular item, the data would be available for psychometric analysis.

Image Scoring Under the Alternate NAEP Designs

If NAEP continues to use constructed-response questions that cannot be scored automatically, then centralized and remote-site image scoring offer similar interactions with the various NAEP designs. Remote-site scoring may enable scoring to be done in a more timely manner since the number of raters would not be dependent on location. Thus the period between data collection and reporting would be reduced. The use of remote-site scoring may also impact the type of questions, since item-specific tools such as overlays (templates) may be too expensive to install at remote sites.

Automated Scoring

Since 1990, the number of open-ended questions in NAEP that ask for one-word answers, phrases, sentences, paragraphs, or essays has increased dramatically. This led, in 1994, to a fundamental change in approach to scoring, from paper-and-pencil scoring

to computer-based scoring. Scoring large numbers of student responses is very expensive and the only current way to achieve real savings is to reduce the number of constructed-response questions contained within an assessment.[3] This, however, may come at the cost of not being able to measure what is considered important in the various domains.

Since the advent of computer-delivered assessments in major testing programs (such as GMAT), there has been added impetus to think of a system that not only delivers the assessment but scores the open-ended items as well.[4] While there are currently no concrete plans to deliver NAEP assessments via the computer, they may well be delivered in this manner in the not-too-distant future. It is imperative, therefore, to investigate automated scoring methodologies that may allow for accurate scoring of NAEP responses, thereby positioning NAEP to take full advantage of the interdependency among computer-based test components.

Many automated scoring systems are currently being developed and/or evaluated.[5] The approaches are numerous and range from creating new item types and devising new methodologies to score the responses, to taking existing responses and adapting off-the-shelf software. While results are mixed, a number of examples will serve to illustrate the various approaches. Braswell and Jackson[6] created a free-response item type in mathematics for the PSAT program that required students to enter their answers on a grid. This proved to be very successful. Braun et al.[7] developed a computer science item for AP students that presented a faulty solution to a computer programming problem. This was successfully scored using a system called PROUST

---

[3] Lazer, S. (1996). *Reductions to NAEP costs.* Internal memorandum.

[4] Bennett, R., & Bejar, I. (1997). *Validity and automated scoring: It's not only the scoring.* Princeton, NJ: Educational Testing Service.

[5] Bennett, R., Gong, B., Kershaw, R., Rock, D., Soloway, E., & Macalalad, A. (1988). *Agreement between expert systems and human ratings of constructed-responses to computer science* (ETS-RR-88-20). Princeton, NJ: Educational Testing Service.
Kaplan, R., Burstein, J., Trenholm, H., Lee, C., Rock, D., Kaplan, B., & Wolff, S. (1995). *Evaluating a prototype essay scoring procedure using off-the-shelf software* (ETS-RR-95-21). Princeton, NJ: Educational Testing Service.
Burstein, J., & Kaplan, R. (1995). *GE FRST evaluation report: How well does a statistically-based natural language processing system score natural language constructed-responses?* (ETS-RR-95-29). Princeton, NJ: Educational Testing Service.

[6] Braswell, J. S., & Jackson, C. A. (1995). *An introduction of a new free-response item type in mathematics.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

[7] Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1989). *Developing and evaluating a machine-scorable constrained constructed-response item* (ETS-RR-89-30). Princeton, NJ: Educational Testing Service.

and MicroPROUST. Short responses stemming from items known as formulating-hypotheses items were examined by Kaplan and Bennett[8] and Burstein and Kaplan[9] using a program called FRST. It was found that FRST assigned scores to correct hypotheses that correlated highly with scores given by human raters; however, it also classified many incorrect answers as correct. Thus, in its present form, this methodology would be unsuitable for use in the field.

Since many testing programs now include an essay component, much effort has also been expended to find accurate ways of scoring essays automatically. For example, Kaplan et al.[10] took essay responses from the Test of Written English that is administered as part of the Test of English as a Foreign Language and used off-the-shelf grammar checking programs to evaluate them. (Essays in this particular program are holistically scored by hand on characteristics that include grammar, style, and the ability to organize and support ideas.) Approximately 30 percent of the essays analyzed could be scored correctly by computer. While encouraging, this percentage is also too low for consideration in the field. While natural-language processing programs to score essays are difficult to develop, researchers are making real advances as evidenced by the proposed use of automatic scoring to rate essays written by candidates taking the GMAT examination.

NAEP responses, while they represent the types of natural-language responses currently being investigated by researchers, do present special challenges. The current NAEP items that require short and extended constructed responses may be too difficult to analyze using current technology given the number of grammatical and syntactical errors that occur in student responses, especially at grades 4 and 8. Researchers tend to work with responses written by adults from programs that are high stakes. Thus, responses may be written more thoughtfully than those in NAEP, which is low stakes

---

[8] Kaplan, R. M., & Bennett, R. E. (1994). *Using the free-response scoring tool to automatically score the formulating-hypothesis item* (ETS-RR-94-08). Princeton, NJ: Educational Testing Service.

[9] Burstein, J. C., & Kaplan, R. M. (1995). *GE FRST evaluation report: How well does a statistically-based natural language processing system score natural language constructed-responses?* (ETS-RR-95-29). Princeton, NJ: Educational Testing Service.

[10] Kaplan, R. M., Burstein, J., Trenholm, H., Lee, C., Rock, D., Kaplan, B., & Wolff, S. (1995). *Evaluating a prototype essay scoring procedure using off-the-shelf software* (ETS-RR-95-21). Princeton, NJ: Educational Testing Service.

and does not generate individual scores. Research using actual NAEP items is sparse; however, a recent feasibility study did look at the automatic scoring of current NAEP mathematics items that asked for numeric answers, diagrams, or graphs.[11] Preliminary results are encouraging, given the fact that these items were not constructed with machine scoring in mind. Many items were amenable to generalization; thus, knowing the parameters for scoring would encourage item writers to create questions appropriate for automatic scoring. Difficulties encountered in the scoring process could have been eliminated by better defining where answers should be placed. For example, students could enter their numeric responses in a box, units could be preprinted so that only the number is scored, or if the question asks for an angle to be drawn, the line plus starting point can be specified. Thus, item specifications can be written that include the parameters essential for an item to be scored automatically.

The Role of Automated Scoring in NAEP

Several scenarios can be envisioned for the role of automated scoring in NAEP. As a first scenario, the assessment could include only those open-ended items that are amenable to accurate and reliable scoring by computer. This, however, would constrain the types of items in the assessment and may exclude part of the domain being tested. As a second scenario, several different types of open-ended items could be included, some of which would be scored automatically while others would be scored at terminals. To achieve this, the types of constructed-response items already seen in NAEP assessments would have to be augmented since current research clearly shows that the most successful automatic scoring involves questions that have been developed knowing the parameters of the automatic scoring methodology. In addition, responses to items in subject areas such as mathematics may be more amenable to automatic scoring than responses to items in other subject areas. The research conducted in mathematics, however, accounts only for correct answers. In the field, many answers

[11] Reid-Green, K.(1997). *Results of research into automatic scoring of NAEP mathematics test questions.* Princeton, NJ: Educational Testing Service.

are scored for partial understanding as well. Clearly, much research needs to be done if automatic scoring is to become a viable part of NAEP.

Costs

If constructed-response items are to remain an integral part of NAEP, the best approach to automatic scoring would be to develop items for which there is already proven success. Each subject area would vary in the types of items that would be appropriate. Some mathematics items can already be scored automatically.[12] In addition, other items could have been scored successfully if constraints had been put upon the area of response. Clearly some of the item types that were successfully scored in mathematics can be transferred across domains. For example, the 1990 science assessment contained figural-response items that are similar to those examined in the mathematics study. The item types that elicit student explanations and thus give evidence of student understanding may still be administered, but the numbers could be reduced as other types of items are considered whose responses can be scored automatically. Scoring rubrics may also take longer to prepare because when there are multiple correct student responses, these must be entered into the scoring system as models of response classifications.[13] Certain responses would be unscorable by machine and so would have to be scored by an individual. This would require spot checking to verify that scores are accurate. Special studies using automatic scoring and hand scoring would have to be performed in order to compare and evaluate the methodologies.

Benefits

Once the systems were worked out, responses could be scored much faster. Responses could be scanned into databases using currently available systems. A minimal number of raters would be needed. Since scoring would be carried out by

---

[12] Ibid.

[13] Burstein, J. C., & Kaplan, R. M. (1995). *GE FRST evaluation report: how well does a statistically-based natural language processing system score natural language constructed-responses?* (ETS-RR-95-29). Princeton, NJ: Educational Testing Service.

machine, factors such as halo effects, rater fatigue, or an ability of a rater to follow the scoring guide would not exist.

Interactions of Automated Scoring With Other Program Areas in the Current NAEP

If automated scoring were adopted, item development would be affected in a major way since new types of items would need to be developed. In addition, more extensive scoring rubrics might have to be developed since rubric data need to be entered prior to scoring. This would add to the cost of NAEP initially, but would realize real savings once the new item types and methodologies were in place. It would also allow for the creation of item variants. Analysis may occur in a more timely fashion if machine-scorable items replace many of the items currently scored by hand. This in turn would reduce the time between data collection and report preparation.

Automated Scoring in a Streamlined NAEP

The adoption of automated scoring would make it possible to maintain the proportion of constructed-response questions with machine-scorable items replacing many of the hand-scored items, provided the items were deemed appropriate for measuring the construct in question. In addition, the different item types could be grouped prior to administration. This would expedite scoring preparation.

Automated Scoring in a Modular NAEP

There could be modules of different types of NAEP items. For example, the core could consist only of machine-scorable items. Special modules could be constructed that included items of similar construct (including hand-scored items), thus streamlining processing, scanning, etc. (Parallel-forms NAEP could be handled in a similar manner.)

Rater Reliability

While rater reliability has improved since the advent of the image-based scoring system, rater errors do occur, and rater effects models are currently being researched to identify aberrant raters and detect types of rater errors. In addition to the identification of "problem raters," there are also statistical means that can be used to calibrate the scores of aberrant raters. The rater error issue will be examined from both perspectives.

The number of constructed-response questions in NAEP assessments has increased dramatically since 1990. Since responses in NAEP are judged using specific criteria, it is important that raters assign ratings accurately, in order to reflect what students in the U.S. know and can do. With the current NAEP scoring methodology, the speed of assigning scores has increased dramatically; thus, models to detect rater aberrance in a timely fashion need to be examined to ascertain the feasibility of inclusion during scoring. Several types of differential rater functioning over time (DRIFT) have been identified by researchers, namely, peeking, rebellion, fatigue, practice, differential centrality, differential extremism, regency, and primacy.[14] Many factors appear to come into play, depending on whether scoring takes place over several days or several weeks. Research, however, is still in its infancy. The approaches seen are based on a normative measurement framework; that is, raters are judged relative to the pool of other raters. Thus, if patterns are observed with one rater, it is not possible to tell if it is the rater or pool of raters who are manifesting the error. Therefore, while the question "Which raters look suspect?" can be answered, the question "Which raters are incorrect?" cannot.[15] Thus it is important to develop criterion-referenced models for assessing DRIFT. These could take the form of rater accuracy indexes that compare ratings obtained from operational raters to those

[14] Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation procedures. In A.R. Ciminero, K.S. Calhoun, & H. E. Adams (Eds.), *Handbook of Behavioral Assessment (2nd ed.)*, pp. 253-324. New York: John Wiley and Sons.
Wolfe, E. W., & Myford, C. M. (1997). *Detecting order-effects with a multi-faceted rating scale model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

[15] Wolfe, E. W., & Chiu, C. W. T. (1997). *Detecting rater effects with a multi-faceted rating scale model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

assigned by a committee[16] or new models that depict specific types of aberrant rating patterns that can be compared to normal patterns.[17]

If both scoring and analysis can use rater effects models, further advantages accrue. If such models are used in the Item Response Theory scaling of the data, a "severity" parameter can be estimated for each rater and conceivably for each item. The scoring results from different raters can be treated as belonging to separate groups. Evidence for different scoring standards among raters is indicated by differences in group means and standard deviations. If a response is rated by more than one rater, the scale-score of the respondent is the weighted mean of the standardized scale scores of each rater. The resulting scores can be adjusted for both the average severity and average variability of each rater.[18] Information about the magnitude of effects and the ability to identify items with larger rater effects can guide efforts to further improve scoring and/or training procedures.

The Current NAEP System

Currently, interrater reliability data are used to detect rater errors in NAEP. While this information is available early in the scoring process, it only reveals where splits occur; responses are then examined to determine which rater is at fault. Since scoring of one item can take from a number of hours to a week or more, it is important to establish procedures to address the concerns that arise from using raters to score responses. This has become more pressing since the advent of state testing, in that certain responses may take a week or more to score and the ability of raters to perform optimally during this time period is questionable without specific intervention. Thus, the better able the system is to detect types of rater errors, the more effectively they can be addressed. This would become even more important if NAEP were to adopt remote-site scoring models. When raters work together in a common location, scoring leaders

---

[16] Englehard, G. J. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56-70.

[17] Wolfe, E. W., & Myford, C. M. (1997). *Detecting order-effects with a multi-faceted rating scale model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

[18] Muraki, E. & Bock, R.D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data.* Chicago, IL: Scientific Software International.

can closely supervise the process. Because this oversight is impossible off-site, methods to detect aberrant raters' performance become more critical. The literature describes ways of addressing errors caused by such factors as fatigue.[19] Strategies for reducing these types of errors are already an integral part of NAEP scoring. Therefore, a more pressing need is to detect aberrant raters and pinpoint what the error is.

Costs

Models that allow early detection of aberrant raters need to be developed and tested in the field. Further, procedures for dealing with aberrant raters must be devised and put in place. Many of these procedures already exist, and if the type of error can be identified, remedial action that addresses the particular error can be implemented. At present the only approach is to retrain aberrant raters and hope that the error is addressed.

Benefits

Identification of aberrant raters can be done in a timely manner that allows for intervention and ultimately renders a more accurate measure of what students know and can do. This would become increasingly important if NAEP responses were scored on-line at remote sites.

Impact on Alternative Designs

The impact of incorporating methodologies that can pinpoint aberrant raters and diagnose which error is in evidence would be similar in any NAEP assessment that utilized constructed-response questions. Such methods would increase the reliability and validity of ratings, a goal that is critical if data are to inform policy.

[19] Ibid.

Recommendations for Constructed-Response Scoring

There are several recommendations about constructed-response scoring that issue from the investigations undertaken in this chapter. These are described below.

Existing image-processing scoring tools should be expanded to include enhanced training and monitoring capabilities. Such expansion represents a logical next step in the development of tools for computer-assisted scoring, and builds on current successes.

A feasibility and cost/benefit study of remote-site Internet-based scoring should be conducted. Such a study should consider fiscal, technical, quality, and security factors.

Automated scoring programs should be evaluated for possible use in components of NAEP. Research into such programs should involve identification of exercises amenable to computerized scoring, development and testing of machine-scorable performance items, and investigation of the effects of separating machine and human scorable exercises into different sections of cognitive instruments.

NAEP should research methods of statistical adjustment that would compensate for any effects introduced by rater unreliability. Such adjustments might at some point routinely be made as part of the NAEP scaling process.

This page intentionally left blank.

CHAPTER 9
ANALYSIS

EXECUTIVE SUMMARY

This chapter contains an examination of NAEP statistical analysis methodologies, and a discussion of ways in which these processes might be streamlined or improved. First, the chapter focuses on possible methods on reducing analytic complexity; this section discusses changes that may lead to real gains as well as those that are likely to have a negative impact on the system as a whole. Following this, the chapter examines analysis techniques that, while not increasing efficiency, would allow NAEP to yield more or different information. The final section considers major analysis changes that, while seeming to introduce efficiencies, would likely prove disastrous for NAEP. The following arguments and recommendations are included in this chapter:

- Expanding student testing time, while seemingly a way to reduce reliance on statistical models, would do irreparable harm to the NAEP system. Therefore, student testing time should not be increased.

