NATIONAL RESEARCH COUNCIL

DESIGNING A MARKET BASKET FOR NAEP

SUMMARY OF A WORKSHOP

DESIGNING A MARKET BASKET FOR NAEP

SUMMARY OF A WORKSHOP

Committee on NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting

Pasquale J. DeVito and Judith A. Koenig, editors

Board on Testing and Assessment

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C.

National Academy Press • 2101 Constitution Avenue, N.W. • Washington, D.C. 20418

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

The study was supported by the U.S. Department of Education under contract number E95083001. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number 0-309-07128-3

Additional copies of this report are available from National Academy Press, 2101 Constitution Avenue, N.W., Lock Box 285, Washington, D.C. 20005. Call (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area).

This report is also available online at http://www.nap.edu

Printed in the United States of America

Suggested citation: National Research Council (2000) *Designing a Market Basket for NAEP: Summary of a Workshop.* Committee on NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. Pasquale J. DeVito and Judith A. Koenig, editors. Board on Testing and Assessment. Washington, D.C.: National Academy Press.

THE NATIONAL ACADEMIES

National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

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

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

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

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

COMMITTEE ON NAEP REPORTING PRACTICES: INVESTIGATING DISTRICT-LEVEL AND MARKET-BASKET REPORTING

PASQUALE DEVITO (*Chair*), Office of Assessment, Rhode Island Department of Education
LINDA BRYANT, Westwood Elementary School, Pittsburgh
C. MELODY CARSWELL, Department of Psychology, University of Kentucky
MARYELLEN DONAHUE, Planning, Research & Development, and District Test Coordination, Boston Public Schools
LOU FABRIZIO, Division of Accountability Services, North Carolina Department of Public Instruction
LEANN GAMACHE, Assessment and Evaluation, Education Services Center, Littleton Public Schools, Littleton, Colorado
DOUGLAS HERRMANN, Department of Psychology, Indiana State University
AUDREY QUALLS, Iowa Testing Program, Iowa City, Iowa
MARK RECKASE, Department of Counseling, Educational Psychology, and Special Education, Michigan State University
DUANE STEFFEY, Department of Mathematical and Computer Sciences, San Diego State University

JUDITH KOENIG, *Study Director*
KAREN MITCHELL, *Senior Program Officer*
KAELI KNOWLES, *Program Officer*
DOROTHY MAJEWSKI, *Senior Project Assistant*

BOARD ON TESTING AND ASSESSMENT

Acknowledgments

At the request of the U.S. Department of Education, the National Research Council (NRC) established the Committee on NAEP Reporting Practices to examine the feasibility and potential impact of district-level and market-basket reporting practices. As part of its charge, the committee sponsored a workshop in February 2000 to gather information on issues related to market-basket reporting for the National Assessment of Education Progress (NAEP). A great many people contributed to the success of this workshop, which brought together representatives from state and local assessment offices, experts in educational measurement, and others familiar with the issues related to market-basket reporting for NAEP. The committee would like to thank the panelists and discussants for their contributions to a lively and productive workshop. The full participant list appears in Appendix A.

Staff from National Center for Education Statistics (NCES), under the direction of Gary Phillips, acting commissioner, and staff from the National Assessment Governing Board (NAGB), under the leadership of Roy Truby, executive director, were valuable sources of information. Peggy Carr, Patricia Dabbs, Arnold Goldstein, Steve Gorman, Andrew Kolstad, and Holly Spurlock of NCES and Roy Truby, Mary Lyn Bourque, Sharif Shakrani, Lawrence Feinberg, and Raymond Fields of NAGB provided the committee with important background information on numerous occasions. Papers prepared for the workshop by Andrew Kolstad and Roy Truby were particularly helpful to the committee.

The committee also is grateful to John Mazzeo and Robert Mislevy, both from the Educational Testing Service (ETS), for the information they supplied as part of their papers for the workshop and their responses to our many questions after the workshop.

Special thanks are due to a number of individuals at the National Research Council who provided guidance and assistance at many stages during the organization of the workshop and the preparation of this report. We thank Michael Feuer, director of the Board on Testing and Assessment (BOTA), for his expert guidance and leadership of this project. We are indebted to BOTA staff officer, Karen Mitchell, for her assistance in planning the workshop and writing this report; she was a principal source of expertise in both the substance and process for this workshop. We also wish to thank BOTA staff members Patricia Morison and Viola Horek for their assistance with this work. Special thanks go to Dorothy Majewski, who capably managed the operational aspects of the workshop and the production of this report. We thank Eugenia Grohman for her deft guidance of the report through the review process and Christine McShane for her assistance in moving the report through the publication process.

The committee is particularly grateful to NRC project staff, Judith Koenig, study director, and Kaeli Knowles, program officer, for their efforts in putting together the workshop and preparing the manuscript for this report. Judith acted as the coordinator of the activities, while Kaeli Knowles played a prominent role in organizing and running the workshop, in obtaining written materials from workshop speakers, and in assisting with the writing of this report.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making the published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their participation in the review of this report: Karen Banks, Wake County Schools, Raleigh, NC; Brian Junker, Department of Statistics, Carnegie Mellon University; Jeffrey M. Nellhaus, Student Assessment, Massachusetts Department of Education; Thanos Patelis, The College Board, New York, NY; Alan Schoenfeld, School of Education, University of California, Berkeley; and Mark Wilson, Graduate School of Education, University of California, Berkeley.

Although the individuals listed above provided constructive comments and suggestions, it must be emphasized that responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Pasquale J. DeVito
Chair

Contents

1

Introduction

THE MARKET-BASKET CONCEPT

The purpose of the National Research Council's Workshop on Market-Basket Reporting was to explore with various stakeholders their interest in and perceptions regarding the desirability, feasibility, and potential impact of market-basket reporting for the National Assessment of Educational Progress (NAEP). The market-basket concept is based on the idea that a relatively limited set of items can represent some larger construct. The most common example of a market basket is the Consumer Price Index (CPI) produced by the Bureau of Labor Statistics. The CPI tracks changes in the prices paid by urban consumers in purchasing a representative set of consumer goods and services. The CPI measures cost differentials from month to month for products in its market basket; therefore, the CPI is frequently used as an indicator of change in the U.S. economy. In the context of the CPI, the concept of a market basket resonates with the general public; it invokes the tangible image of a shopper going to the market and filling a basket with a set of goods that is regarded as broadly reflecting consumer spending patterns.

The general idea of a NAEP market basket draws on a similar image: a collection of test questions representative of some larger content domain; and an easily understood index to summarize performance on the items. There are two components of the NAEP market basket, the collection of items and the summary index. The collection of items could be large (longer than a typical test form given to a student) or small (small enough to be considered an administrable test form). The summary index currently under consideration is the percent correct score.

At present, several alternatives have been proposed for the NAEP market basket. Figure 1 provides a diagram of the various components of the market basket and shows how they relate to two alternate scenarios under which the market basket would be assembled and used. Under one sce-

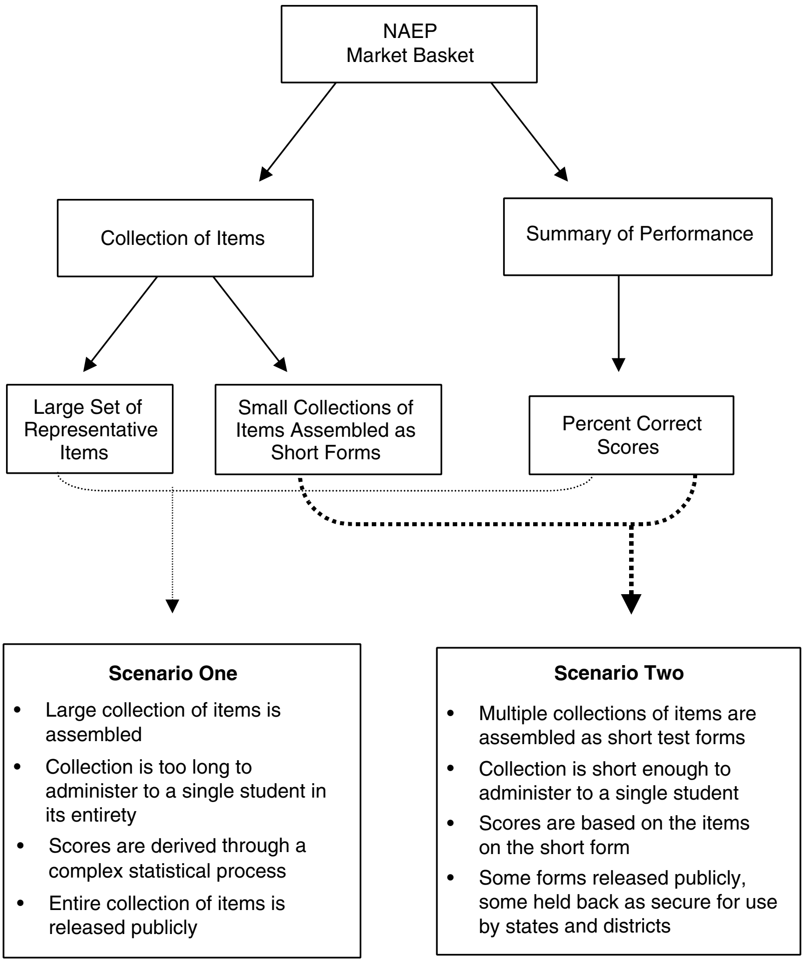

FIGURE 1 Components of the NAEP market basket.

nario, a large collection of items would be assembled and released publicly. To adequately cover the breadth of the content domain, the collection would be much larger than any one of the forms used in the test and probably too long to administer to a single student at one sitting. This presents some challenges for the calculation of the percent correct scores. Because no student would take all of the items, complex statistical procedures would be needed for estimating scores. This alternative appears in Figure 1 as "scenario one."

A second scenario involves producing multiple, "administrable" test forms (called "short forms"). Students would take an entire test form, and scores could be based on students'performance for the entire test in the manner usually employed by testing programs. Although this would simplify calculation of percent correct scores, the collection of items would be much smaller and less likely to adequately represent the content domain. This scenario also calls for assembling multiple test forms. Some forms would be released to the public, while others would remain secure, perhaps for use by state and local assessment programs, and possibly to be embedded into or administered in conjunction with existing tests. This alternative appears in Figure 1 as "scenario two."

THE COMMITTEE ON NAEP REPORTING PRACTICES

At the National Research Council (NRC), the study on market-basket reporting is being handled by the Committee on NAEP Reporting Practices. The NRC established this committee in 1999 at the request of the United States Department of Education to examine the feasibility, desirability, and potential impact of district-level and market-basket reporting practices. Because issues related to these reporting practices are intertwined, the committee is examining them in tandem. The committee's study questions regarding district-level reporting are as follows:

1. What are the proposed characteristics of a district-level NAEP?
2. If implemented, what information needs might it serve?
3. What is the degree of interest in participating in district-level NAEP? What factors would influence interest?
4. Would district-level NAEP pose any threats to the validity of inferences from national and state NAEP?
5. What are the implications of district-level reporting for other state and local assessment programs?

District-level reporting was the focus of a workshop held in September 1999. The proceedings from this workshop were summarized and published (see National Research Council, 1999c).

The committee's study questions with respect to market-basket reporting are as follows:

1. What is market-basket reporting?
2. How might reports of market-basket results be presented to NAEP's audiences? Are there prototypes?
3. What information needs might be served by market-basket reporting for NAEP?
4. Are market-basket results likely to be relevant and accurate enough to meet these needs?
5. Would market-basket reporting pose any threats to the validity of inferences from national and state NAEP? What types of inferences would be valid?
6. What are the implications of market-basket reporting for other national, state, and local assessment programs? What role might a NAEP short form play?

On February 7 and 8, 2000, the committee convened the Workshop on Market-Basket Reporting to begin to address the questions outlined above regarding NAEP market-basket reporting. The committee's further consideration of both district-level and market-basket reporting will be reflected in its final report, scheduled for release in November 2000.

WORKSHOP ON MARKET-BASKET REPORTING

The workshop opened with a panel of representatives from the organizations involved in setting policy for and operating NAEP: the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES). Also included were individuals from the Educational Testing Service (ETS), the contractual agency that works on NAEP. Panel members talked about the perceived needs that led to consideration of the market basket and plans for conducting research on the market basket. The panel included:

- Roy Truby, executive director of NAGB

- Andrew Kolstad, senior technical advisor for the Assessment Division at NCES
- Robert Mislevy, distinguished research scholar with ETS
- John Mazzeo, executive director of ETS's School and College Services

Prior to the workshop, each panel member prepared a paper addressing questions specified by the committee. At the workshop, time was allotted for panel members to present their papers orally. The initial questions posed to these panel members and synopses of their papers and presentations appear in Chapter 4 of this report.