- Research conducted for this paper shows that item parameters drift over time. Therefore, items should be recalibrated after each administration.

- NAEP should continue to cover broad content domains while limiting individual testing time. Accomplishing these goals implies matrix sampled instruments and continued use of Item Response Theory and marginal estimation techniques.

- NAEP should research methods that might be used off critical analysis paths to gain enhanced or different information about student performance.

This page intentionally left blank.

CHAPTER 9
ANALYSIS

- Eugene G. Johnson / James E. Carlson -

Introduction

In this chapter, we identify innovative analysis techniques that might lead to benefits for NAEP. The techniques we consider group into three sets. The first category of topics covers techniques that minimize analytic complexity and processing and thus lead to increases in the efficiency of analysis. These include an examination of the costs and benefits of **lengthening testing time**. In addition, there are analytic techniques that might bring reporting and cost efficiencies. These include **precalibration of items, two-phase assessment analysis**, and **market-basket based analyses**.

The second category of issues related to NAEP analysis concerns the search for techniques that are innovative and produce more or different information. In this general area, issues about the usability of new modes of data summarization and interpretation, such as **rule-space analyses,** will be discussed. Also, adjustments to **item response theory and other model-based procedures** should be investigated as possible sources of efficiency. Finally, given concerns over motivation and participation rates, any psychometric investigation must examine potential improvements in **nonresponse adjustments**. It is important to understand that not all these analysis procedures can be immediately implemented in a streamlined NAEP. Rather, they should be part of a carefully thought-out research plan. They should also be used where appropriate to meet specific program priorities. Complex psychometric procedures often increase analysis time. Although these procedures potentially increase the amount of information from the assessment, they should not be used as a part of the initial analysis of data. Rather, these techniques should be considered for subsequent investigations that are accomplished off the critical reporting path.

The final category of issues related to NAEP analysis concerns procedures which, while potentially simplifying analytic complexity in some instances, do so at the cost of less flexibility in other instances. These include the possibility of **eliminating IRT scaling** and **elimination of plausible values.**

Analysis Procedures in the Current NAEP

Since NAEP's inception, item pools have been too large to present all items to each student. The major reason for this is that the entire NAEP instrument requires many hours of testing; if students were asked to take entire assessments, overall results would be affected by fatigue, possibly limited motivation, and potential difficulties with nonresponse at the school and student level. On the other hand, it was widely felt that a set of items that could be fit into 45 to 60 minutes of testing time could not adequately cover the detailed and extensive NAEP *Frameworks.* Consequently, matrix sampling is employed to present every assessed student with a subset of the assessment items under a design where the entire item pool—deemed large enough to adequately cover the *Framework*—is presented to representative samples of students.

The basic information from an assessment consists of the responses of students to the items presented. The earliest assessments focused solely on item-level statistics for reporting and analysis. This is the most direct manner of presenting the assessment results, is the least complex in terms of analysis, and is the fastest in terms of processing time. However, because of the vast amount of information, separate results for each of the items in the assessment pool can hinder the comparison of the general performance of various population subgroups.

It rapidly became clear that some means of summarization over tasks was needed as a way of determining underlying patterns and trends in the data. An obvious measure of achievement within a domain of interest is the average item score across all presented items within that domain. The advantage of averaging is that it tends to cancel out the effects of idiosyncrasies in items which can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more

easily the general performances of subgroups. Finally, averaging is not much more complex or much slower than item-level reporting.

Despite these advantages, there are a number of significant problems with average item scores. First, the interpretation of these results depends on the selection of items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, because the average item score metric is related to the particular items comprising the average, direct comparisons in performance between subpopulations require that those subpopulations be administered the same set of items. Third, because this approach limits comparisons to average scores on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, the average item score provides no estimate of the distribution of proficiency in the population when each student is administered only a fraction of the items. Average item score statistics describe the mean performance of students within subpopulations, but provide little other information about the distributions of skills among students.

These limitations were overcome in the second design of NAEP implemented in 1984 by the use of Item Response Theory (IRT) scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale with this method, even though none of the respondents take all of the items within the pool.

Using the scale, it becomes possible to discuss distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables. While scaling is analytically more complex and takes more time than simple mean scores, the benefits of these procedures likely outweigh these costs.

Still more complex is the use of marginal estimation procedures to obtain estimates of subpopulation distributions and associations with other background

variables. Since NAEP continually changes its test length, test difficulty, and balance of content in order to satisfy the *Frameworks*, NAEP needs to use methods that can accommodate substantial updating from assessment to assessment while remaining sensitive enough to detect small but real changes in subgroup performance in situations of nonnegligible measurement error. Marginal estimation procedures, which estimate population characteristics directly from item responses, while complex, meet these needs and control for the variations in test properties caused by matrix sampling and changes in test length, content, and difficulty. Numerical approximations of the appropriate marginal estimation procedures can be obtained for a wide variety of analyses by constructing, from the results of an extensive marginal solution, plausible values of student proficiencies.

The essential idea of plausible values methodology is to represent what the true proficiency for an individual might have been, had it been observed, with a small number of random draws from an empirically derived distribution of proficiency values.[1] This distribution is conditional on the observed values of the assessment items and on an extensive collection of background variables for each sampled student. The draws from the conditional distribution can be considered to be representative values from the distribution of potential scale scores for all students in the population with similar characteristics and identical patterns of item responses. The several draws are different from one another in a way that quantifies the degree of precision (the width of the spread) in the underlying distribution of possible proficiencies that could have generated the observed performances on the items.

Plausible values technology was developed to satisfy a requirement of the NAEP *Frameworks* and the concomitant designs. It is important to note that the *Frameworks* and designs do not have to be constrained by the requirements of the scaling/plausible values techniques—the scaling/plausible values procedures are necessitated by the

[1] Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Education Statistics, 17,* 131-154.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Frameworks, data collection, and reporting requirements. Plausible values, once developed, allow for any analysis of the NAEP data that is consistent with the decisions made for reporting the data. Scaling and plausible values add limited time to analysis and reporting of initial results, but they substantially improve the response time for secondary reports. However, primary reporting might be accelerated at the expense of secondary reporting and analysis if specific priorities were made clear. This will be discussed below in the section about two-phase analysis.

Issues Related to Analysis

Lengthening Testing Time

An obvious way to increase test reliability while minimizing analytic complexity and processing time is to invoke a design in which lengthened NAEP forms are presented to students. Given forms of sufficient length, individual student proficiencies could be determined using standard IRT methods, thereby eliminating the need for the conditioning/plausible values step of analysis. ETS and ACT have already conducted an experiment where students were administered a double-length assessment booklet.[2]

As a partial check on the validity of the achievement levels set for the 1994 NAEP in geography, eighth-grade teachers on the panel which set the achievement levels were asked to classify their students on the basis of achievement level descriptions. These students were also administered a double-length form of the NAEP grade 8 geography assessment. This study thus provided data to allow the empirical evaluation of the need for plausible values for a test of this length. It also provided information about student fatigue effects. It should be noted that the students in this study were better prepared and more motivated than the general population of grade 8 students. Thus these students might be more willing to participate in the assessment and to try harder than the average student assessed in NAEP. Furthermore, the teachers and schools involved in the study were also potentially more

[2] Johnson, E.G., Liang, J-L., Norris, N., Rogers, A., & Nicewander, A. (1996, April). *Directly estimated NAEP scale scores from double-length assessment booklets—A replacement for plausible values?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

amenable to the extended testing time than would be the typical school. This could have an impact on school and student participation rates.

Even given the higher motivation of students, there was a noticeable fatigue effect across the 100 minutes of testing. The presence of the fatigue effect provides an underestimate of student performance averaging 3 scale score points lower for this sample of students than that which would be obtained from administering a standard length assessment of 2 blocks in 50 minutes of testing time and applying the plausible values machinery. Such a fatigue effect might well be more pronounced for less well prepared and less motivated students such as would be found in a typical assessment. For reference purposes, a three-point change in performance between two time points or between two subpopulations within a particular time point is generally statistically significant and viewed worthy of reporting.

Unfortunately, eliminating such a fatigue effect would appear to require either the application of model-based procedures (such as plausible values) or the administration of the lengthened instrument over two sessions, with consequent increase in administration expense and respondent and school burden.

Precalibration of Items

Since the first time that item response theory methods were applied to NAEP, the calibration of the items has taken place after the collection of the operational data. There are two reasons for this. First, the field test samples are too small and unrepresentative to allow for stable and accurate parameter estimates. Second, a nontrivial proportion of the items are changed enough between the field test and the operational assessment to render field test item parameters useless. Cost considerations led to these design elements: Increasing the size of the field test or including an additional field test to obtain final item parameter estimates would add significant expense. Nevertheless, if items could be calibrated before the assessment, there would be a decrease in the analysis time required after the operational assessment.

Precalibrated items are a possibility for both the comprehensive and the standard assessments. In order to be used for the comprehensive assessments,

precalibration would require a nationally representative field test of sufficient size for stable parameter estimates. The requirement of a nationally representative field test stems from NAEP's experience that item calibration needs to be performed on representative samples of the population to which inferences are to be applied. That is, the typical assumptions of local independence and invariance of item parameters across subpopulations have been found to only approximately hold in a large-scale assessment setting. NAEP scaling procedures on representative samples have evolved to produce IRT results which are close enough to be robust with respect to the context in which the data are collected and the inferences that are to be drawn.[3]

These same precalibrated items would then be used in subsequent cycles of a standard assessment which uses the same items as in the comprehensive assessment. In order for precalibration of items to produce acceptable results, the items' relative operating characteristics must remain constant over time. Studies of the NAEP reading anomaly suggest that this condition may be approximately true in adjacent assessments when the assessment forms are held constant.[4]

However, even though the long-term trend assessments hold their forms fixed, item parameter calibration for long-term trend data currently involves using data from both the current assessment and the previous assessment. The aim of this is to control for item parameter drift, that is, the differential change in the values of item parameters over time.[5] We have detected such drift in item parameter estimates in the course of analysis even in this most controlled situation.

Figures 9-1 and 9-2 and Tables 9-1 and 9-2 show the results of a study designed to evaluate the potential effects of item parameter drift in the case of the long-term trend assessments of mathematics and reading. The data analyzed for both subjects are the responses to the assessment instruments given by 13-year-old students who were

---

[3] Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Education Statistics, 17,* 131-154.

[4] Beaton, A.E., & Zwick, R. (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: Educational Testing Service.

[5] Donoghue, J. R., & Mazzeo, J. (1992, April). *Comparing IRT-based equating procedures for trend measurement in a complex test design.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

assessed in 1994. The assessment instruments and administration procedures have been held rigidly identical since the 1986 assessment. These data were analyzed in two ways. The first way used the item parameters from the 1994 assessment; the second used the item parameters from the 1990 assessment. The parameters from the two distinct calibrations were set to be on comparable metrics.

Figure 9-1 shows the comparison of the item difficulty parameter estimates from the 1990 and 1994 administrations of the long-term trend assessment of mathematics. Figure 9-2 provides the same comparison for the long-term trend reading assessment. While the difficulty estimates for the mathematics data appear fairly similar between 1990 and 1994, there are a number of reading items which display substantial item parameter drift since 1990.

Table 9-1 shows the result of estimating the mathematics proficiency distribution of the 1994 age 13 students using the 1990 item parameters instead of the 1994 parameters. Shown in the table are the means, standard deviations, and percentiles of subgroup proficiency distributions, based on the 1990 item parameters, minus the same statistics for the proficiency distributions based on the 1994 item parameters. The standard errors of the differences are given in parentheses. While none of the differences exceed their standard errors, it is worth noting that reported results would be somewhat different if the 1990 parameter estimates were used rather than the 1994 estimates.

Table 9-2 shows equivalent information for the 1994 age 13 long-term trend assessment in reading. Again, none of the differences are statistically significant. However, there are a number of cases where reported results would be noticeably different. For example, if the 1990 item parameters were used, the Hispanic mean would be about 2.3 points higher than the reported value; the private school mean would be 2.1 points lower. Aggregated over several assessments, such drifts might be sufficient to turn a nonsignificant trend into a significant trend. Also of interest is the fact that the overall 95th percentile would be 4.5 points lower using the 1990 item

Figure 9-1: Comparison of Item Difficulty Parameter Estimates from the 1990 and 1994 Mathematics Long-Term Trend Assessments

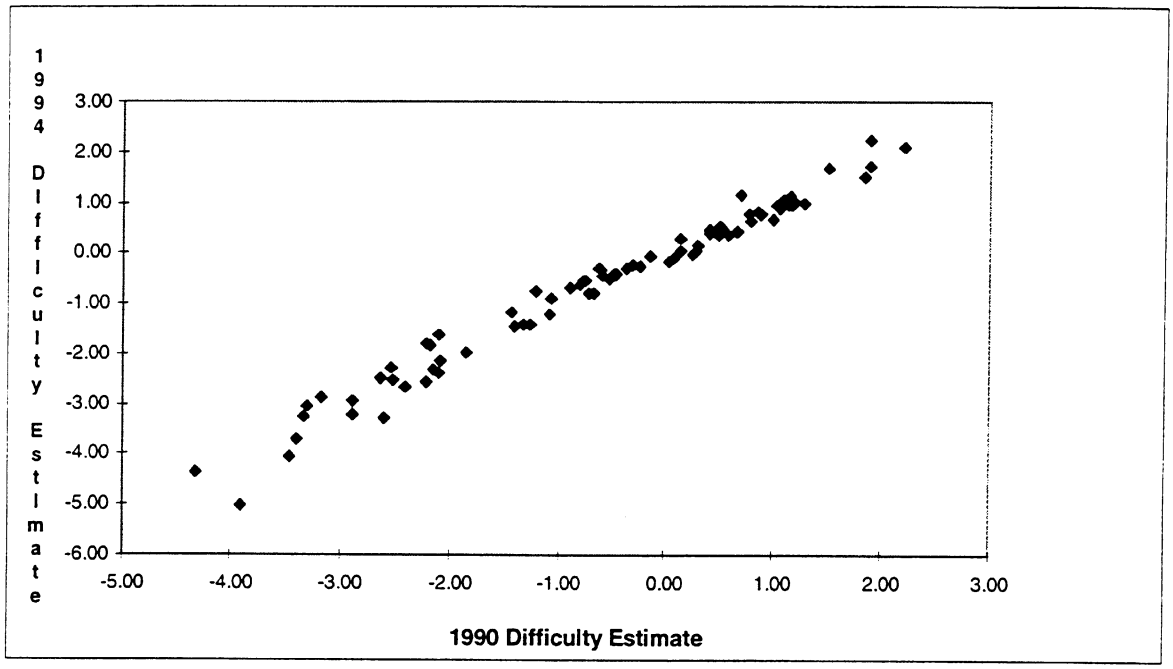


Figure 9-2: Comparison of Item Difficulty Parameter Estimates from the 1990 and 1994 Reading Long-Term Trend Assessments

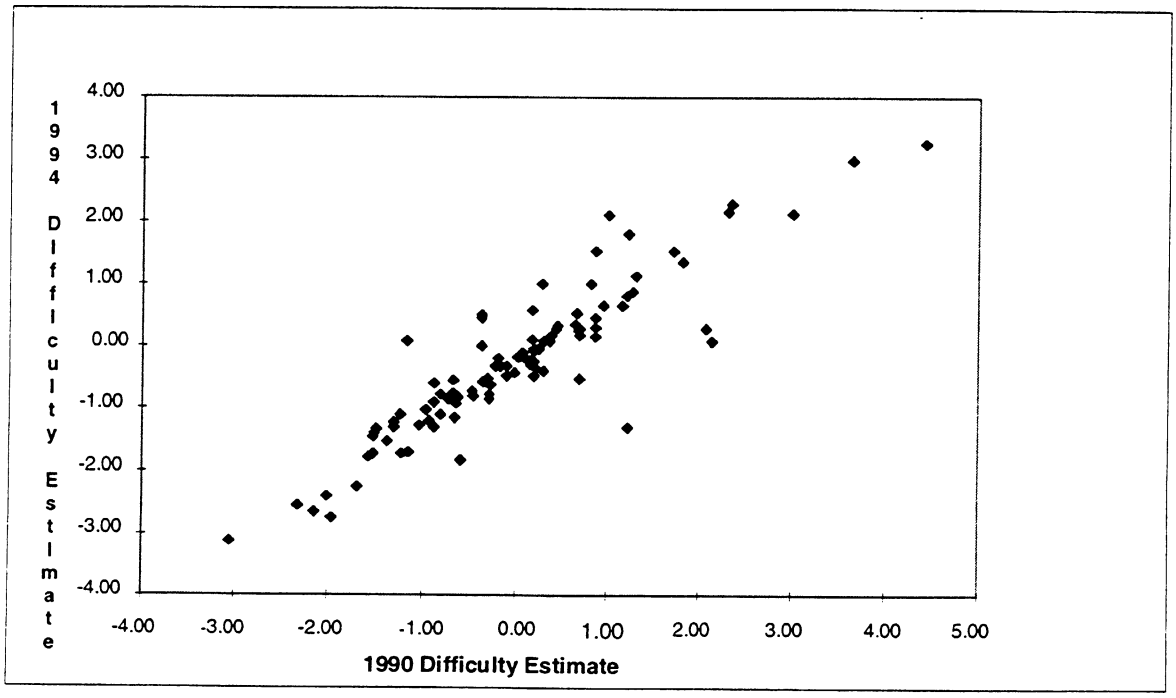


Table 9.1: Comparison of the Mathematics Proficiency Distributions for 1994 Age 13 Students Estimated with the 1990 and 1994 Item Parameters Differences of Means, Standard Deviations, and Percentiles Based on 1990 Parameters Minus Those Based on 1994 Parameters (Standard Errors in Parentheses)

	MEAN	STD DEV	5TH	10TH	25TH	50TH	75TH	90TH	95TH
-- TOTAL --	0.161(1.45)	0.174(0.76)	0.131(2.44)	-0.273(1.87)	-0.316(1.78)	0.181(1.33)	0.435(1.85)	0.397(1.79)	0.252(2.23)
SEX									
MALE	0.147(1.79)	0.116(1.17)	0.178(3.39)	-0.290(3.16)	-0.486(2.44)	0.130(1.58)	0.618(2.08)	0.522(2.79)	0.226(2.78)
FEMALE	0.176(1.49)	0.237(0.74)	0.064(2.33)	-0.295(2.64)	-0.414(1.19)	0.069(1.49)	0.556(2.07)	0.933(2.21)	0.933(3.30)
REGION									
NORTHEAST	0.269(2.06)	0.312(0.97)	-0.879(2.76)	-0.120(2.67)	-0.008(2.63)	0.463(2.64)	0.546(3.33)	0.777(3.90)	0.786(3.22)
SOUTHEAST	0.415(2.84)	0.247(1.02)	0.465(4.80)	0.004(4.56)	-0.507(4.03)	0.204(4.19)	0.925(2.57)	0.784(2.73)	0.450(3.50)
CENTRAL	-0.083(4.81)	0.049(2.18)	-0.140(8.53)	-0.315(6.23)	-0.146(5.27)	0.018(4.71)	-0.047(4.45)	0.091(4.05)	-0.257(5.88)
WEST	0.119(2.42)	0.161(1.41)	0.115(4.78)	-0.246(4.54)	-0.444(2.93)	-0.037(2.32)	0.538(2.38)	0.361(3.21)	0.494(2.74)
RACE/ETHNICITY									
WHITE	0.146(1.38)	0.276(0.87)	-0.654(1.78)	-0.459(2.28)	-0.202(1.56)	0.331(1.50)	0.491(1.98)	0.413(1.60)	0.413(2.60)
BLACK	0.400(5.02)	0.074(3.06)	0.908(9.09)	0.743(4.91)	0.217(5.15)	-0.014(3.62)	0.581(4.75)	1.197(5.59)	1.462(11.88)
HISPANIC	-0.013(2.34)	-0.010(1.52)	0.979(5.05)	0.139(3.74)	-0.339(3.69)	-0.459(2.72)	0.206(2.92)	0.184(2.62)	0.377(7.72)
ASIAN	-0.076(5.38)	0.160(2.14)	-0.390(15.79)	-1.446(7.37)	0.050(5.02)	0.767(10.55)	-0.057(6.92)	-0.651(3.96)	-0.945(7.22)
PACIFIC ISLANDER	0.086(4.69)	0.072(5.03)	0.863(13.21)	-0.932(11.94)	-0.490(15.34)	-0.833(6.71)	0.710(5.67)	-0.390(12.89)	-1.066(23.56)
AMERICAN INDIAN	0.949(11.07)	0.308(10.28)	0.057(45.70)	0.428(36.81)	1.251(19.04)	0.003(9.00)	1.171(11.18)	2.395(8.02)	2.594(15.86)
PARENTS' EDUCATION LEVEL									
LESS THAN H.S.	0.042(2.99)	0.035(2.39)	0.444(12.44)	0.395(6.64)	-0.056(5.31)	-0.127(5.12)	-0.051(4.53)	0.321(2.84)	0.492(3.05)
GRADUATED H.S.	0.109(1.49)	0.102(1.25)	0.379(2.19)	-0.022(2.60)	-0.469(2.01)	-0.072(1.54)	0.765(1.86)	0.320(2.37)	0.491(3.81)
SOME EDUC AFTER H.S.	0.115(2.28)	0.331(1.24)	-0.360(4.09)	-0.442(3.86)	-0.391(3.43)	-0.095(2.50)	0.360(2.44)	0.278(3.57)	0.759(4.52)
GRADUATED COLLEGE	0.216(1.68)	0.222(0.85)	-0.500(3.56)	-0.626(2.59)	-0.143(2.07)	0.556(1.30)	0.552(2.47)	0.504(2.63)	-0.061(2.12)
UNKNOWN	0.258(3.41)	-0.104(2.18)	0.425(7.55)	0.859(6.47)	0.346(4.32)	-0.235(6.63)	0.093(3.32)	0.629(7.19)	0.409(8.77)
TYPE OF SCHOOL									
PUBLIC	0.136(1.56)	0.133(0.76)	0.187(1.96)	-0.327(1.76)	-0.314(1.45)	0.217(1.43)	0.521(1.94)	0.443(2.31)	0.514(2.37)
PRIVATE AND CATHOLIC	0.353(3.48)	0.432(3.87)	-0.532(6.24)	-0.090(6.94)	0.179(3.92)	0.499(4.17)	0.909(4.58)	1.170(6.10)	0.699(3.84)
PRIVATE ONLY	0.371(9.51)	0.539(3.31)	-2.200(18.44)	-0.913(46.62)	0.521(5.11)	1.380(11.13)	0.616(12.99)	0.084(10.30)	0.674(8.22)
CATHOLIC ONLY	0.349(3.96)	0.410(4.65)	-0.272(7.89)	-0.374(6.57)	0.208(5.91)	0.406(4.26)	1.335(4.89)	2.165(11.74)	0.781(6.85)
TYPE OF LOCATION									
CENTRAL CITY	0.274(2.98)	0.110(1.29)	0.663(2.92)	0.150(3.47)	-0.040(3.35)	0.131(3.20)	0.821(2.71)	0.685(5.22)	0.971(6.44)
URB FRINGE/LARG TOWN	0.137(2.24)	0.290(1.32)	-0.619(4.29)	-0.461(4.06)	-0.263(2.83)	0.185(2.65)	0.576(2.70)	0.526(2.36)	0.602(2.71)
RURAL/SMALL TOWN	0.064(2.49)	0.117(1.12)	0.057(4.89)	-0.301(5.22)	-0.311(3.57)	0.112(2.27)	0.409(2.71)	0.536(4.06)	-0.182(2.65)

Table 9.2: Comparison of the Reading Proficiency Distributions for 1994 Age 13 Students Estimated with the 1990 and 1994 Item Parameters
Differences of Means, Standard Deviations, and Percentiles
Based on 1990 Parameters Minus Those Based on 1994 Parameters
(Standard Errors in Parentheses)

	MEAN	STD DEV	5TH	10TH	25TH	50TH	75TH	90TH	95TH
-- TOTAL --	-0.062(1.16)	-2.551(0.91)	5.090(5.46)	3.651(2.22)	1.733(1.59)	-0.618(1.46)	-1.730(1.90)	-3.266(1.87)	-4.495(2.03)
SEX									
MALE	0.475(1.73)	-2.419(1.14)	4.791(5.03)	3.930(3.61)	1.908(2.37)	-0.420(2.42)	-1.367(1.83)	-2.942(3.01)	-3.289(3.13)
FEMALE	-0.631(1.48)	-2.573(1.17)	5.591(3.82)	2.606(3.96)	0.577(2.03)	-0.983(1.80)	-2.233(2.37)	-3.545(3.24)	-4.605(3.38)
REGION									
NORTHEAST	-1.151(2.86)	-2.271(1.57)	3.164(5.10)	2.775(8.70)	1.122(4.83)	-1.116(3.28)	-3.505(3.40)	-3.956(3.42)	-4.300(3.08)
SOUTHEAST	0.517(3.44)	-1.937(1.78)	5.598(8.73)	4.128(3.64)	1.376(3.29)	0.738(4.06)	-1.229(4.96)	-1.791(4.31)	-3.496(7.29)
CENTRAL	0.192(4.45)	-2.542(2.62)	6.174(12.57)	4.580(9.16)	1.837(6.73)	-0.424(4.58)	-1.158(4.29)	-2.778(2.84)	-4.578(3.07)
WEST	0.030(2.86)	-2.995(2.13)	4.891(6.44)	4.389(6.68)	1.511(5.10)	-0.262(4.25)	-1.726(3.58)	-3.887(3.72)	-4.277(3.92)
RACE/ETHNICITY									
WHITE	-0.721(1.37)	-2.269(1.16)	3.658(3.46)	2.231(2.92)	0.600(1.43)	-1.044(1.99)	-1.817(2.29)	-3.412(1.81)	-4.416(2.35)
BLACK	1.284(2.87)	-2.196(2.17)	3.978(7.28)	4.151(5.73)	4.293(6.64)	0.229(3.75)	-0.749(4.92)	-1.717(3.79)	-2.878(6.82)
HISPANIC	2.267(3.02)	-2.405(2.88)	5.029(13.24)	5.743(6.15)	3.193(6.19)	3.025(5.30)	0.596(4.22)	-0.059(6.93)	-3.127(10.01)
ASIAN	-0.740(6.28)	-3.619(4.25)	0.002(14.51)	0.997(14.28)	4.090(11.03)	2.671(9.88)	-5.682(11.24)	-7.884(12.44)	-7.905(7.66)
PACIFIC ISLANDER	0.732(6.88)	-2.696(5.25)	3.208(12.23)	6.849(14.21)	1.261(9.65)	-0.079(11.21)	-3.239(9.28)	-2.272(9.12)	-6.337(24.99)
AMERICAN INDIAN	3.380(16.70)	-4.913(13.48)	10.954(55.94)	19.606(56.99)	-6.504(29.17)	16.964(98.91)	3.344(37.98)	-2.705(74.98)	-9.123(28.39)
PARENTS' EDUCATION LEVEL									
LESS THAN H.S.	1.883(3.53)	-2.276(3.53)	8.275(13.80)	4.809(6.93)	4.874(10.19)	2.454(5.26)	0.899(6.20)	0.160(12.14)	-4.145(11.51)
GRADUATED H.S.	0.531(1.97)	-2.253(1.53)	4.675(6.04)	5.815(4.62)	2.750(3.57)	0.011(2.92)	-1.405(2.74)	-1.492(3.66)	-2.783(6.19)
SOME EDUC AFTER H.S.	-1.059(2.70)	-0.628(2.51)	-0.832(9.43)	-0.921(7.04)	-1.389(4.52)	-0.751(5.96)	-2.095(3.44)	-1.038(6.95)	1.346(8.45)
GRADUATED COLLEGE	-1.042(1.66)	-2.472(1.25)	3.234(3.76)	1.933(4.00)	0.846(3.06)	-0.795(2.17)	-2.852(2.12)	-4.262(2.38)	-5.116(2.48)
UNKNOWN	2.478(4.40)	-2.347(3.06)	10.716(11.45)	5.426(9.68)	2.954(10.77)	1.018(5.88)	1.092(7.17)	1.490(6.89)	-0.458(7.01)
TYPE OF SCHOOL									
PUBLIC	0.043(1.29)	-2.511(1.01)	5.456(4.88)	4.235(2.52)	1.618(1.91)	-0.425(1.93)	-1.818(1.98)	-3.437(2.19)	-3.937(2.98)
PRIVATE AND CATHOLIC	-0.903(4.06)	-2.708(4.37)	0.677(19.83)	2.020(6.59)	0.320(4.28)	-0.749(5.32)	-2.904(4.75)	-4.053(4.36)	-4.160(8.81)
PRIVATE ONLY	-2.148(6.81)	-2.181(3.99)	4.176(23.12)	0.585(10.83)	-0.867(9.85)	-2.151(12.25)	-5.962(9.69)	-3.507(10.73)	-3.323(17.60)
CATHOLIC ONLY	-0.550(4.43)	-2.766(5.26)	0.953(16.38)	0.887(7.56)	0.820(5.80)	-0.735(3.99)	-2.344(4.22)	-2.810(5.84)	-3.717(6.53)
TYPE OF LOCATION									
CENTRAL CITY	0.616(2.94)	-2.473(1.48)	4.544(3.91)	4.853(3.13)	2.536(2.59)	0.555(3.60)	-1.137(4.59)	-2.959(4.91)	-3.829(5.57)
URB FRINGE/LARG TOWN	-0.626(2.51)	-2.592(1.83)	3.655(6.45)	2.474(5.61)	0.931(3.69)	-1.146(2.87)	-2.031(1.51)	-3.171(2.66)	-4.236(6.22)
RURAL/SMALL TOWN	-0.008(3.69)	-2.392(2.30)	4.860(12.65)	3.582(7.63)	0.939(5.09)	-0.291(4.03)	-1.565(4.91)	-3.211(3.38)	-4.317(5.41)

parameters. Since the advanced achievement level is often near the 95th percentile, this result suggests a potential impact in the setting and reporting of achievement levels.

The implication of this study is that, while not of a statistically significant magnitude, indications of item parameter drift were observed over a four year period for assessment instruments and administration procedures constrained to be identical. Furthermore, these instruments contain very few open-ended items, and the open-ended items that are included are all short answer. It is a distinct possibility that item parameter drift would become a more severe problem in an assessment instrument that contained a substantial proportion of extended constructed-response items, due, in part, to variability of scoring over time.

Since the item parameter calibration only takes a few weeks time, it may be more prudent to conduct a recalibration after the collection of the data. Such a calibration would be very quick in a situation where the items have been previously calibrated, since a large portion of the time spent in item calibration in current NAEP is spent resolving problems with new items which had not been previously calibrated.