The committee invited individuals representing a variety of perspectives to serve as discussants at the workshop and to react to the material presented by the opening panel speakers. These discussants received copies of the speakers' papers several weeks in advance of the workshop along with a set of questions to address in their comments (see Agenda in Appendix A for these questions). Each discussant made an oral presentation during the workshop and subsequently submitted written copy of his or her remarks.

Policy Issues

The first discussant panel responded from a policy perspective, highlighting the impact the market-basket proposal might have on state and local education policy for instruction and assessment programs. This panel included Wayne Martin, director of the Council of Chief State School Officers' State Education Assessment Center; Marilyn McConachie, a member of the Illinois State Board of Education and former member of NAGB; Carroll Thomas, superintendent of public schools in Beaumont, Texas; and Marlene Hartzman, an evaluation specialist with the public school system in Montgomery County, Maryland.

Assessment and Curriculum

A second discussant panel explored the perspectives of data users and practitioners. This panel considered the impact and uses of the market basket with respect to state and local assessment programs, curricula, and instructional practices. Panel members included Scott Trimble, director of assessment with Kentucky's Department of Education; Joseph O'Reilly, director of assessment for the public school system in Mesa, Arizona, and past president of the National Association of Test Directors (NATD); and Ronald Costello, assistant superintendent for public schools in Noblesville, Indiana.

Measurement Issues

A third panel of discussants responded to technical and measurement issues related to the market basket. This panel comprised well-known statisticians who have worked extensively with NAEP. Panel members were Darrell Bock, a faculty fellow with the University of Chicago's Department of Psychology; David Thissen, professor of psychology at the University of North Carolina-Chapel Hill; and Donald McLaughlin, chief scientist with the American Institutes of Research (AIR).

Content and Skill Coverage

Patricia Kenney, research associate with the Learning Research and Development Center at the University of Pittsburgh, was invited to speak with respect to the content and skill coverage of the collections of items included in the market basket. Kenney has extensive knowledge of the NAEP mathematics frameworks through her work with the National Council of Teachers of Mathematics (NCTM), her role as co-director of the NCTM NAEP Interpretive Reports Project, and her work comparing the congruence between NAEP and state assessments in North Carolina and Maryland.

First in the World Consortium

Two school district superintendents, Paul Kimmelman (of West Northfield, Illinois) and Dave Kroeze (of Northbrook, Illinois), discussed the work of the "First in the World Consortium." This group of 20 school districts in Illinois participated in and received results from the Third International Mathematics and Science Study (TIMSS) as part of their efforts to achieve their education goals and to work toward world-class standards. Kimmelman and Kroeze described the ways in which consortium members have used TIMSS results to guide changes in instruction and assessment. Their comments provided insights into how individual school districts might use the released short-form version of the NAEP market-basket.

A Newspaper Reporter's Perspective

The committee was interested in what the newspaper-reading general public wants to know about student achievement. Committee members wanted to hear a reporter's perspective on how the press might use results from the market basket and the types of conclusions that might be drawn from the information. Richard Colvin, education reporter with the *Los Angeles Times,* served as a discussant at the workshop.

The Consumer Price Index

The committee also solicited information about market baskets and summary indicators as they are used in other contexts, such as the Consumer Price Index (CPI), which is most frequently cited as an analogy for the NAEP market basket. Kenneth Stewart, chief of the Information and Analysis Section at the Bureau of Labor Statistics, spoke about how the CPI is formed using a market basket of representative consumer goods and services and how the CPI influences other economic measures.

ORGANIZATION OF THIS REPORT

The purpose of this report is to capture the discussions and major points made during the market-basket workshop in order to assist NAEP's sponsors in their decision making about the feasibility, desirability, and potential impact of the NAEP market-basket proposal. The workshop permitted a considerable amount of open discussion by presenters, as well as by participants, much of which is woven into this summary report. As a summary, this report is intended to highlight the key issues identified by the various stakeholders who attended the workshop, but it does not attempt to establish consensus on findings or recommendations. The Committee on NAEP Reporting Practices will publish its final report in November 2000 that will include findings from the entire study and will offer recommendations with respect to district-level and market-basket reporting.

The concept of a NAEP market basket was first addressed by NAGB in "Redesigning the National Assessment of Educational Progress" (National Assessment Governing Board, 1996) and is again discussed in "National Assessment of Educational Progress: Design 2000-2010 Policy" (National Assessment Governing Board, 1999b). Background on these two redesign efforts appears in Chapter 2 to provide readers with an understanding for the motivations behind the market-basket proposal.

Since the NAEP market-basket concept has often been compared to the Consumer Price Index, the information provided to workshop participants by Kenneth Stewart appears in the next chapter. Chapter 3 is intended to acquaint the reader with summary indicators in other contexts and establish background for the exploration of an analogous summary indicator for NAEP.

Chapter 4 contains synopses of the papers and presentations by NAEP's sponsoring agencies (NAGB and NCES) and test development contractor (ETS). Because this material set the stage for discussants' presentations, these summaries are provided to help the reader understand the basis for discussants' remarks.

Chapters 5 through 7 highlight the comments made by discussants and other workshop participants. These chapters consider features of the market basket in relation to the NAEP redesign objectives that market-basket reporting has been conceptualized to address, that is, using innovations in measurement and reporting and simplifying NAEP's technical design (as described in Chapter 2). Because there was considerable overlap in the nature of the comments made during the workshop, Chapters 5 through 7 are organized around the central issues raised during the workshop rather than according to the chronological delivery of discussants' remarks panel by panel.

Chapter 8 concludes the report by highlighting issues to consider and resolve as NAEP's sponsors develop future plans for the market basket.

2

Origin of the Market-Basket Concept

This chapter traces the evolution of the NAEP market-basket concept. The first part of the chapter briefly describes NAEP's design between 1969 and 1996, providing foundation for material that appears later in the report. Discussion of a NAEP market basket began with the redesign effort in 1996 (National Assessment Governing Board, 1996). The second part of the chapter explores aspects of the 1996 redesign that relate to the market basket. The final section of the chapter discusses NAGB's most recent proposal for redesigning NAEP, focusing on the redesign objectives that pertain to the market basket (National Assessment Governing Board, 1999b).

NAEP'S DESIGN: 1969-1996

During the 1960s, the nation's desire grew for data that could serve as independent indicators of the educational progress of American children. With the support of the U.S. Congress, NAEP was developed and first administered in 1969 to provide a national measure of students' performance in various academic domains.

In the first decade of NAEP's administration, certain political and social realities guided the reporting of results. For example, at the time, there was strong resistance on the part of federal, state, and local policymakers to any type of federal testing, to suggestions that there should be a national curriculum, and to comparisons of test results across states (Beaton and

Zwick, 1992). To assuage these policymakers' concerns, NAEP results were reported in aggregate for the nation as a whole and only for specific test items, not in relation to broad knowledge or skill domains. In addition, to defuse any notion of a national curriculum, NAEP was administered to 9-, 13-, and 17-year-olds, rather than to students at specific grade levels.

In the early 1980s, the educational landscape in the United States began to change and, with it, the design of NAEP. The nation experienced a dramatic increase in the racial and ethnic diversity of its school-age population, a heightened commitment to educational opportunity for all, and increasing involvement by the federal government in monitoring and financially supporting the learning needs of disadvantaged students (National Research Council, 1999b). These factors served to increase the desire for assessment data that would help gauge the quality of the nation's education system. Accordingly, in 1984, NAEP was redesigned. Redesign, at this time, included changes in sampling methodology, objective setting, item-development, data collection, and analysis. Sampling was expanded to allow reporting on the basis of grade levels (fourth, eighth, and twelfth grades) as well as age.

Administration and sponsorship of NAEP has evolved over the years. Congress set the general parameters for the assessment and, in 1988, created the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP (Beaton and Zwick, 1992). NAGB is an independent body comprising governors, chief state school officers, other educational policymakers, teachers, and members of the general public. The Commissioner of the National Center for Education Statistics (NCES) directs NAEP's administration. NCES staff put into operation the policy guidelines adopted by NAGB and manage cooperative agreements with agencies that assist in the administration of NAEP. On a contractual basis, scoring, analysis, and reporting are handled by ETS, and sampling and field operations are handled by Westat.

Over time, as policy concerns about educational opportunity, the nation's work force needs, and school effectiveness heightened, NAGB added structural elements to NAEP's basic design and changed certain of its features. By 1996, there were two components of NAEP, *trend* NAEP and *main* NAEP.

Trend NAEP consists of a collection of test questions in reading, writing, mathematics, and science that have been administered every few years (since the first administration in 1969) to 9-, 13-, and 17-year-olds. The purpose of trend NAEP is to track changes in education performance over time, and thus, changes to the collection of test items are kept to a minimum.

Main NAEP includes questions that reflect current thinking about what students know and can do in certain subject areas. The content and skill outlines for these subject areas are updated as needed. Main NAEP encompasses two components: *state* NAEP and *national* NAEP. State and national NAEP use the same large-scale assessment materials to assess students' knowledge in the core subjects of reading, writing, mathematics, and science. National NAEP is broader in scope, covering subjects not assessed by state NAEP, such as geography, civics, U.S. history, world history, the arts, and foreign languages. National NAEP assesses fourth, eighth, and twelfth graders, while state NAEP includes only fourth and eighth graders.

NAEP's mechanisms for reporting achievement results have evolved over the years, but since 1996, two methods have been used: scale scores and achievement levels. Scale scores ranging from 0 to 500 summarize student performance in a given subject area for the nation as a whole and for subsets of the population based on demographic and background characteristics. Results are tabulated over time to provide trend information. Academic performance is also summarized using three achievement-level categories based on policy definitions established by NAGB: *basic*, *proficient*, and *advanced*. NAEP publications report the percentages of students at or above each achievement level as well as the percentage that fall below the basic category.

THE 1996 REDESIGN OF NAEP

The overall purpose of the 1996 redesign of NAEP was to enable assessment of more subjects more frequently, release reports more quickly, and provide information to the general public in a readily understood form. In the "Policy Statement for Redesigning the National Assessment of Educational Progress" (National Assessment Governing Board, 1996), NAGB articulated three objectives for the redesign:

1. Measure national and state progress toward the third National Education Goal[1] and provide timely, fair, and accurate data about student

[1]The third goal states: "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

achievement at the national level, among states, and in comparison with other nations.

2. Develop, through a national consensus, sound assessments to measure what students know and can do as well as what they should know and be able to do.

3. Help states and others link their assessments to the National Assessment and use National Assessment data to improve education performance.

The policy statement laid out methods for accomplishing these objectives including one that called for the use of innovations in measurement and reporting. Discussed was the use of domain-score reporting in which "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." Domain-score reporting was cited as an alternative to reporting results on "an arbitrary and less meaningful scale like the 0 to 500 scale" (National Assessment Governing Board, 1996:13).

The concepts of domain-score reporting and market-basket reporting were explained and further developed in a report from NAGB's Design and Feasibility Team (Forsyth et al., 1996). In this document, the authors described a market basket as a collection of items that would be made public so that users would have a concrete reference for the meaning of the score levels. They noted that the method for reporting results on the collection of items could be one that is more comfortable to users who are "familiar with only traditional test scores," such as a percent-correct metric (Forsyth et al, 1996: 6-26).

Forsyth and colleagues explored three options for the market basket. One involved creating a market basket the size of a typical test form (like scenario two in Figure 1), and a second called for a market basket larger than a typical test form (like scenario one in Figure 1). Their third option drew on Bock's (1993) idea of *domain referenced reporting*. With this option, a sufficient quantity of items would be developed so as to constitute an operational definition of skill in the targeted domain, perhaps as many as 500 to 5,000 items. All of the items would be publicly released. They

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" (National Education Goals Panel, 1994:13).

explain further that "having specified how to define a score based on a student responding to all of these items, it would be possible to calculate a predictive distribution for this domain score from a student's response to some subset of the items" (Forsyth et al., 1996:6-29).

Forsyth et al. (1996:6-26) also described the conditions under which market-basket items could be embedded into existing tests and stated that, under some plans, the market basket might allow for "embedding parallel 'market baskets' of items within more complex assessment designs. . . . Results from market basket forms would support faster and simpler, though less efficient, reporting, while information from broader ranges of items and data could be mapped into its scale using more complex statistical methods. . . . [R]eleased market basket forms could be made available to embed in other projects with strengths and designs that complement NAEP's." This use of the market basket falls under the second scenario in Figure 1 where the market basket is the size of a typical test form.

In 1997, NAGB adopted a resolution supporting market-basket reporting, which was defined as making NAEP "more understandable to a wide public by presenting results in terms of percent correct on a representative group of questions called a market basket." Additionally, the resolution stated that the market basket "may be useful in linking NAEP to state assessments" (National Assessment Governing Board, 1997:1).

NAEP DESIGN 2000-2010

Since the 1996 redesign, NAGB has continued to support extensive study of NAEP. Evaluation reports, reviews by experts, and commissioned papers highlight issues that bear on the 1996 redesign. Among these are when to change test frameworks, how to simplify NAEP's technical design, how to improve the process for setting achievement levels, and how NAEP results might be used to examine factors that underlie student achievement (National Assessment Governing Board, 1999b).