Two-Phase Analysis

In our most recent proposal for the scoring, analysis, and reporting of NAEP, we recommended a two-phase analysis system as a way to ensure the rapid production of initial results. By restricting initial analysis to student-level data, we felt that it would be possible to accomplish rapid reporting. The second phase of analysis would involve the more complex and problematic data (including special studies and teacher and school questionnaires) that tend to slow down reporting. While the results from the second wave of analysis would be statistically equivalent to those from the first wave, for any subpopulations specifically included in the first-phase analysis, the results might appear slightly different. We do not view this as a flaw in the two-phase analysis system. An alternative is to use marginal maximum likelihood methods for the first wave of reporting and to provide a program designed to conduct marginal maximum likelihood analysis for secondary analysis. The result would be that no official set of

plausible values would be released, and primary and secondary users would be responsible for their own subsequent analyses. Any of their analyses that included the initial variables would generally also give slightly different results, but the appearance of conflict would be avoided because no "official," seemingly "more correct," analysis would follow. Secondary users could recreate the "official" initial numbers using the "official" model setup if they wanted to. Costs of such a strategy are discussed in a later section.

It should be noted that the utility of a two-phase analysis is greatly reduced under a design where the teacher questionnaire is decoupled from the student questionnaire.

Market-Basket Based Analysis

Variants of market-basket reporting must also be considered in the quest for improving the speed and usefulness of reporting. (See Chapter 10 for a more lengthy discussion of market-basket reporting.) It is important to realize that the term "market-basket" has a variety of meanings. At one extreme, it is simply a mechanism for reporting results from an instrument that is identical to the current NAEP instruments. At the other extreme, it is a core of booklets, each designed to be rigidly parallel so that results can be reported on a simple, total-booklet score metric. Reporting results from such an assessment would be very fast, but the speed would come at a cost.

The latter approach has both costs and benefits. Benefits include fast and clear reporting. Costs include the constraints associated with the construction of rigidly parallel forms; the difficulty of including performance items; and the need for a pilot test and a large-scale field test before the actual assessment. In addition, market-basket approaches may not be equally applicable to various subject areas.

Rule-Space Analysis

Tatsuoka's rule-space model[6] is an analysis technique with which ETS has been successfully experimenting. The rule-space model uses information about the cognitive and skill requirements of the assessment tasks to categorize students into groups based on what they know and can do. This is a probabilistic approach with the purpose of identifying students' state of knowledge or skills and is based on an in-depth analysis of the task's cognitive requirements or attributes. After the cognitive states (ideal-item-score patterns) are determined, actual item response patterns of students can be analyzed and mapped onto the cognitive states. Once students' item response patterns have been classified, attributes that a given examinee or group of examinees have mastered at a specified probability level can be identified.

New Item Response Theory Applications and Other Model-Based Procedures

It is worth considering IRT procedures that include additional parameters to identify and control for item drift and, in the case of professionally scored constructed-response items, to identify and control for rater effects. It is also worth considering alternative ways of measuring differential item functioning (DIF).[7]

DIF occurs when an item is differentially difficult for one group relative to another, when both groups have been matched on an overall measure of ability. In examining DIF, researchers[8] have determined circumstances where the probabilities of false positives (that is, declaring DIF where there is none) are higher than the nominal

---

[6] Tatsuoka, K. (1990). Toward an integration of item response theory in cognitive area diagnoses. In N. Frederiksen, R.L. Glaser, A. M. Lesgold, and M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition*. Hillsdale, NJ: Lawrence, Erlbaum Associates.

[7] Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*, 231-251.

Kulick, E., Donoghue, J. R., & Allen, N. L. (1995, April). *Subscale-level DIF analyses based on complex samples of items.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Allen, N. L., & Donoghue, J. R. (1994, April). *DIF analysis based on complex samples of dichotomous and polytomous items.* Paper presented in the symposium "DIF for Polytomously Scored Items" at the annual meeting of the American Educational Research Association, New Orleans.

Isham, S. P., & Donoghue, J. R. (1995, April). *An investigation of the sampling distributions of measures of IRT Drift/DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

[8] Chang, H., Mazzeo, J., & Roussos, L. (1997). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement*, in press.

values, and circumstances where the probabilities of false negatives are also different from the nominal values.

Issues of fit of the IRT model to the data can sometimes be cast into the framework of DIF. For example, item parameter drift may be due to differential item functioning (DIF) between two assessments. This type of DIF has been parameterized in the dichotomous item response model.[9] The extension to the polytomous item response model is straightforward[10] although there are issues of indeterminacy of the parameters of the full model.

Methods for identifying individuals whose patterns of response do not accord with those of other students in the assessment should be considered. These person-fit statistics could be used to identify students whose response patterns are atypical due, perhaps, to differential motivation. They would also be potentially useful for identifying students who, while they were assessed with various accommodations, could nevertheless be combined with the main population for reporting purposes.

Finally, most of the NAEP assessment subject area framework documents specify that data be reported along multiple dimensions. In current operational NAEP analyses the items created for each such dimension are separately scaled, unidimensionally, using a combination of the PARSCALE [11] and BILOG[12] computer programs. This PARSCALE/BILOG program allows for the simultaneous unidimensional scaling of dichotomously and polytomously scored items. There is an underlying assumption that, in responding to the items, the interaction of the assessment population members with the items is truly unidimensional. Although the

---

[9] Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275-285.
Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory.* New York: Springer-Verlag.

[10] Muraki, E. (1997). *Application of multigroup polytomous item response models to differential item functioning.* Unpublished manuscript. Princeton, NJ: Educational Testing Service.

[11] Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks.* Chicago: Scientific Software International.

[12] Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models (2nd ed.).* Chicago: Scientific Software International.

fit of the items to the model is systematically studied for each assessment, and the fit has been judged acceptable for almost all items used to date, there is the possibility that even better fit could be established by using multidimensional scaling.

Muraki has developed software (POLYFACT)[13] for scaling item response data assuming that assessment respondents interact with the items using several different dimensions. The model allows both dichotomously and polytomously scored items to be scaled. Muraki and Carlson[14] described, and illustrated with NAEP 1992 writing assessment data, the use of the multidimensional model for fitting item response data using POLYFACT. They used 1992 fourth-grade NAEP writing assessment data collected from 9,136 students, each of whom wrote responses to two writing prompts. The prompts, following usual NAEP practice, were administered to students according to a Balanced Incomplete Block design. There were a total of nine different prompts in the data analyzed in this study. The prompts were designed to measure the three purposes of writing (narrative, informative, and persuasive), and three of the nine prompts in the analyzed data represented each of these purposes. Although results showed that there was a strong first dimension, thereby justifying the unidimensional scaling used in operational NAEP, specifying an additional dimension significantly (chi-squared goodness-of-fit statistic) improved the fit of the data to the model. Systematic patterns of IRT discrimination parameters on the two dimensions were observed. These patterns showed that the three purposes of writing could be interpreted as separate (but correlated) dimensions.

In addition to the aforementioned model, Muraki, Carlson, and Bolt[15] developed another IRT model that allows for confirmatory analysis of item response data assumed to be multidimensional.

---

[13] Muraki, E. (1993). *POLYFACT [Computer program].* Princeton NJ: Educational Testing Service.

[14] Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73-90.

[15] Muraki, E., Carlson, J.E., & Bolt, D.M. (1996). *Multidimensional partial credit model: Confirmatory application.* Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Performance on an assessment item may depend simultaneously on one or more content-defined proficiencies and/or one or more process-defined proficiencies. Such proficiencies may be specified in NAEP assessment frameworks. As further experience is gained with the multidimensional IRT models described above, it may become feasible and desirable to use them in NAEP operational scaling and analysis to account for proficiencies described in the assessment frameworks. Further research must first be conducted in subject areas other than writing, and directed at implementing the multidimensional scaling results in NAEP operational analyses that depend on the scaling.

Nonresponse Adjustments

Research with NAEP data has shown that student nonresponse to cognitive items is an increasing source of concern. In particular, nonresponse to constructed-response items can be markedly higher than for multiple-choice items. Using homogeneous blocks (with respect to item format) appears to diminish this effect, possibly because the opportunity to pick and choose in responding among items of different formats is not available. Current analysis practice is to treat the omitted items as either missing information or as incorrect. Both of these choices are probably suboptimal.

Techniques Which May or May Not Lead To Efficiencies or Cost Savings

For the reasons given at the beginning of this chapter, the current NAEP psychometrics involve the use of complex scaling models followed by the use of plausible values methodology to estimate features of the proficiency distributions of subpopulations. Because both of these procedures are psychometrically sophisticated and since both procedures admittedly take time and resources to implement, our report examines the possibility of their simplification or elimination. However, for the reasons given in the introduction to this section, we believe that something very similar to these

techniques must be employed unless there is a radical change in the instrumentation, administration, and testing conditions of NAEP. In this section, we discuss the possibility of eliminating the IRT model altogether and eliminating conditioning.

Eliminate IRT Scaling

Current NAEP technology uses multiple-imputation methodology incorporating IRT scaling of items, and background information on the student, the school, and the teacher, to estimate distributions of proficiency. Although NAEP does not report individual students' test scores, the use of such scores to estimate mean proficiencies of various groups of students is possible. Through the use of rigidly parallel forms, the distribution of test scores could be presented.[16] On the face of it, this would imply that IRT scaling would no longer be needed.

Test scores, as usually conceived, are combinations of scores on the items of an assessment instrument. On an instrument comprised solely of dichotomously scored items, one common scoring procedure is to denote incorrect item scores as zeros, correct scores as ones, and the test score as the sum of the item scores. This is usually referred to as the number-right score. Other suggestions have been made, including various schemes of weighting each item score before summing to form the test score.[17]

Test developers often assume that each item contributes equally to the test scores when number-right scoring is used. Exactly what is meant by "contribute" in this context is ambiguous. Occasionally different items' (or sets of items') scores are multiplied by some constant before summing, under the assumption that this weighting changes the contribution to the test score. The following quote from Gulliksen correctly specifies the fact that these kinds of intuitive weights often used in test assembly are essentially irrelevant to the "contribution" of items or subtests to the total score.

---

[16] As argued elsewhere, however, the ability to cover an extensive framework is seriously compromised with the use of rigidly parallel forms, particularly when performance tasks are to be included.

[17] Gulliksen, H. (1987). *Theory of mental tests.* (Reprint of work originally published in 1950). Hillsdale, NJ: Lawrence Erlbaum Associates. Chapter 18.

In most amateur discussions of weighting of tests the first factors considered are the number of items in the test and the average magnitude of the score. It is believed, for example, that if gross scores are added, the effect will be to give a 100-item test twice the weight of a 50-item test. That such is not the case can be seen for example by assuming that the 100-item test was a very easy one on which everyone obtained scores ranging from 95 to 100. Adding scores on this test to a student's record would then, at the most, make a 5-point difference in the total score. If, on the other hand, the 50-item test were composed of fairly difficult items and were fairly reliable, it could easily be that scores on it would range from 20 to 50. In other words, adding this test would make a 30-point difference in extreme cases, and a 10- or 20-point difference in the majority of cases, so that the total score would agree rather closely with the score on the 50-item test and not correlate with the score on the 100-item test.[18]

When polytomously-scored items are included in an assessment instrument the most common extension of the number-right scoring procedure is to score each polytomous item from zero to some positive integer and simply add these item scores, along with the dichotomous item scores, to form the test score.

Contributions of items to total score

The question of interest is, How should we combine scores on individual items or sets of items (subtests) to form an appropriate total test score? In particular, how should we form a total test score from N multiple choice items and M performance tasks? The answer to this question is not an easy one for several reasons.[19]

Gulliksen's above-cited paragraph would suggest that "contribution" be defined in terms of contribution to variation in scores. Variation is usually represented by variance. It is easy to show that when combining scores (whether item scores or subtest scores) the variance of the composite is equal to the sum of the variances of each component plus twice the sum of the covariances between each pair of component

[18] Ibid., pp. 36-37.

[19] Ibid., Chapter 20.

scores. This fact leads many to define the variance of a component plus the sum of all covariances involving that component as the contribution of a component to the composite.

For example, the variance of a composite test, C, made up of two subtests, X and Y, is the sum of the variances of the two subtests and twice the covariance between them. The formula for the relationship is

$$\sigma_C^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \qquad (1)$$

where σ_C^2, for example, represents a variance and σ_{XY} a covariance between X and Y. Using the aforementioned definition we would define the contributions of subtests X and Y as

$$\begin{aligned} X &: \sigma_X^2 + \sigma_{XY} \\ Y &: \sigma_Y^2 + \sigma_{XY} \end{aligned} \qquad (2)$$

respectively, so that the two contributions sum to the composite variance. A problem with this definition is that the covariance term in (1) is arbitrarily divided into two equal parts in the definitions in (2). Note that the covariance is the product of the two standard deviations and the correlation coefficient,

$$\sigma_{XY} = \sigma_x \sigma_Y \rho_{XY} . \qquad (3)$$

One could argue that the covariance term ought to be divided into two components in the ratio of the standard deviations, rather than being divided by two. This would yield "contributions" in the form

$$\begin{aligned} X &: \sigma_X^2 + \sigma_X (2\sigma_X \sigma_Y \rho_{XY}) / (\sigma_X + \sigma_Y) \\ Y &: \sigma_Y^2 + \sigma_Y (2\sigma_X \sigma_Y \rho_{XY}) / (\sigma_X + \sigma_Y) \end{aligned} . \qquad (4)$$

Although these expressions may appear somewhat complex they, like the expressions in (2), sum to the right-hand side of (1). The important point is that the partitioning in (4) is every bit as defensible as that in (2).

Various other definitions of contribution of a component to a composite have been proposed and studied.[20] As mentioned by Carlson, "there is a lack of agreement among the different writers as to what is meant by the contribution of a predictor variable to a criterion variable, or a part score to a composite."[21] There is some agreement in the case that all of the components correlate zero with one another. In that case all of the covariance terms become zero. For test items this would occur only if each item represented a unique dimension with no single proficiency variable underlying the set of items—in other words, a poorly constructed test.

IRT Scoring

Although NAEP does not report scores of individual students, IRT scaling is used, as mentioned above, as one component in the estimation of score distributions. As has been pointed out by Lord, the number-correct scores yield a sufficient statistic for proficiency estimation only in the case that a one-parameter IRT model fits the data for each item.[22] And our experience tells us that such is not the case for any of the NAEP data we have analyzed. For a two-parameter model the sufficient statistic is a weighted sum of item scores where the weights are the discrimination parameters of the IRT model. The polytomous IRT model used in NAEP is a generalization of the two-parameter model, so a similar situation pertains. And for the three-parameter model there is no sufficient statistic. The likelihood equations for estimation of proficiencies according to that model, however, involve the discrimination parameters. In practice, of course, the item parameters are unknown, and we use estimates of these parameters.

---

[20] Creager, J.A., & Valentine, L.D., Jr. (1962). Regression analysis of linear composite variance. *Psychometrika, 27,* 31-37.

Horst, P. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika, 1,* 53-60.

Horst, P. (1941). *The prediction of personal adjustment* (Social Science Research Council Bulletin 48). New York: SSRC.

Richardson, M.W. (1941). The combination of measures. Supplementary study D in Horst, P. (Ed.). *The Prediction of Personal Adjustment.* Social Science Research Council Bulletin 48. New York: SSRC, 377-401.

Wilks, S.S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23-40.

[21] Carlson, J. E. (1968). *Effects of differential weighting on the inter-reader reliability of essay grades.* Unpublished dissertation, University of Alberta., p. 64.