During extensive deliberations, NAGB recognized that NAEP was "being asked to do too many things, some even beyond its reach to do well, and was attempting to serve too many audiences" (National Assessment Governing Board, 1999b:2). Governing Board members found that NAEP's design was being overburdened in many ways. In its most recent redesign plan, "National Assessment of Education Progress: Design 2000-2010" (National Assessment Governing Board, 1999b), NAGB proposed to remedy these problems by refocusing the national assessment on what it does best, i.e., measure and report on the status of student achievement and change over time. NAGB also drew distinctions among the various audiences for NAEP products. Their report pointed out that the primary audience for NAEP *reports* is the American public, whereas the primary *users of its data* have been national and state policymakers, educators, and researchers (National Assessment Governing Board, 1996:6).

The Design 2000-2010 policy stated five over-arching principles for the conduct and reporting of NAEP (National Assessment Governing Board, 1999b:3):

1. conduct assessments annually, following a dependable schedule
2. focus NAEP on what it does best
3. define the audience for NAEP reports
4. report results using performance standards
5. simplify NAEP's technical design

Details of the initiative to develop a short form appeared under the policy objective of simplifying NAEP's technical design (National Assessment Governing Board, 1999b:7):

> Plans for a short-form of [NAEP], using a single test booklet, are being implemented. The purpose of the short-form test is to enable faster, more understandable initial reporting of results, and possibly for states to have access to test instruments allowing them to obtain NAEP assessment results in years in which NAEP assessments are not scheduled in particular subjects.

Like the 1996 redesign policy, the 2000-2010 design policy sought to use innovations in the measurement and reporting of student achievement, citing the short form as one means for accomplishing this objective. Further, the NAEP 2000-2010 design repeated the earlier objective of helping states and others link to NAEP and use NAEP data to improve education performance. (While this objective is not explicitly tied to the short form, suggestions for this use of the short form appeared in Forsyth et al., 1996.) The 2000-2010 policy goes a step beyond the 1996 policy in that it encourages states designing new state assessments to have access to NAEP frameworks, specifications, scoring guides, results, questions, achievement levels, and background data.

In addition, NCES has instituted a special program that provides grants for the analysis of NAEP data. NCES is now encouraging applications from states (and other researchers) to conduct analyses that will be of practical benefit in interpreting NAEP results and in improving education performance. The Design 2000-2010 Policy contains examples of studies in which NAGB has collaborated with states, such as Maryland and North Carolina, to examine the content of their state mathematics tests in light of the content of NAEP (National Assessment Governing Board, 1999b).

3
The Consumer Price Index Market Basket

Summary indicators are used in many contexts other than education. The Committee on NAEP Reporting Practices was interested in learning more about them and the experiences of other fields in making the results of complex summary measures understandable to the public. For example, although few people know how the Dow Jones Industrial Average Index of 30 "blue-chip" U.S. stocks is computed, most recognize it as an indication of the status of the stock market and understand what it means when the Dow Jones goes up or down. Similarly, calculation of unemployment rates is based on complex processes, but the end result is a single number that the public believes has immediate meaning.

Because parallels have been drawn between the CPI and the NAEP market basket, the committee arranged for a briefing on the CPI. At the committee's invitation, Kenneth Stewart from the Bureau of Labor Statistics (BLS) addressed committee members and workshop participants about the processes and methods used for deriving and utilizing the CPI (www.stats.bls.gov). Stewart's remarks are summarized below.

MAJOR USES OF THE CPI

Stewart explained that the CPI is a measure of the average change over time in the prices paid by urban consumers in the United States for a fixed basket of goods in a fixed geographic area. The CPI is widely used as an economic indicator and a means of adjusting other economic series (e.g., retail sales, hourly earnings) and dollar values used in government programs. It is the most widely used index for measuring inflation and aids in the formulation of fiscal and monetary policies and in economic decision-making. Stewart noted that the CPI measures the rates of changes in prices, not absolute levels.

CONSTRUCTION OF THE CPI MARKET BASKET

The BLS develops the CPI market basket on the basis of detailed information provided by families and individuals on their actual purchases. The market basket is reconstructed every decade using government survey data. The current CPI market basket is based on the Consumer Expenditure Survey conducted between 1993 and 1995. Approximately 30,000 families responded to this survey, providing information on their spending habits through quarterly interviews and by keeping comprehensive diaries of purchases.

Using the information supplied by these families, the BLS classified their expenditures into more than 200 item categories arranged into eight major groups: food and beverages; housing; apparel; transportation; medical care; recreation; education and communication; and other goods and services. The BLS then constructed a market basket of goods and services and assigned each item in the market basket a weight, or importance, based on its share of total family expenditures.

COMPUTATION OF THE MONTHLY INDEX

The BLS produces the monthly CPI using a sampling process. First, using decennial U.S. Census data, the BLS specifies a sample for the urban areas from which prices are to be collected and chooses housing units within each area for inclusion in the housing component of the CPI. A second sample of about 16,800 families each year serves to identify the places (outlets) where households purchase various types of goods and services. The final stage in the sampling process involves selecting the specific detailed item within each item category to be priced each month in a particular outlet. This selection is made using a random probability sampling method that reflects an item's relative share of sales at that particular store.

REPORTING AT SUBNATIONAL LEVELS

In addition to monthly release of the national CPI estimates, the BLS publishes monthly indexes for the four principal regions of the nation (Northeast, Midwest, South, and West), as well as for collective urban areas classified by population size. The BLS also publishes indexes for 26 local areas on monthly, bimonthly, or semiannual schedules. An individual area index measures how much prices have changed over a specific time interval in that particular area. However, due to the specifics of the design and sampling, indexes cannot be used for relative comparisons of the level of prices or the cost of living in different geographic areas. In fact, the composition of the market basket generally varies substantially across areas because of differences in purchasing patterns.

PARALLELS WITH EDUCATIONAL SETTINGS

In response to Stewart's presentation, workshop participants attempted to draw parallels between the CPI and the NAEP market-basket proposal. In doing so, they realized that the construction and measurement of the CPI market basket is somewhat different than that envisioned for the NAEP market basket. Creating a NAEP market basket using procedures modeled after the CPI would involve a process like the following: identify samples of teachers to participate in a survey; collect information from teachers (or schools) on the content and skills that they teach; classify the content and skills and sample from this listing to create the "market basket;" then, test students to determine their level of performance on this market basket of content and skills. This is quite different from the approach planned for the NAEP market basket. While the NAEP frameworks are developed by committees of experts familiar with school-level curricula, they are not based on surveys of what schools actually teach.

4
The Perspectives of NAEP's Sponsors and Contractors

In preparation for the market-basket workshop, the Committee on NAEP Reporting Practices asked NAEP's sponsors and contractors to respond to a series of questions regarding components of the market-basket concept. Specifically,

- What are the primary objectives for market-basket administration, market-basket reporting, and the short form?
- Who are the proposed users of market-basket materials and the short form?
- What types of inferences are expected to be supported by short-form and market-basket results?
- What is the status of research and development work on the market basket and short form?
- What are the Board's plans for pursuing work on the market basket/short form—with regard to the 2000 assessment and beyond?

Representatives from the sponsoring agencies (NAGB and NCES) and contracting agency (ETS) responded to these questions by preparing papers prior to the workshop. At the workshop, each representative made an oral presentation of the material covered in his paper. The committee asked workshop discussants to respond to papers as well as to the oral presentations. A summary of each paper and presentation is provided below to give the reader a context for the discussants' remarks. Summaries of discussants' remarks appear in Chapters 5 through 7.

A MARKET BASKET FOR NAEP: POLICIES AND OBJECTIVES OF THE NATIONAL ASSESSMENT GOVERNING BOARD

Roy Truby

During his presentation, Roy Truby, executive director of NAGB, explained the rationale for exploring the market-basket concept. Truby reminded the audience of the overall purpose of the NAGB redesign adopted in August 1996: to enable NAEP to assess more subjects more frequently, release reports more quickly, and provide information to the general public in a form that is readily understood. With these goals in mind, NAGB began considering alternatives including a NAEP market basket.

Under one alternative, students' results on a representative set of NAEP test items would be presented using percent correct scores (like scenario one in Figure 1). According to Truby, reporting NAEP results using a percent-correct metric would be more understandable for the general public and would allow for more timely reporting of NAEP results. Furthermore, the released items would be representative of the NAEP frameworks and would provide more clarity to the public about the content and skills tested by NAEP.

A second alternative involves the construction of a short, administrable NAEP test, the "short form," that would be representative of the content domain tested on NAEP (like scenario two in Figure 1). Results on the short form could be summarized using a percent-correct metric. The short form would provide additional data collection opportunities to state-NAEP users that are not part of the standard NAEP schedule, such as testing in off years or in other subjects not assessed at the state level. Truby described how some people envision using a short form:

> If short forms were developed and kept secure, they could provide flexibility to states and any jurisdiction below the state level that were interested in using NAEP for surveying student achievement in subjects, grades, and times that were not part of the regular state-NAEP schedule. Once developed, such market-basket forms should be faster and less expensive to administer, score, and report than the standard NAEP, and could provide score distributions without the complex statistical methods on which NAEP now relies. This might help states and others link their own assessments to NAEP, which is another important objective of the Board's redesign policy.

Truby noted that NAGB has approved a policy for "market-basket reporting" and has approved a pilot for a "market-basket short form," but added that the details associated with these components of the market-basket concept have not yet been thoroughly investigated.

Truby concluded by explaining that ETS is currently investigating the market-basket concept by conducting a pilot study in grade four mathematics as part of NAEP 2000. This study involves preparation of NAEP short forms (scenario two in Figure 1). Details of the study are described below (see section entitled "NAEP's Year 2000 Market Basket Study: What Do We Expect to Learn?"). Based on the findings from the pilot study, NAGB might pursue similar studies in other content areas and grades.

SIMPLIFYING THE INTERPRETATION OF NAEP RESULTS WITH MARKET BASKETS AND SHORTENED FORMS OF NAEP

Andrew Kolstad

Andrew Kolstad, senior technical advisor for the assessment division of the National Center for Education Statistics, traced the history of NAEP's reporting methods. During the 1970's, NAEP reported its results in terms of the percentage of students who correctly answered each test item (item p-values) as well as the average percent of items answered correctly (average percents correct). Since many items were released along with information on the percentages of students who answered them correctly, reporting item p-values offered specific and concrete information to data users. While this procedure gave data users a good sense of what was covered on the test and how students performed, it had at least two drawbacks: first, it required the development of a substantial number of new items in each assessment cycle in order to replace those released; and second, data users had a hard time understanding the overall picture of student performance on collections of items.

Reporting the average percent correct over a set of items helped to overcome the second problem because this gave an overall picture of student performance. Nevertheless, several drawbacks remained. First, this method provided only one piece of information about the performance distribution, namely, the mean percent correct, and did not provide any information about the rest of the performance distribution. Second, the percent-correct summary statistic also suffered from the limitation that if the set of items changed, then the average percent correct would also change. In other words, if the sample of items was relatively easy, the average percent correct would be higher than if the sample of items was relatively hard. This created interpretation problems, particularly with interpretations of trends in performance. The composition of items changed from one assessment to the next as items were dropped from the assessment pool (because they had been released). Third, making generalizations about students' performance on a fixed collection of *administered* items to their expected performance on other *non-administered* items, albeit items from the same frameworks, was problematic. The idea that test questions were sampled from a pool of potential items was not yet formalized.

In the early 1980's, ETS became the test development contractor for NAEP. ETS began using item response theory (IRT)[1] scaling, which alleviated many problems of interpretation deriving from the practice of reporting percent corrects for subsets of items. Item response theory scales items according to the probability of a correct answer, given the proficiency level of the examinee and the item's discriminating power, difficulty, and susceptibility to examinee guessing. It relies on assumptions that, if met, result in proficiency estimates that, theoretically, are not dependent on the particular subset of items administered and that yield item parameter estimates that are relatively independent of the group of students taking the items.

With the introduction of IRT scaling, average IRT-based scale scores replaced average percent correct scores for NAEP reporting. However, many data users regard IRT-based scale scores as substantially less interpretable than percent correct scores. While NAEP still releases a few items as illustrative of the assessment, a substantially smaller proportion of items are released (reducing development costs). Also, item p-values have been replaced by IRT-based item mapping. Item mapping provides an interpretation of the relative difficulty of test items, as well as of the performance of examinees relative to items of differing difficulty. However, the item mapping procedure has been subject to controversy because it requires somewhat arbitrary decisions about the probability thresholds used.

Kolstad believes that the use of market-basket reporting and percent correct scores in conjunction with IRT-based scaling—as supported by NAGB and suggested in an NRC report, *Grading the Nation's Report Card*, (National Research Council, 1999b)—could improve understanding of

[1]Item response theory is a statistical model that calculates the probability each student will get a particular item correct as a function of underlying ability; for further discussion of IRT modeling, see Lord (1980).