[22] Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

This leads to an additional uncertainty. Note, however, that the same is true of variances and covariances entering into the variance contribution formulas associated with number-correct scoring as described above.

Lord also points out that, "It is generally agreed that when a sufficient statistic *s* exists for θ, any statistical inference for θ should be based on some function of *s* and not on any other statistic." He further states:

> There is no sufficient statistic when there is guessing ... This means that there is no sufficient statistic in cases, frequently reported in the literature, where the Rasch model ... is (improperly) used when the items can be answered correctly by guessing.[23]

Another important point made by Lord about IRT scoring is that the test information defined for the maximum likelihood estimator of proficiency (assuming an IRT model) "provides an (attainable) upper limit to the information that can be obtained from the test, no matter what method of scoring is used."[24] He also points out that test information, so defined, is a simple sum of item information, the information provided by the individual items comprising the test. Further:

> In classical test theory, . . . [t]here is no way to identify the contribution of a single item to test validity; the contribution of the item depends in an intricate way on the choice of items included in the test. The same may be said of an item's contribution to coefficient alpha . . . and to other test reliability coefficients.[25]

Lord goes on to discuss scoring weights and proves that optimal scoring weights (in terms of minimum error variance) are functions of the IRT model, and that the optimal weights yield an information function that is identical to the IRT information function. And Lord shows that this is the maximum information attainable by any scoring method.[26]

---

[23] Ibid., p. 58.

[24] Ibid., p. 71.

[25] Ibid., p. 72.

[26] Ibid., p. 74.

The conclusion that we reach from consideration of the points outlined above is that a simpler scoring scheme for NAEP instruments, while possible to use, would be no more defensible than the current methodology. In fact there is a suggestion that, for assessment purposes in general, the IRT-based methodology yields results that are statistically and psychometrically more defensible than those based on a classical test theory model.

Eliminate Plausible Values

As noted in the introduction, the plausible values approach makes it possible to estimate features of the proficiency distribution. This technically sophisticated approach was designed to produce accurate and statistically unbiased estimates of subpopulation characteristics under the current NAEP design where there is considerable unreliability in individual measurement. The plausible values technology admittedly takes time and money to implement, however. Two ways to eliminate the need for plausible values are discussed below.

Develop Ways of Increasing the Precision of Individual Measurement

This would involve a design change either to increase testing time or to employ targeted or adaptive assessment. Either of these would reduce individual measurement error and thereby (given sufficient reduction) allow the elimination of the plausible values technology. However, there are significant problems with either of these approaches. As noted earlier, increasing testing time could lead to marked fatigue effects, which would suppress performance, and to increases in nonresponse at the student and school level. Targeted or adaptive testing could result in an increase in the complexity of instrumentation, administration, and analysis.

Use Direct Estimation Techniques (Such As Marginal Maximum Likelihood) Instead of Plausible Values

Features of the proficiency distributions for various subpopulations can certainly be estimated without using plausible values. A specialized procedure could be used to obtain

such distributions in narrowly defined content areas without the need to generate plausible values for individual students. This procedure would have to be performed for each subpopulation separately, however, through numerous single runs—making this approach more time consuming than plausible-values technology. An added advantage of the plausible-values approach is that it handles all the potential interrelationships among proficiency scales and background variables. The plausible-values approach solves the estimation problem once (albeit with more work than any one or two or even ten of the simpler single runs) and permits the efficient completion of the hundreds of analyses required by the extensive number of NAEP background variables.

A final advantage of plausible-values methodology should be mentioned. The student-level data provided by this approach, upon which the NAEP reports are based, enables secondary researchers to carry out the full range of NAEP analyses. The specialized approach, in contrast, does not yield detailed student-level information for use by a broad range of secondary researchers relying on standard statistical packages.

We note that while the initial plans for the analysis and reporting of TIMSS called for the use of the specialized approach, the actual analysis of TIMSS used the plausible values approach.

Interactions Between Analysis Procedures and the Other Program Areas

The statistical and psychometric procedures needed to analyze the NAEP data are determined by the instrumentation, administration, and sampling, which are, in turn, determined by the *Frameworks* and other statements of the goals and priorities of NAEP. The statistical and psychometric procedures used affect reporting, since these procedures determine the results that can be reported. In addition, they affect instrumentation, sampling, and data collection in that BIB designs impose some complexity. A global redesign of NAEP would affect the analysis procedures used.

Specifically, different instrument, sample, and contextual data designs would change the nature of NAEP analysis.

Recommendations for Analysis

There are several recommendations about NAEP analysis that issue from the investigations undertaken in this chapter. These are described below.

Student testing time should not be increased. Expanding student testing time, while seemingly a way to reduce reliance on statistical models, would do irreparable harm to the NAEP system.

Item parameters should be re-estimated after each administration. Research conducted for this paper shows that item parameters drift over time, and such drift might well affect NAEP's ability estimates. Since recalibration represents a relatively simple step with previously-used items, such a step adds little cost.

NAEP should continue to rely on matrix sampling, Item Response Theory, and Marginal Estimation techniques. NAEP should continue to cover broad content domains while limiting individual testing time. Accomplishing these goals implies matrix sampled instruments and continued use of Item Response Theory and marginal estimation techniques.

NAEP should research methods that might be used off critical analysis paths to gain enhanced or different information about student performance. Rule-space analyses, improved non-response adjustments, and other analytic enhancements can provide more and different information for reports and secondary analysts. These analyses can be implemented in ways that do not interfere with main reporting schedules.

This page intentionally left blank.

CHAPTER 10
REPORTING

EXECUTIVE SUMMARY

This chapter discusses the ways in which NAEP data are and should be reported to the public. After an initial review of current NAEP reporting processes, reporting innovations are discussed. Specifically, the use of "market-basket" reporting metrics is examined. Other reporting innovations are also identified, as are potential sources of cost and schedule efficiency. Finally, the ways in which NAEP redesign may affect reporting are examined. The following arguments and recommendations are included in this chapter:

- Market-basket reporting metrics may be developed. However, such metrics may lead to misunderstandings for some users of NAEP data, and may prove more useful in some subjects than in others. Therefore, focus-group studies of market basket metrics should be undertaken before such reporting techniques are adopted. In addition, the implications of different types of market-baskets for the NAEP system must be studied.

- Pre-approval of report shells before analyses are completed would substantially speed the release of reports.

- CD-ROM and Internet tools may effectively be used to fashion a new generation of NAEP products; research should continue into the development of these products.

- As NAEP changes, reporting assumptions must change as well. For example, as states get more choice of which components of NAEP to administer the current state reporting system may need to be reevaluated. It may be wise and efficient to replace the current computer-generated reports with a combination of shorter reports and enhanced tools that will allow states to examine their own data.

This page intentionally left blank.

CHAPTER 10
REPORTING

- Stephen Lazer / Eugene G. Johnson -

Introduction

Providing information for the public is the *sine qua non* of the National Assessment. Reporting represents NAEP's public face and is the part of the program that influences and affects the greatest number of people. The amount and type of data to be reported drive the remainder of the NAEP system: only once these are determined can a system that will allow such reporting be effectively designed.

Consistent with our systemic focus on the overall NAEP project, several reporting issues are addressed in this chapter. First, we examine the ramifications of **market-basket reporting.** Market baskets may be a way to make NAEP reporting simpler and more meaningful to a broad public. However, there are also pitfalls in market-basket approaches that must be considered. Second, the **ways in which current processes impact the release of reports** are considered. Third, we examine **whether or not current analysis and instrumentation is appropriate to support new reporting goals.** Fourth, we investigate whether the assumptions underlying **the nature and amount of NAEP reporting** should be revisited. Specifically, we discuss the augmentation of current NAEP reports by a new generation of materials, and examine whether other reports might, at some point, be replaced by data tools that allow secondary users to derive any data they wish and produce their own reports.

Reporting in the Current NAEP

NAEP reports have changed over the years. In the early years of NAEP, instrument construction and analysis techniques permitted only item-level reporting, so program reports were compendia containing items and statistics on student performances. In the 1980s, NAEP adopted instrument and analysis designs that allowed for scaling and the maintenance of assessment-level trends, and these

techniques led to short *Report Cards* that summarized national data. In the early 1990s, the rise of state assessments had a number of impacts on reporting. National *Report Cards* became longer and included large volumes of tabular data. Computer-generated state reports were designed; these gave natural-language results for individual states. Finally, the increase in special studies and performance testing led to a series of focused reports on particular topics.

Toward the middle of the 1990s, NAEP began once again to change reporting plans. The *First Look* reports in 1994 were succinct and aimed at general audiences. The 1996 reports have been designed to target specific audiences: The *Report Cards* are written for the press and the general public; the *State Reports* are intended for state education officials; the *Focused Reports* are still written for researchers and special-interest communities; the *Instructional Reports* are designed to be of interest to teachers and curriculum developers; and the *Update Reports* are intended to be of interest to parents.

However, reporting has also led to certain systemic problems. The reporting process is one of the most contentious and time-consuming in NAEP. Different publics imply different conditions for the reporting of NAEP data, and resolving these differences is both slow and expensive. The purposes of reports and the amount of interpretive space that should be left to authors have remained open to question.

Reporting is interconnected with the rest of the system. What one can say in reports is a function of cognitive and background instruments, samples, data collection, and analysis. And what one *plans* to say must influence the designs of all other assessment components. Thus reporting plans are an important part of any redesign effort.

However, there are areas in reporting in which improvements can probably be made without affecting the remainder of the system. For example, Wainer and Hambleton and Slater have each written papers[1] describing ways in which NAEP data

[1] Hambleton, R., & Slater, S. (1995). Are the reports understandable and how can they be improved? *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board (NAGB) and the National Center* for Education Statistics (NCES). Washington, DC: U.S. Government Printing Office.
Wainer, H. (1997). Some multivariate displays of NAEP results. *Psychological Methods*, in press.

might be made more comprehensible to the public. In addition, ETS has made use of new technologies, including the Internet, to improve the efficacy of NAEP reporting. While these are important, they are not addressed here. Instead, we have focused on aspects of reporting that will be affected by changes to other components of the NAEP system.

Market-Basket Reporting

The idea of building NAEP to support market-basket reporting has become quite popular. Yet, market baskets mean many things to many people. To some, they are solely a reporting metric. To others, they represent a way of building parallel test forms that will fundamentally change the way assessments are conceptualized, developed, and field tested. We will view the market-basket solely as a reporting mechanism.[2]

The market-basket as a reporting mechanism has a compelling simplicity: Identify a set of items typical of the assessment *Frameworks* and express assessment results in terms of a total test score on that exemplar instrument. The resulting report is simple and intuitive. For a short enough form (Variation #1 below), the public is provided the complete set of items on the exemplar test, with NAEP results (such as average scale scores or achievement levels) expressed as scores on that test.

Mislevy[3] identifies three variations of the market-basket reporting metric, where the variants are in terms of the size of the market basket:

> **Variation #1: Market-Basket the Size of a Typical Test Form**
> In this variation, a first market-basket collection would be used to establish a reporting metric—observed scores on this set of items, and other sets that are very much like it—and released to the public. Replicate market-basket collections would be administered in the same assessment, and, having been built to be parallel to the original collection, could be linked to it with standard equating functions.

[2] Throughout this report, we use the phrase "short form" to refer to linking forms and other limited block uses, and the phrase "parallel forms" to refer to a reconceptualization of NAEP in which BIB designs are discontinued and replaced with a series of tests amenable to true-score type analyses.

[3] Mislevy, R.J. (1997, May). *Implications of market-basket reporting for achievement level setting.* Washington, DC: Committee on the Evaluation of National and State Assessments of Educational Progress.

Variation #2: Market-Basket Larger than a Typical Form
A disadvantage of Variation #1 is that the breadth of the subject domain could probably not be fully represented with a set of items a student would be administered in a testing session. But multiple one-period collections of tasks might together adequately convey the mix of formats, skills, and topics specified in the content framework. This larger collection could serve as a market-basket for reporting results. The resulting advantages would be better representativeness and communication of content coverage. The disadvantage is that observed scores on a typical administered booklet no longer provide unbiased estimates of population distributions other than central tendency—in particular, not proportions of students at or above proficiency level cut points. More complex statistical procedures would be required for such inferences in terms of these larger market-baskets.

Variation #3: Market-Basket = Subject Domain
Darrell Bock's Domain Referenced Scoring[4] marshals a sufficient number of items to constitute an operational definition of skill in that domain, and reports results in terms of an expected score on the collection as a whole, which could be anywhere from 500 to 5000 items. To accomplish this requires establishing IRT scales or similar models for the items in the domain. Since no student would ever be able to take the entire item pool, scaling models would be the vehicle through which predictive distributions on the domain as a whole were calculated. Bock envisages releasing the entire domain of items immediately to the public in order to stimulate discussion and learning in the subject area (although its use by various individuals and groups might raise performance on these kinds of tasks at the expense of skills that cannot be addressed in the NAEP setting). This variation requires large item development and item calibration efforts at the beginning of, say, a ten-year period.

ETS has already experimented with the market-basket idea as a metric for reporting.[5] The data used for this reporting experiment were taken from the 1992 mathematics assessment. The market basket for each grade consisted of three of the released blocks of items given to that grade in 1992. Since students in the 1992 assessment were each given three cognitive blocks, this is an example of Market-Basket Variation #1. However, this is the projected version of Variation #1 because the three blocks never appeared together in an assessment booklet. Consequently, IRT and conditioning technology were essential to produce market-basket scores for the students.

Tables 10-1 and 10-2 compare the existing reporting system with the market-basket reporting system. Table 10-1 uses the scale scores to present average

[4] Bock, R.D. (1996). *Domain-referenced reporting in large-scale educational assessments.* Commissioned paper to the National Academy of Education, for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment.

[5] Johnson, E. G. (1996, August). *A demonstration of market-basket reporting.* Presentation to the National Assessment Governing Board, Washington DC.

proficiency, the cutpoints for the basic, proficient, and advanced achievement levels, and the percentages of students at or above each of those cutpoints. Table 10-2 presents the same information in terms of the percent correct metric on the market-basket. Thus, for example, the mean score for grade 8 students would be 42 percent of the total possible market-basket score rather than a mean proficiency of 268. The advanced cutpoint for grade 8 would be 73 percent, rather than a scale score of 333. However, the percentage of grade 8 students above the advanced cutpoint is 3.8 in either metric.

Table 10-1: 1992 NAEP Mathematics Assessment Results in Terms of Scale Scores

	MEAN SCORE	ADVANCED		PROFICIENT		BASIC	
Grade		Cut Point	Percent At Or Above	Cut Point	Percent At Or Above	Cut Point	Percent At Or Above
4	220	282	3.1	249	19.1	214	61.5
8	268	333	3.8	299	23.6	262	55.6
12	300	367	2.8	336	14.9	288	63.1

Table 10-2: 1992 NAEP Mathematics Assessment Results in Terms of Market-Basket (Percent of Total Score)

	MEAN SCORE	ADVANCED		PROFICIENT		BASIC	
Grade		Cut Point	Percent At Or Above	Cut Point	Percent At Or Above	Cut Point	Percent At Or Above
4	41	80	3.1	58	19.1	34	61.5
8	42	73	3.8	55	23.6	37	55.6
12	40	75	2.8	57	14.9	33	63.1

Market-basket reporting of this type might have several real advantages. Percentages of possible score on a defined set of items might prove intrinsically

meaningful to readers. In addition, market-baskets would likely have the effect of grounding interpretations firmly in the observed data.

If NAEP planned to generate market-basket scores booklets consisting of the three market-basket blocks could be presented to students as intact instruments.[6] In fact, this could be a desirable feature of market-basket reporting since initial reports could be based only on those booklets with results determined by the direct computation of student-level total test scores. Subsequent analyses and reporting could be based on projecting results from other assessment booklets onto the exemplar market basket.