NAEP reporting. Kolstad's conception of market-basket reporting is one in which IRT-based scaling would be used to project the expected percent correct on a market basket of items (scenario one in Figure 1), an approach that does not require the actual administration of those items (provided that the IRT-based item parameters are known).

Kolstad pointed out that the proposed market-basket reporting of expected percent correct scores on a market-basket collection of items is better than the average percent correct used in the early days of NAEP for several reasons. One is that the IRT-based approach would include publication of the market-basket set of items that constitute the pool of questions, which could improve understanding of item content. Because the items need not be administered during each assessment cycle in order to be used for this kind of reporting, developmental costs would be minimized. Furthermore, IRT-based projections can differentiate between performance on easy and hard test questions. If the difficulty composition of the items in the market-basket set changes, the results can be appropriately adjusted through the use of IRT-based projections. Unlike the use of average percents correct in the early days of NAEP, the use of IRT-based projections of expected percent correct on a market basket of items enables prediction of performance on other items from the same framework that did not happen to be included in NAEP's assessment instrument.

Kolstad believes that focus groups and empirical studies should be conducted to verify that the market-basket metric—expected percent correct—is indeed simpler for consumers to understand. Kolstad also cautioned that invalid inferences about achievement-level performance would be drawn from empirical average percent correct scores, unless they are based on IRT projections, and suggested careful consideration of potential misinterpretations.

EVIDENTIARY RELATIONSHIPS AMONG DATA-GATHERING METHODS AND REPORTING SCALES IN SURVEYS OF EDUCATIONAL ACHIEVEMENT

Robert Mislevy

During his presentation, Robert Mislevy, distinguished research scholar with ETS, laid the conceptual groundwork for the technical and measurement issues involved in market-basket reporting. Mislevy distinguished between data collection methods and data reporting methods. To Mislevy, the term *data collection methods* refers to the means of gathering performance data, including information that bears on the test questions comprising the market basket. The term *data reporting methods* refers to the mechanisms used for translating performance data into a reporting metric, including performance on a market basket of items.

Mislevy described five approaches for collecting data: (1) a single test form, (2) parallel test forms, (3) tau-equivalent test forms, (4) congeneric test forms, and (5) arbitrary test forms. The first two—a single form and parallel forms—are the formats typically associated with testing programs. Under a single test form approach, one form of the test is developed, and all students take the same form. Under a parallel test forms approach, multiple equivalent forms are developed. The forms contain different items but are sufficiently similar to be considered interchangeable. They contain the same kinds and numbers of items, tap the same mix of underlying skills, are used for the same purposes, and are administered under similar conditions. Forms that are considered parallel have equal raw score means, equal standard deviations, and equal correlations with other measures for any given population.

Test forms that are either tau-equivalent or congeneric measure similar constructs but do not meet the stringent criteria of parallel forms. Tau-equivalent forms are closely related but not strictly parallel. For example, they may have the same mix of items but may differ with regard to the numbers of items. Congeneric forms are less closely related and, for example, may include the same essential mix of knowledge and skills but may differ in terms of the number, difficulty, and sensitivities of the items included.

Arbitrary forms are only generally related to the same content domain, and, for example, may differ considerably as to the mix, number, format, or content of items. While arbitrary forms may be similar with respect to timing, balance, or other characteristics, they have not been constructed to be parallel, tau-equivalent, or congeneric. For instance, one arbitrary form may focus on multiple-choice items while another may primarily use constructed response items.

Mislevy drew distinctions among three reporting metrics: the observed score metric, the true score metric, and the latent trait metric. The observed score metric is based on a simple tally of the number of right answers or the number of points received. Observed scores can quickly be converted to a percent correct scale by dividing the number correct score by the total number of questions or points. However, observed scores have the problem, mentioned by Kolstad, of being tied to the composition and difficulty of the particular test form.

Reporting on a true score metric involves making a transformation from the observed score to the *expected* or *predicted* distribution of an individual's true score (it is a *predicted* score, since an individual's true score is never known). There are a number of advantages to reporting on an IRT-based true-score scale since such scores can be placed on a percent correct scale. However, given that reporting on a true score metric means working with predictive distributions of individuals' true scores, the transformation is much more complex. In particular, there is no one-to-one mapping between an observed score and an expected score.

Finally, the latent trait metric refers to the IRT-based proficiency estimates. Using this metric requires estimation of the latent trait distribution. While this process involves a complicated transformation from observed scores, it has the advantage that, when IRT assumptions are met, the distributions are not content specific. Further, the latent trait distributions could be transformed to an expected percent correct metric. NAEP currently estimates latent trait distributions that are converted to scaled score distributions for reporting. Current procedures for NAEP do not, however, transform latent trait distributions to expected percent correct metric.

Market-basket scores could be based on intact, *administrable* forms (like scenario two in Figure 1), like the proposed short forms. To support inferences about performance on one version of the short form to performance on another version of the short form, the short forms would need to be strictly parallel in the full technical sense. Creating parallel short forms would not be a sufficient condition to support inferences from scores on the short forms to the main NAEP scale, however. Much more complex statistical procedures would be needed to enable generalizations about performance on main NAEP based on performance on the short form.

Alternatively, market-basket scores could be based on *synthetic* forms (scenario one in Figure 1). A synthetic form is a form proportionately representative of the content and skill domain but too long to administer in its entirety to a single student. The concept of a synthetic form is similar to the concept of a market basket as it is used in other settings; i.e., a sampling of items intended to be representative of some larger whole (e.g. the content and skills tested by NAEP). Summarizing performance on synthetic forms using a percent-correct or observed-score reporting metric would be quite complex, as no one student would take the entire test. This approach to market-basket reporting would have to be based on hypothetical observed scores for the synthetic form. And results would be modeled projections from data on some other forms.

Mislevy then proceeded to develop a framework for analyzing the complexities involved in collecting and reporting data. He identified various ways of collecting data and of reporting it, then described the kinds of inferences supported by various reporting procedures and their appropriateness for the different collection methods. Throughout the paper, Mislevy emphasized that all combinations of collection and reporting procedures involve tradeoffs. Some methods are simpler and quicker than others but do not support the desired inferences. Other methods yield generalizable results but at the expense of simplicity. A key issue for the NAEP market-basket concept is the desire to have market-basket results that are comparable to main NAEP results. The goal is to be able to make inferences about performance on the market-basket collection of items compared to a national benchmark (main NAEP). This goal becomes particularly challenging under scenario two (see Figure 1), where a short form is released to states and districts for their use and scores are to be derived quickly and are intended to be comparable to main NAEP.