It should be noted that this direct reporting of total test scores on the exemplar market basket can only be done once. The reason is that such direct reporting over two cycles of the assessment would necessitate either the readministration of released exercises for the second assessment, or the non-release of the market basket items after the first assessment. The former choice would leave results open to artifactual increases in performance due to knowledge of the test items; the latter would largely eliminate the utility of the market-basket metric, since the public would not be able to judge the meaning of a particular score on an unknown test. Consequently, some mechanism for strongly linking other forms to the exemplar market basket must be employed.

It is important to note that market-basket reporting approaches are not without certain negative implications. The Redesign External Advisory Panel[7] felt that the public might have difficulty with an assessment where the advanced achievement level is expressed as 73 percent of the total possible score on the market basket, or where the basic level is set at 34 percent. The problem is that a score of 73 would be widely perceived by the public as C-level work. Clearly, readers of NAEP reports would likely require some explanation that the market basket represents a difficult test would be necessary.

---

[6] If IRT analysis techniques are used, the presentation of intact market-baskets to students is not strictly necessary.

[7] The panel consisted of Johnny Blair, Associate Director of the Survey Research Center at the University of Maryland; Robert Linn, Co-Director of the Center for Research on Educational Standards and Student Testing; Dori Nielson, Montana Director of Testing; and David Thissen, Director of the Quantitative Psychology Program at the University of North Carolina at Chapel Hill.

In addition, it is not immediately clear that the public would prefer market-basket reporting to the current scale scores used. Our assumption is that these scores would be more meaningful, but this is an hypothesis that should be tested. We therefore recommend that NCES undertake focus-group research into the use of market-basket metrics. Such studies, involving users of NAEP data, could help determine whether market-baskets should be used and what form they should take.

Besides these questions, the following issues related to the use of certain market-basket methodologies in NAEP need to be considered before such instrumentation and reporting are operationally adopted:

- ***How could other forms be linked to the exemplar market basket?*** Ways to accomplish such a linking would range from the use of rigidly parallel forms to some variety of scaling.

- ***If an approach using multiple parallel market-baskets is adopted, what costs are related to the field work needed to develop rigidly parallel forms?*** These would include issues and costs related to a design that requires an initial pilot to screen items (of around the size of a current NAEP field test), followed by a large-scale field test (of the general size and complexity of a current main assessment) to allow for scaling.

- ***How can performance tasks best be included in parallel market baskets?*** This raises issues of cost, task-by-person interaction (leading to problems in linking), and increased measurement error.

- ***How representative of a broad domain is an exemplar market basket?*** This includes the coverage of the *Framework* possible for a given market basket length, the similarity of meaning of parallel forms of the market basket, and the consistency of meaning of the exemplar market basket from the first assessment cycle to an assessment many years later.

Ways In Which Current Processes Impact the Release of Reports

The current NAEP reporting process involves extensive government review after the actual data has been analyzed and reports have been written. This process can take over three months. NAEP has experimented in recent years with pre-approval of "report shells." In these cases, "dataless" report shells are submitted to the

government for review prior to the completion of analyses. In principle, these shells can be reviewed more extensively than is currently the case. For example, the shells might be sent to a range of governmental and external reviewers prior to the availability of data. This could lead to substantial decreases in reporting time. In fact, operational testing programs that report rapidly have prior agreement from all interested parties on the format and content of score reports. Implementing this sort of a process in NAEP would speed reporting. However, it is important to note that speed would be gained at the cost of a loss of flexibility in reporting and analysis.

Whether or Not Current Analysis or Instrumentation Is Appropriate to Support New Reporting Goals

Reporting priorities continue to change for NAEP, and these priorities should drive other program decisions. Therefore, the implications of different reporting plans and priorities for the other components of the NAEP system need to be examined.

For example, if the goal of NAEP is to report rapid and inexpensively-obtained scale scores, one would likely design instruments that made only limited use of open-ended assessment. On the other hand, if the goal is for reports to show examples of student work on a variety of tasks, then the inclusion of performance assessments, either in core assessments or in modules, is of paramount importance.

While these examples seem obvious, they do raise an essential point: Changes to any part of the NAEP system must be evaluated in terms of the effort and schedule they imply for reporting of data, and in terms of whether the instruments will yield the data that users of NAEP wish to see in reports. Once again, trade-offs will be made depending on the program priorities set, and on the audiences for whom assessment results are targeted.

The Nature and Amount of NAEP Reporting

The specific structure of NAEP national and state reports has, as we have stated above, evolved over time. But one thing has been consistent: Most jurisdictions

participating in NAEP have used essentially the same instrumentation in a given subject. In other words, where a state program was present, all participants have used all available components. State results have been reported through the use of an artificial-intelligence system that produces natural language reports that summarize and discuss data separately for each participating jurisdiction.

This system has worked well. However, there are likely to be two serious changes in NAEP that both have implications for the reporting of state data. In the first place, states have opted in or out of components of the state program over time. In other words, a state may have participated in the 1990 and 1996 assessments, but not in the 1992 assessment. As the number of trend points increases, the years in which particular states participate varies, meaning that different reports will be needed for different states. Similarly, some states test in public schools only, while others include private schools in their samples. Again, this creates the need for a range of state reports. This increasing complexity has already occurred.

In addition, in coming years demands for greater flexibility may lead NAEP to allow states to pick and choose among assessment options. Such choice will likely call into further question the utility of artificial-intelligence reporting. Such expert systems may simply not be operationally feasible as the number of assessment options proliferates.

This may mean that certain NAEP designs will necessitate a rethinking of both the nature and amount of reporting. For example, NAEP has already developed software for secondary analysis of data: the NAEPEX for Windows Data Generation Program and the NAEP/SPSS Analysis modules. NAEP might develop more such tools that can be used by participating jurisdictions to produce their own reports. NAEP contractors or grantees might then provide, as part of their specified work, regular and ongoing assistance to external users of NAEP data. The Internet provides a low-cost means of distributing such data. It also provides a way to develop data interrogation tools, including those that allow review and retrieval of data and those that allow active analysis. Such potential changes would likely lead to major savings for the program, and would allow for the release of data in a more timely fashion.

While it may be necessary to limit reporting in some areas, in others the array of NAEP products and reports can be usefully expanded. For example, individuals involved with school recruitment efforts for national and state NAEP report that schools and teachers are often not very knowledgeable about NAEP and often feel the program and its results are not relevant to their day-to-day instructional activities. Additional Internet or computer-based products could be of direct use to the nation's teachers. One such product could involve released test forms, with item rationales, answer keys for multiple-choice items, scoring rubrics with exemplar papers for constructed-response items, and national results for these items. Such a product could be made available on CD-ROM or through the Internet, in the same way that NAEP Data Almanacs are currently provided. Teachers and schools could administer all or parts of these tests in classroom instruction or evaluation activities and compare results for their students to students across the nation on an item-by-item basis.

CD-ROM or Internet staff development tools could be designed to illustrate performance tasks and scoring rubrics and to convey a common understanding of high standards for performances on these tasks. The tools could include large numbers of exemplar responses or portfolio submissions, annotated to illustrate the admirable or deficient features of a response relevant to a set of grading standards. Teachers could explore these tasks and responses to learn about large- scale performance testing and to better understand the standards embodied in scoring.

NAEP has a wealth of information about the performance of students on constructed-response exercises. Finding more effective ways to share this information with teachers would increase the educational value of the NAEP program.

Recommendations for Reporting

There are several recommendations about NAEP reporting that issue from the investigations undertaken in this chapter. These are described below.

Focus-group studies should be conducted to evaluate market-basket reporting plans before implementation. Market-basket reporting metrics may be developed. However, such metrics may lead to misunderstandings for some users of NAEP data,

and may prove more useful in some subjects than in others. Therefore, focus-group studies of market basket metrics should be undertaken before such reporting techniques are adopted.

Pre-approval of report shells before analyses are completed would substantially speed the release of reports. Further streamlining of the cumbersome adjudication process would lead to more timely release of NAEP reports.

CD-ROM and Internet tools may effectively be used to fashion a new generation of NAEP products. Research should continue into the development of these products, and should focus on the provision of information of use to teachers and educators.

The nature and amount of state reporting must be re-examined. As NAEP changes, reporting assumptions must change as well. For example, as states get more choice of which components of NAEP to administer the current state reporting system may need to be reevaluated. It may be wise and efficient to replace the current computer-generated reports with a combination of shorter reports and enhanced tools that will allow states to examine their own data.

This page intentionally left blank.

APPENDIX

This page intentionally left blank.

THE NATION'S REPORT CARD NAEP

National Assessment Governing Board

National Assessment of Educational Progress

Policy Statement

on

Redesigning

The National Assessment of Educational Progress

Adopted Unanimously by
The National Assessment Governing Board, August 2, 1996

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

This page intentionally left blank.

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. The National Assessment Governing Board believes that twelfth grade achievement is important to monitor at the national level, because the 12th grade marks the end of elementary and secondary education, the transition point for most students from school to work, to college, or to technical training. The National Assessment is the only source of nationally representative data at the 12th grade. 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. And to date, virtually no state-based assessment program tests 12th graders.

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.

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.

National Assessment Redesign

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. These changes are described in this policy statement. Undergirding these changes is an explicit statement of the purposes, objectives, audiences, and limitations of the National Assessment.

While change is in order, many current policies should continue. For example, reliability, validity, and quality of data will remain hallmarks of the National Assessment. The sample of tested students will be as representative as possible, using policies and procedures that maximize the number of students included who are disabled or are of limited English proficiency. And reporting on trends over time will remain a central commitment of the National Assessment.

The intent of this policy statement is to guide current operations of the National Assessment, the development of new requests for proposals for contracts for conducting the National Assessment and the activities and structure of the National Assessment Governing Board. Contracts for current operations extend through assessments to be conducted in 1998. New contracts would cover assessments as early as 1999 and thereafter.

Purpose and Objectives 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 <u>should</u> 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 specific changes in the design of the National Assessment described below are discussed in relation to these objectives.

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. Although more comprehensible, direct, and useful, the reports will not trade accuracy for simplicity. The tradition of high quality of National Assessment reports will be continued, with no erosion of validity and reliability. Assessment questions and samples of student work that illustrate performance standards are likely to receive heightened prominence in reports.

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

Limitations: 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.

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.

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

Some have suggested that a solution is to combine into a single assessment several related subjects (e.g. reading and writing and/or history, geography, civics, and economics). Under such an approach, assessment data would be reported using both an overall score and subscores for the respective disciplines. Although such an approach has the appeal of

reducing the number of separate assessments, its feasibility, desirability, and costs are unknown. Also, such an approach has far-reaching implications for the test frameworks that guide the development of each assessment and for reporting results. These implications must be considered carefully. For the immediate future, subjects will continue to be assessed separately. However, the National Assessment Governing Board is committed to providing the public with more information as efficiently as possible. The Governing Board will consult with technical experts and education policymakers, in conjunction with the development of assessment frameworks, to determine the feasibility, desirability, and costs of combining several related subjects into a single assessment.

- **The National Assessment shall be conducted annually, two or three subjects per year, in order to cover all required subjects at least twice a decade.**

- **The National Assessment shall assess all subjects listed in the third National Educational Goal—reading, writing, mathematics, science, history, geography, civics, the arts, foreign language and economics—according to a publicly released schedule adopted by the National Assessment Governing Board, covering eight to ten years, with reading, writing, mathematics, and science tested more frequently than the other subjects.**

- **The National Assessment Governing Board shall consult with technical experts and with education policymakers, in conjunction with the development of assessment frameworks, to determine the feasibility, desirability, and costs of combining several related subjects into a single assessment.**

Provide National Assessment results for states

In 1988, testing at the state level was added to the National Assessment as a trial, with participation strictly voluntary, subjects and grades specified in law, and an independent evaluation required. Previously, the National Assessment had 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. The independent evaluation concluded that the trial state assessments produced valid and reliable data. The evaluation report recommended, and Congress agreed, that state-level assessments, with continued evaluations, be included in the 1994 reauthorization of the National Assessment.

Currently, the National Assessment draws a separate sample to obtain national results in addition to the samples drawn for individual state reports. Keeping the schools drawn for national samples completely partitioned from the state samples increases costs and creates additional burdens on states, particularly small states. Options should be identified for making the national and state samples more efficient and less burdensome. For example, it may be possible to reduce the current state sample size of 100 schools to a smaller number (e.g. 65-75) without a great loss in precision.

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 participate in the National Assessment to get state level information at grades 4 and 8 in reading, writing, mathematics, and science. The level of interest in participating in the National Assessment varies with respect to the other subjects (i.e., history, geography, civics, economics, the arts, and foreign language) and at grade 12, where state officials say that obtaining cooperation from high schools and 12th grade students is difficult.

Some states, however, would like to be able to use National Assessment tests in the other subjects and at grade 12. Such use of National Assessment tests would be conducted as a service, with the reporting of results and maintenance of data under the control of the state. States will be able to use National Assessment tests if they adhere to requirements to protect the integrity of the National Assessment program and pay the additional costs. At the present time, states that participate in the National Assessment to get state level information at grades 4 and 8 in reading, writing, mathematics, and science provide in-kind support to cover the cost of in-state coordination and test administration. The National Assessment program covers the majority of costs, including test development, sampling, analysis and reporting. States that wish to use National Assessment tests in other subjects and at grade 12 would pay for much of these additional costs.

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.

- National Assessment state-level assessments shall be conducted on a reliable, predictable schedule according to an eight to ten year plan adopted by the National Assessment Governing Board.

- Reading, writing, mathematics, and science at grades 4 and 8 shall be given priority for National Assessment state-level assessments.

- States shall have the option to use National Assessment tests in other subjects and at grade 12 by assuming a larger share of the costs and adhering to requirements that protect the integrity of the National Assessment program. However, the National Assessment Governing Board shall seek ways to make such use of National Assessment tests attractive and financially feasible.

- Where possible, changes in national and state sampling procedures shall be made that will reduce burden on states, increase efficiency, and save costs.

Vary the amount of detail in testing and in reporting results

More subjects can be assessed if different strategies are used. Currently, each time the National Assessment is conducted, it uses a similar approach, regardless of the nature of the subject or the number of times an assessment in a subject has been administered. 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, the cost of this thoroughness is that fewer subjects are assessed, assessments occur less frequently, and reports take longer to produce.

The different strategies needed might include several approaches to testing and reporting, all of which should be designed in ways that maintain the National Assessment's commitment to providing valid and reliable data of high quality. 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 could 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. 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.

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. Comprehensive reporting in a particular subject would occur infrequently, perhaps once in ten years, but under a planned schedule of assessments.

Special, focused assessments on timely topics also would be conducted. 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 versus out-of-school youth), and (3) examining skills and knowledge across several subjects (e.g. readiness for work).

The use of background surveys also would be varied. The three kinds of background surveys—student, teacher and principal questionnaires—would not necessarily all be employed each time a subject is assessed. Instead, the use of such surveys would be limited and selective, with reports of results focused on a core of background questions addressing the most essential issues. Also, background surveys used for standard report cards in a particular year would be designed to complement, rather than duplicate, background surveys used for comprehensive reports in the same year.

- **National Assessment testing and reporting shall vary, using standard report cards most frequently, comprehensive reporting in selected subjects about once every ten years, and special, focused assessments.**

- **National Assessment results shall be timely, with the goal being to release results within 6 months of the completion of testing for standard report cards and within 9 months for comprehensive reports.**

Simplify the National Assessment design

The current design of the National Assessment is very complex and, in fact, has grown more complex over the years. Here are just three examples of this complexity. (1) 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. Scores are calculated using sophisticated statistical procedures. (2) Students, teachers, and principals complete separate background questionnaires and may submit them for scoring at different times. Data from the questionnaires are used in calculating results of the assessments. (3) Current requirements for data analysis demand that test scores be calculated for every background variable collected by the National Assessment before any report can be produced. This lengthens the time from data collection to reporting and adds significantly to cost.

The design became more complex, in part, because the National Assessment's purposes and audiences had proliferated and the amount of background information collected had expanded. Specifying the purposes, audiences, and limitations of the National Assessment, as well as providing for varied means for testing and reporting, will result in opportunities for simplifying the National Assessment design.

- **Options shall be identified to simplify the design of the National Assessment.**

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. Monitoring long-term trends in educational achievement, both for the population as a whole and for significant sub-groups, is a capacity unique to the National Assessment and should be continued as a central mission. However, as the National Assessment approaches its third decade, it must address the problem of how to assess trends in achievement when curricula continue to evolve and change. An assessment in a subject must be kept stable to monitor trends. However, stable assessments may not reflect important changes in curricula. Over time, there develops a legitimate concern about the relevance of the content of the assessment versus the ability to track change in achievement.

As a solution to this problem, since 1990, the National Assessment has reported achievement trends using two unconnected assessment 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 assessments 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 assessed in the main National Assessment is unrelated to the schedule of subjects under the second testing program.

The second assessment program reports long-term trends that go as far back as 1970. Only four subjects are covered: reading, writing, mathematics, and science. The assessments 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 assessment 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 should be recognized as a continuing policy matter for the National Assessment, requiring efficient and balanced design solutions. Among the factors to consider are: (1) setting a standard period of time for a long-term trend (e.g. 15-20 years) using a particular "metric" in a subject; (2) providing for overlapping administrations of old and new assessments and "bridge" studies to determine whether the new can be linked to the old assessment; and (3) periodic administration of older assessments (e.g. once every ten years once a new trend-line has been established so that it would be possible to compare performance in 2010 with that in 1970 on the old trend line and with that in 1990 on a new trend line).

- **A carefully planned transition shall 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.**

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.

First employed in 1990, the achievement levels have been the subject of several independent evaluations and some controversy. Information from these evaluations, as well as from other experts, has been used over the last six years to improve and refine the procedures by which achievement levels are set. Although the current procedures may be among the most comprehensive and sophisticated standard-setting procedures used in education, the Governing Board remains committed to improving the process and to the continuing conduct of validity studies.

- **The National Assessment shall 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. Although there are technical hurdles to overcome, consideration of such qualitative information can be used to good effect. 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. This, in turn, should take into account problems in making international comparisons truly comparable, such as differences in the samples of students tested, differences in the curricula, and differences in the translated test questions.

- **National Assessment test frameworks, test specifications, achievement levels, and data interpretations shall take into account, where feasible, curricula, standards, and student performance in other nations.**

- **The National Assessment shall 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. Without practical incentives, even at grades 4 and 8, the willingness of district and school administrators and staff to participate in the National Assessment may diminish over time.

- **The National Assessment shall 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 shall be permitted when deemed appropriate and when necessary for international comparisons and for long-term trends, should the National Assessment Governing Board decide to continue long-term trends in their current form.**

- **Grade 12 results shall 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 shall seek to improve school and student participation rates and student motivation at grade 12.**

- **The National Assessment shall provide practical incentives for school and district participation at grades 4, 8, and 12.**

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 on 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. Students are then administered test exercises that are in the range of difficulty indicated by the pre-test. 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.

- **The National Assessment shall assess the merits of advances related to technology and the measurement and reporting of student achievement.**

- **Where warranted, the National Assessment shall 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, shall 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 an assessment will be put together. The 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 assessment 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.).

Since 1989, the National Assessment Governing Board has conducted a national consensus process to develop new test frameworks and specifications. 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 test questions developed for the National Assessment are reviewed for relevance and quality by representatives from participating states.

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.

- **Test frameworks and test specifications developed for the National Assessment generally shall 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 shall 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 National Assessment 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 shall be estimated. The National Assessment 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 produce 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." Both types of questions can vary in difficulty and the richness of information they provide, and 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. They range in length of test taking time from a short-answer or fill-in-the-blank format requiring about a minute of response time, to items requiring about 5 minutes of

response time, to writing exercises that may allow 15 to 50 minutes response time. Although they may be desirable, performance items are more expensive than multiple-choice to develop, administer, and score. In addition, much larger proportions of students fail to respond to performance items, particularly as the amount of required response time increases.

Multiple-choice questions can be challenging and are desired because they are efficient in collecting information about student knowledge. 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 exercises. In other subjects, the mix of multiple-choice and performance items varies. The appropriate mix of items for each subject should be determined by the nature of the subject, the range of skills to be assessed, and cost.

- **Both multiple-choice and performance items shall 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 shall 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 frameworks, specifications, scoring guides, results, questions, achievement levels, and background data—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 also would have a means for comparing their tests and standards with those of other states.

- **The National Assessment shall develop policies, practices, and procedures that assist states, school districts, and others who want to do so at their own cost to link their test results to the National Assessment.**

- **The National Assessment shall be designed so that others may access and use National Assessment test frameworks, specifications, scoring guides, results, questions, achievement levels, and background data.**

- **The National Assessment shall employ safeguards to protect the integrity of the National Assessment program, prevent misuse of data, and ensure the privacy of individual test takers.**

This page intentionally left blank.

DRAFT
November 4, 1996

National Center for Education Statistics

An Operational Vision for NAEP—Year 2000 and Beyond

Background

The time is right for a new, forward-looking, and comprehensive look at the National Assessment of Educational Progress (NAEP). NAEP's current design has developed largely because the program has responded to a wide array of important, but diverse, purposes and interests. It is appropriate to examine whether NAEP's design and approach effectively reflect its most important missions and purposes.

The concepts put forward here are built upon and consistent with the principles for the redesign of NAEP developed by the National Assessment Governing Board, offering an operational way of meeting those principles. As the agency responsible for conducting NAEP, the National Center for Education Statistics has the responsibility of considering NAEP's design in light of its most important purposes and constructing a workable vision for NAEP in the future that operationalizes its most important roles. Responding to this responsibility of NCES, this document assembles and puts forward such a vision. It tries to present an operational approach that addresses explicitly all of the major dimensions or parameters involved in planning for NAEP in the future. The document presents in a single statement a plan that acknowledges the needs and opportunities confronting NAEP, as well as the constraints and parameters under which an operational plan must be constructed. In one set of proposals, the global redesign possibilities for NAEP are addressed, along with the practical concerns of national and state information needs and the resources available to support NAEP.

Premises

This operational vision for NAEP is premised on several parameters and assumptions:

Long-term trend lines must be maintained.

Data are needed at the national level in core and other subjects on a regular and reliable basis: some results should be released each year, and certain results should be released at the same time each year.

States should have access to regular state-level information on NAEP produced according to sound methods, and they should be able to acquire sufficient data to meet their needs for normative information.

Results should be reported according to NAGB-established achievement levels, among other ways.

NAEP's design must be rethought to some extent in order to make it more effective in responding to major needs.

An Operational Vision for NAEP

As indicated above, the vision expressed here consists of several elements which work together to meet many needs:

- Four types of assessment would be included: long-term trend assessments similar to current practice; comprehensive assessments using new frameworks and instruments; "standard" assessments which would replicate comprehensive assessments but be less bulky and more efficient in some areas; and "market baskets"—assessment blocks which would represent the cognitive measures and be used for obtaining state-level and additional national level results.

- At least to start, **Long-term Trend** assessments would be similar to those used now in NAEP to maintain long-term trend lines; however, more efficient methods for integrating and relating long-term trend assessments and other forms of the assessment will be explored.

- **"Comprehensive"** assessments would be similar to, but a modification of, the complete assessments done now. They would be based on new frameworks and incorporate new cognitive measures, and they would be used at least for 8-10 years. Modifications would be made from current models to make analysis and reporting more targeted and efficient. Also, comprehensive assessments would be designed to carry other, important parts of the assessment program.

- "**Standard**" assessments would replicate comprehensive assessments cognitively, but they would be streamlined versions of the comprehensive assessments in other respects. They would replicate the cognitive measures of comprehensive assessments in order to be sure that their results were sound and equivalent to the comprehensive assessments. This approach provides for sound, comparable data from these assessments without necessitating designing, developing, and verifying new assessments to "represent" the main assessments. By measuring and reporting on much-reduced background information, standard assessments would allow for more efficient administration between comprehensive assessments, to generate short-term trends and much focused background analyses.

- "**Market baskets**" would be blocks of assessment items designed to be equivalent to one another and to represent the cognitive scales of comprehensive and standard

assessments. They would be administered with comprehensive assessments in order to be scaled and validated, and then they would be released and used as needed to provide representation of the cognitive scales—for state-level reporting, for indications of performance in a subject in years other than a main assessment, and for other purposes.

- National assessments in the core subjects (reading, writing, mathematics, and science) would be done every eight years using comprehensive assessments and at four-year intervals in between, using replicating "standard" assessments. (see Table 1.) Other subjects would be assessed on varying schedules, as shown in the chart. This is an overall approach for the national-level assessments which corresponds to the design principles and schedule suggested by the National Assessment Governing Board in its redesign recommendations. However, here an approach for realizing these principles and goals is detailed and made explicit: "standard" assessments would be complete replications of one another, in terms of cognitive items, background questions, and procedures; standard assessments would replicate the cognitive portion of comprehensive assessments, in order to avoid technical development and validation which would be necessary, even for abbreviated new cognitive assessments used to represent comprehensive assessments; and, standard assessments could and would provide subscale information, since they would replicate the cognitive scales of the comprehensive assessments. This would make this instructionally-useful information available on a much wider basis.

- NAEP would continue to offer **States** the approach to State-level assessments which was used in the Trial State Assessment. However, in the future, NAEP would offer two additional features that should be very helpful to States. The first is that a "market basket" booklet would be offered as part of the routine assessment. It would simply be another booklet imbedded in the normal BIB design of the State assessment. At the end of the assessment, the market-basket booklet would be released to the public and could be used by districts, schools or teachers to conduct their own assessments.

- The second feature is that in addition to offering **market basket** as part of the State Assessments (primarily paid for by the Federal Government) NAEP would offer market-baskets in the main subjects not included in the current assessment. For example, if NAEP is testing in reading and writing, market baskets in math and science could still be offered. These additional market basket administrations would be paid for by the States (but monitored by the Government).

- In a typical **comprehensive** assessment of a subject, a number of elements would be present. The current cognitive assessment would be used in its entirety. Background items would be selected and revised from current items, to address the most important background areas, and they would be analyzed and reported more efficiently to

address the most salient issues in education. Something on the order of fifteen market basket blocks would be included in the comprehensive assessment, so they could be administered and verified as yielding the same measures as the overall assessment. Subsequent to the comprehensive assessment, these market baskets would be kept secure and used as intervening measures at the national level and as optional, normative measures at the state level.

- **Standard** assessments administered in between comprehensive assessments would include fewer background questions than current assessments, but would include the full, current cognitive assessment. This is to avoid technical development of a new assessment approach and to ensure the technical soundness of the standard assessment's cognitive measures. Analysis and reporting from standard assessments would be quicker than the current assessments or the comprehensive model, permitting reporting to be done much sooner than at present.

- **Market Basket** assessments would be additional cognitive assessment blocks, designed and verified to be parallel to the overall assessment. (This is a proven technology which has been used in large-scale assessments for many years.) Market baskets would be used for several purposes. Some would be used in intervening years between comprehensive assessments to provide national results, without necessitating administration of the entire assessment. Some would be used to provide nationally-benchmarked State-level results, in lieu of the trial-State model; in this use they would constitute NAEP's "gold standard." Some would be made available to States to use on a more frequent basis as part of a State's own program to obtain State-level results. (If used appropriately, this application could be benchmarked to the "gold standard.") Market baskets are essentially shorter, parallel forms of the main assessment, representing the domain being assessed. They are conceived as a way of obtaining data equivalent to the full assessment more cost-effectively. They would obviate the need for much sampling and administration of assessments for various purposes. In a comprehensive assessment, typically twelve regular assessment blocks would be used and kept secure, one market basket would be used and released for public use, and fifteen market baskets would be administered, scaled and verified in terms of their relationship to the overall assessment, to be kept secure for various uses.

The accompanying Table 1, provided by the National Assessment Governing Board, displays how this approach would work. Schedules are shown for long-term, comprehensive, and standard assessments at the national level and for optional trial-State and market basket assessments at the State level. Table 2 describes some of the components of the comprehensive, standard and market basket assessments. The details outlined in Table 2 are for discussion purposes only. They will change as we firm up our plans next year.

Conclusion

What would this approach look like from the State or local level? For the first time, national results from NAEP would be forthcoming on a predictable and reasonably frequent schedule, showing both short-term and long-term trends. It would now be possible to obtain State-level measures on the NAEP standard—as often as every year if that was wanted. Options would be offered for obtaining State-level data: from a traditional and comprehensive approach yielding all cognitive and background information, to a new approach offering normative, cognitive information in a more efficient way, reducing some of the testing time for students and offering the possibility of linkage to the State program to generate NAEP-referenced data on schools or districts.

This approach offers several concrete benefits and advantages to the NAEP program. It provides a predictable, frequent schedule for both national and State-level information. It capitalizes on the possibilities of rethinking the assessment into approaches that are more efficient in addressing NAEP's main needs. It also offers creative ways for NAEP to give States State-comparable results and benchmarking to the national assessment.

The ideas are put forward in the spirit of advancing the discussion of how NAEP can be reconsidered to meet its most important needs. Discussion and comment are welcomed.

Attachments

This page intentionally left blank.

Table 1
Proposed NAEP Schedule-for Discussion

YEAR	NATIONAL	STATE	STATE MARKET BASKET
1996	Math Science	Math (4, 8) Science (8)	
1997	Arts (8)		
1998	Reading Writing Civics	Reading (4, 8) Writing (8)	
1999	Long-term trend		
2000	Math Science	Math (4, 8) Science (4, 8)	Reading (4, 8) Writing (4, 8)
2001	U.S. History Geography		
2002	Reading Writing	Reading (4, 8) Writing (4, 8)	Math (4, 8) Science (4, 8)
2003	Civics **FOR. LANG (12)** Long-term trend		
2004	MATH Science	Math (4, 8) Science (4, 8)	Reading (4, 8) Writing (4, 8)
2005	**WLD** **HSTRY (12)*** **ECONOMICS (12)**		
2006	**READING**** Writing**	Reading (4, 8) Writing (4, 8)	Math (4, 8) Science (4, 8)
2007	Arts Long-term trend		
2008	Math **SCIENCE**	Math (4, 8) Science (4, 8)	Reading (4, 8) Writing (4, 8)
2009	**U.S.** **HISTORY***** Geography***		
2010	Reading****** **WRITING******	Reading (4, 8) Writing (4, 8)	Math (4, 8) Science (4, 8)

1. Assessments are conducted annually beginning in the year 2000.

2. National assessments are at grades 4, 8, and 12 unless otherwise indicated.

3. State assessments are at grades 4 and 8 in reading, writing, math and science only. They would be conducted under standard conditions as they are now.

4. State market baskets would be made available only to states participating in regular state assessments. Subjects would be limited to reading, writing, math, and science and would be available for grades 4 and 8. Market baskets would cover the two subjects NOT being tested that year in the regular state assessment program

5. President Clinton's literacy proposal calls for state testing every other year in reading and provides additional funds to NAEP to pay the cost. Upon enactment, the schedule for state assessments will be revised accordingly.

6. Reading and writing are paired, as are math and science. Assessments in these subjects are conducted once every four years in alternating even-numbered years.

7. All other subjects are conducted once every eight years in odd-numbered years.

8. Generally, two subjects are tested each year. Because of its scope, the arts assessment is conducted alone.

9. Long-term trend assessments are conducted once every four years.

10. Comprehensive assessments are indicated in BOLD UPPER CASE: standard assessments are in lower case.

* If the Governing Board decides to combine history, geography, civics and economics into a single assessment with subscale reporting by area, 2005 is the earliest likely date for implementation.

** If the Governing Board decides to combine reading and writing into a single assessment with subscale reporting by area, 2006 is the earliest likely date for implementation.

*** This may instead be the second administration of a combine history, geography, civics, and economics assessment and would be conducted as a standard assessment.

**** This may instead be the second administration of a combined reading and writing assessment and would be conducted as a standard assessment.

This page intentionally left blank.

Table 2 describes the characteristics of comprehensive, standard, long term trend, and market basket assessments. The table shows the differences and similarities between each of these types of assessments.

Table 2

Operational Implications of Survey Concepts Used in NAGB's NAEP Redesign*

Draft

	Long Term Trend	Current Cross Sectional	Comprehensive**	Standard**	Market Basket***
Assessment cycle (in years)	2	5	6	3	2
# school per grade/age	300	500	500	300 (?)	(Comp, St)
# schools pilot per grade	0	100	200	0	(Comp)
# schools NR field test per grade	0	0	500	0	(Comp)
# schools state sample per grade	0	100	75	75	75
Item development	no	yes	yes	no	(Comp)
Achievement levels	no	yes	yes	yes	yes
Oversampling	no	yes	yes	no (?)	(Comp, St)
Subscales	no	yes	yes	yes (?)	no
Constructed Response	yes	yes	yes	yes	yes
Extended Constructed Response	no	yes	yes	yes	yes (?)
Released items	no	yes	no	no	yes
Background - Student, General	45	26	26	13	(Comp, St)
Student , Subject	45	26	26	13	(Comp, St)
Student, Motivation		5	5		(Comp, St)
Teacher- Background		30	30	(?)	(Comp, St)
Teacher- Classroom		40	40	(?)	(Comp, St)
School - General	80	80	80	40	(Comp, St)
Conditioning	yes	yes	yes	yes	no
Scaling	yes	yes	yes	yes	no (?)
Equating	yes	yes	yes	yes	yes
Differential item functioning	yes	yes	yes	yes	yes
Weighting	yes	yes	yes	yes	yes

*The characteristics of the surveys in this table are rough working estimates used for ***discussion purposes only***.
**Cognitive portion of the assessment are identical
***Market Basket is version 1 in Design Feasibility Report

This page intentionally left blank.

ACKNOWLEDGMENTS

This report is the result of the effort of many individuals who contributed their considerable knowledge, experience, creativity, and enthusiasm to the Redesign Project. In addition to us, major sections of this report were authored by Nancy Caldwell, James Carlson, Keith Rust, Robert Mislevy, Juliet Shaffer, Kim Whittington, William Ward, and Gita Wilder. Data analyses for this report were performed by Norma Norris and Jim Ferris. Budget analyses were performed by Kim Whittington and Lauren Fried.

Essential input to guide the thinking that resulted in this report came from weekly meetings of the Redesign Planning Team consisting of Nancy Allen, John Barone, Nancy Caldwell, James Carlson, John Donoghue, Elizabeth Durkin, John Fremer, David Freund, John Mazzeo, Robert Mislevy, Linda Reynolds, Keith Rust, Juliet Shaffer, Bradley Thayer, William Ward, and Gita Wilder. Barbara Klish organized these meetings. Additional input was received from John Burke, James Green, and Brent Studer.

The document was considerably improved by the advice of the External Advisory Panel, which consisted of Johnny Blair, Graham Kalton, Robert Linn, Dori Nielson, Gerald Shelton, and David Thissen. Rosemary Loeb organized our meeting with the panel.

Support and valuable input also came from Henry Braun, Paul Ramsey, and Paul Williams.

Editorial assistance was provided by Lynn Jenkins, Debra Kline, and Elissa Greenwald. The design and production of the report was overseen by Carol Errickson and Terry Schoeps. The text processing for the report was most ably done by Terry Schoeps.

We thank you all.

Eugene G. Johnson
Stephen Lazer
Christine Y. O'Sullivan

This page intentionally left blank.

Listing of NCES Working Papers to Date

Please contact Ruth R. Harris at (202) 219-1831
if you are interested in any of the following papers

Number	Title	Contact
94-01 (July)	Schools and Staffing Survey (SASS) Papers Presented at Meetings of the American Statistical Association	Dan Kasprzyk
94-02 (July)	Generalized Variance Estimate for Schools and Staffing Survey (SASS)	Dan Kasprzyk
94-03 (July)	1991 Schools and Staffing Survey (SASS) Reinterview Response Variance Report	Dan Kasprzyk
94-04 (July)	The Accuracy of Teachers' Self-reports on their Postsecondary Education: Teacher Transcript Study, Schools and Staffing Survey	Dan Kasprzyk
94-05 (July)	Cost-of-Education Differentials Across the States	William Fowler
94-06 (July)	Six Papers on Teachers from the 1990-91 Schools and Staffing Survey and Other Related Surveys	Dan Kasprzyk
94-07 (Nov.)	Data Comparability and Public Policy: New Interest in Public Library Data Papers Presented at Meetings of the American Statistical Association	Carrol Kindel
95-01 (Jan.)	Schools and Staffing Survey: 1994 Papers Presented at the 1994 Meeting of the American Statistical Association	Dan Kasprzyk
95-02 (Jan.)	QED Estimates of the 1990-91 Schools and Staffing Survey: Deriving and Comparing QED School Estimates with CCD Estimates	Dan Kasprzyk
95-03 (Jan.)	Schools and Staffing Survey: 1990-91 SASS Cross-Questionnaire Analysis	Dan Kasprzyk
95-04 (Jan.)	National Education Longitudinal Study of 1988: Second Follow-up Questionnaire Content Areas and Research Issues	Jeffrey Owings
95-05 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses of NLS-72, HS&B, and NELS:88 Seniors	Jeffrey Owings

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
95-06 (Jan.)	National Education Longitudinal Study of 1988: Conducting Cross-Cohort Comparisons Using HS&B, NAEP, and NELS:88 Academic Transcript Data	Jeffrey Owings
95-07 (Jan.)	National Education Longitudinal Study of 1988: Conducting Trend Analyses HS&B and NELS:88 Sophomore Cohort Dropouts	Jeffrey Owings
95-08 (Feb.)	CCD Adjustment to the 1990-91 SASS: A Comparison of Estimates	Dan Kasprzyk
95-09 (Feb.)	The Results of the 1993 Teacher List Validation Study (TLVS)	Dan Kasprzyk
95-10 (Feb.)	The Results of the 1991-92 Teacher Follow-up Survey (TFS) Reinterview and Extensive Reconciliation	Dan Kasprzyk
95-11 (Mar.)	Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work	Sharon Bobbitt & John Ralph
95-12 (Mar.)	Rural Education Data User's Guide	Samuel Peng
95-13 (Mar.)	Assessing Students with Disabilities and Limited English Proficiency	James Houser
95-14 (Mar.)	Empirical Evaluation of Social, Psychological, & Educational Construct Variables Used in NCES Surveys	Samuel Peng
95-15 (Apr.)	Classroom Instructional Processes: A Review of Existing Measurement Approaches and Their Applicability for the Teacher Follow-up Survey	Sharon Bobbitt
95-16 (Apr.)	Intersurvey Consistency in NCES Private School Surveys	Steven Kaufman
95-17 (May)	Estimates of Expenditures for Private K-12 Schools	Stephen Broughman
95-18 (Nov.)	An Agenda for Research on Teachers and Schools: Revisiting NCES' Schools and Staffing Survey	Dan Kasprzyk
96-01 (Jan.)	Methodological Issues in the Study of Teachers' Careers: Critical Features of a Truly Longitudinal Study	Dan Kasprzyk

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-02 (Feb.)	Schools and Staffing Survey (SASS): 1995 Selected papers presented at the 1995 Meeting of the American Statistical Association	Dan Kasprzyk
96-03 (Feb.)	National Education Longitudinal Study of 1988 (NELS:88) Research Framework and Issues	Jeffrey Owings
96-04 (Feb.)	Census Mapping Project/School District Data Book	Tai Phan
96-05 (Feb.)	Cognitive Research on the Teacher Listing Form for the Schools and Staffing Survey	Dan Kasprzyk
96-06 (Mar.)	The Schools and Staffing Survey (SASS) for 1998-99: Design Recommendations to Inform Broad Education Policy	Dan Kasprzyk
96-07 (Mar.)	Should SASS Measure Instructional Processes and Teacher Effectiveness?	Dan Kasprzyk
96-08 (Apr.)	How Accurate are Teacher Judgments of Students' Academic Performance?	Jerry West
96-09 (Apr.)	Making Data Relevant for Policy Discussions: Redesigning the School Administrator Questionnaire for the 1998-99 SASS	Dan Kasprzyk
96-10 (Apr.)	1998-99 Schools and Staffing Survey: Issues Related to Survey Depth	Dan Kasprzyk
96-11 (June)	Towards an Organizational Database on America's Schools: A Proposal for the Future of SASS, with comments on School Reform, Governance, and Finance	Dan Kasprzyk
96-12 (June)	Predictors of Retention, Transfer, and Attrition of Special and General Education Teachers: Data from the 1989 Teacher Followup Survey	Dan Kasprzyk
96-13 (June)	Estimation of Response Bias in the NHES:95 Adult Education Survey	Steven Kaufman
96-14 (June)	The 1995 National Household Education Survey: Reinterview Results for the Adult Education Component	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-15 (June)	Nested Structures: District-Level Data in the Schools and Staffing Survey	Dan Kasprzyk
96-16 (June)	Strategies for Collecting Finance Data from Private Schools	Stephen Broughman
96-17 (July)	National Postsecondary Student Aid Study: 1996 Field Test Methodology Report	Andrew G. Malizio
96-18 (Aug.)	Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children	Jerry West
96-19 (Oct.)	Assessment and Analysis of School-Level Expenditures	William Fowler
96-20 (Oct.)	1991 National Household Education Survey (NHES:91) Questionnaires: Screener, Early Childhood Education, and Adult Education	Kathryn Chandler
96-21 (Oct.)	1993 National Household Education Survey (NHES:93) Questionnaires: Screener, School Readiness, and School Safety and Discipline	Kathryn Chandler
96-22 (Oct.)	1995 National Household Education Survey (NHES:95) Questionnaires: Screener, Early Childhood Program Participation, and Adult Education	Kathryn Chandler
96-23 (Oct.)	Linking Student Data to SASS: Why, When, How	Dan Kasprzyk
96-24 (Oct.)	National Assessments of Teacher Quality	Dan Kasprzyk
96-25 (Oct.)	Measures of Inservice Professional Development: Suggested Items for the 1998-1999 Schools and Staffing Survey	Dan Kasprzyk
96-26 (Nov.)	Improving the Coverage of Private Elementary-Secondary Schools	Steven Kaufman
96-27 (Nov.)	Intersurvey Consistency in NCES Private School Surveys for 1993-94	Steven Kaufman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
96-28 (Nov.)	Student Learning, Teaching Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection	Mary Rollefson
96-29 (Nov.)	Undercoverage Bias in Estimates of Characteristics of Adults and 0- to 2-Year-Olds in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
96-30 (Dec.)	Comparison of Estimates from the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-01 (Feb.)	Selected Papers on Education Surveys: Papers Presented at the 1996 Meeting of the American Statistical Association	Dan Kasprzyk
97-02 (Feb.)	Telephone Coverage Bias and Recorded Interviews in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-03 (Feb.)	1991 and 1995 National Household Education Survey Questionnaires: NHES:91 Screener, NHES:91 Adult Education, NHES:95 Basic Screener, and NHES:95 Adult Education	Kathryn Chandler
97-04 (Feb.)	Design, Data Collection, Monitoring, Interview Administration Time, and Data Editing in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-05 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1993 National Household Education Survey (NHES:93)	Kathryn Chandler
97-06 (Feb.)	Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)	Kathryn Chandler
97-07 (Mar.)	The Determinants of Per-Pupil Expenditures in Private Elementary and Secondary Schools: An Exploratory Analysis	Stephen Broughman
97-08 (Mar.)	Design, Data Collection, Interview Timing, and Data Editing in the 1995 National Household Education Survey	Kathryn Chandler

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-09 (Apr.)	Status of Data on Crime and Violence in Schools: Final Report	Lee Hoffman
97-10 (Apr.)	Report of Cognitive Research on the Public and Private School Teacher Questionnaires for the Schools and Staffing Survey 1993-94 School Year	Dan Kasprzyk
97-11 (Apr.)	International Comparisons of Inservice Professional Development	Dan Kasprzyk
97-12 (Apr.)	Measuring School Reform: Recommendations for Future SASS Data Collection	Mary Rollefson
97-13 (Apr.)	Improving Data Quality in NCES: Database-to-Report Process	Susan Ahmed
97-14 (Apr.)	Optimal Choice of Periodicities for the Schools and Staffing Survey: Modeling and Analysis	Steven Kaufman
97-15 (May)	Customer Service Survey: Common Core of Data Coordinators	Lee Hoffman
97-16 (May)	International Education Expenditure Comparability Study: Final Report, Volume I	Shelley Burns
97-17 (May)	International Education Expenditure Comparability Study: Final Report, Volume II, Quantitative Analysis of Expenditure Comparability	Shelley Burns
97-18 (June)	Improving the Mail Return Rates of SASS Surveys: A Review of the Literature	Steven Kaufman
97-19 (June)	National Household Education Survey of 1995: Adult Education Course Coding Manual	Peter Stowe
97-20 (June)	National Household Education Survey of 1995: Adult Education Course Code Merge Files User's Guide	Peter Stowe
97-21 (June)	Statistics for Policymakers or Everything You Wanted to Know About Statistics But Thought You Could Never Understand	Susan Ahmed
97-22 (July)	Collection of Private School Finance Data: Development of a Questionnaire	Stephen Broughman

Listing of NCES Working Papers to Date--Continued

Number	Title	Contact
97-23 (July)	Further Cognitive Research on the Schools and Staffing Survey (SASS) Teacher Listing Form	Dan Kasprzyk
97-24 (Aug.)	Formulating a Design for the ECLS: A Review of Longitudinal Studies	Jerry West
97-25 (Aug.)	1996 National Household Education Survey (NHES:96) Questionnaires: Screener/Household and Library, Parent and Family Involvement in Education and Civic Involvement, Youth Civic Involvement, and Adult Civic Involvement	Kathryn Chandler
97-26 (Oct.)	Strategies for Improving Accuracy of Postsecondary Faculty Lists	Linda Zimbler
97-27 (Oct.)	Pilot Test of IPEDS Finance Survey	Peter Stowe
97-28 (Oct.)	Comparison of Estimates in the 1996 National Household Education Survey	Kathryn Chandler
97-29 (Oct.)	Can State Assessment Data be Used to Reduce State NAEP Sample Sizes?	Steven Gorman
97-30 (Oct.)	ACT's NAEP Redesign Project: Assessment Design is the Key to Useful and Stable Assessment Results	Steven Gorman
97-31 (Oct.)	NAEP Reconfigured: An Integrated Redesign of the National Assessment of Educational Progress	Steven Gorman