In his paper, Mislevy systematically laid out the issues that need to be resolved before decisions are made on data gathering and data reporting models for the market basket. Through his analysis, Mislevy explored the competing goals of simplicity of methods versus generalizability of results. The simplest methods would use parallel, intact forms for data collection and observed scores for reporting. Questions remain as to how generalizable the forms and scores would be to the content domain, if based on this approach. The most generalizable results would be based on a system of arbitrary forms, with performance reported as the latent trait distribution, as is currently done with NAEP. However, this is also one of the most complex of the possibilities.

NAEP'S YEAR 2000 MARKET-BASKET STUDY: WHAT DO WE EXPECT TO LEARN?

John Mazzeo

John Mazzeo, executive director of ETS's School and College Services, told workshop participants that the ETS year-2000 study on the market basket was designed with three goals in mind: (1) to produce and evaluate a market-basket report of NAEP results; (2) to gain experience with constructing market-basket short forms; and (3) to conduct research on the methodological and technical issues associated with implementing a market-basket reporting system. The study involves the construction of two test forms (also referred to as *administrable* or *short* forms) for grade four mathematics. Although these forms were designed to be parallel, some of the research will evaluate the extent to which the forms meet the necessary assumptions to be considered parallel.

According to Mazzeo, the test developers hope that the study will serve as a learning experience regarding the construction of alternate short forms. Whereas creating intact test forms is a standard part of most testing programs, this is not the case with NAEP. Due to its many content areas and the need to limit the length of the testing time, NAEP uses a matrix sampling design to obtain a representative sample of students taking each subject-area assessment. Under this design, blocks of items within each content domain are administered to groups of students, making it possible to administer a large number and range of items during a relatively brief testing period. Consequently, each student takes only a few items in a given content area—too few to serve as a basis for individual scores.

Because NAEP's current system for developing and field testing items was set up to support the construction of a system of "arbitrary" test forms in an efficient matter, it does not yet have guidelines for constructing market baskets or intact tests. That is why study of the creation of such forms is under way.

The short forms were constructed by a NAEP test development committee that had been instructed to try to identify a set of secure NAEP items that were high quality exemplars of the pool; that matched the pool with respect to content, process, format, and statistical specifications; and that could be administered within a 45-minute time period. The committee constructed two forms with approximately 30 items organized into three distinct blocks, each to be given during separately timed 15-minute test sessions. One of the short forms contains previously administered secure items; the other contains new items. Both forms will be given to a random sample of 8,000 students during the NAEP 2000 administration. These forms will be spiraled[2] with previously administered NAEP materials to enable linking to NAEP.

Mazzeo said that the year-2000 study is expected to result in three

[2]Spiraling is an approach to form distribution in which one copy of each different form is handed out before spiraling down to a second copy of each form and then a third and so forth. The goals of this approach are to achieve essentially random assignment of students to forms while ensuring that an essentially equal number of students complete each form.

products: (1) one or more secure short forms; (2) a research report intended for technical audiences that examines test development and data analytic issues associated with the implementation of market-basket reporting; and (3) a report intended for general audiences.

At the time of the workshop, ETS's plans for the market-basket reports had not been formalized. According to Mazzeo, some of the features being considered include

- National and state-level NAEP results (average scores and achievement level percentages) expressed in a market-basket metric (e.g. percent correct). The reporting of such results could be confined to "total-group" scores or it could be extended to include national and state results by gender, race/ethnicity, parental education, and other standard NAEP reporting groups.
- Release of all, or a sample, of the items that make up the short form as well as performance data. Mazzeo noted that the text of the items, scoring rubrics, and sample student responses might also be provided.
- A format and writing style appropriate for a general public audience.
- Electronic reporting.

The research study will investigate market-basket reporting under two configurations, one in which the short form would be made available, and one in which it would not. ETS researchers will continue to study alternative analytic and data collection methods. One of the studies planned involves conducting separate analyses of the data using methods appropriate for arbitrary forms, methods appropriate for congeneric forms, and methods appropriate for parallel forms. Each of these sets of analyses will produce results in an observed score metric as well as a true score metric.

The study calls for comparing results from the arbitrary forms with results from other approaches to obtain empirical evidence about which data gathering options are most viable for the market-basket concept. These comparisons will focus on the degree of similarity among the sets of results. If the congeneric and parallel forms models (which are based on strong assumptions but involve less complex analytic procedures) produce the same results as the arbitrary forms model (which makes the weakest assumptions but involves the most complicated analysis), then the simpler data collection and analytic procedures may be acceptable. Comparisons of observed score and true score results for each of the approaches will inform decisions about which type of reporting scale should be used.

The study will also provide data that can be used to evaluate context effects. The administration design will yield multiple estimates of item parameters for some of the market-basket items. Comparisons of the parameter estimates will enable investigation of the magnitude of context effects.

The year-2000 study will entail evaluation of the potential benefit of using longer market baskets. According to Mazzeo, the 31-item short forms were chosen out of consideration for school and student burden, increasing difficulties in obtaining school participation in NAEP, and the conviction that, "to be effective, a publicly released market basket of items should be of modest size." Other decisions regarding test length could also be made, such as Darrell Bock's domain score reporting approach. Under this approach, the entire item pool is released, and the reporting scale is defined in terms of scores on the full item pool. Mazzeo reminded participants that a longer collection of items would permit more adequate domain coverage and produce more reliable results.

5
Using Innovations in Measurement and Reporting: Releasing a Representative Set of Test Questions

A basic component of the NAEP market basket is the release of a representative set of test items. While NAEP has always released samples of items, under plans for the market basket, many more items would be released with the goal of representing the content domain for a specific subject. For example, the collection of items for fourth grade mathematics would consist of the appropriate number and mix of test items needed to represent the domain of fourth grade mathematics, as defined by NAEP's frameworks. The Committee on NAEP Reporting Practices was interested in the purposes that might be served by such a release of items and requested that workshop discussants consider who might use this information and how it might be used. The discussion below attempts to summarize the major points made by the speakers.

DEMYSTIFYING THE TEST

Workshop discussants remarked that an aura of mystery surrounds testing. In their interactions with the public, they have found that people question why so much time is devoted to testing and are unsure of how to interpret the results. The public is not fully aware of the material that is covered on achievement tests, the skills that students are expected to demonstrate, and the inferences about student achievement that can be drawn from test results. Moreover, the public does not always see the link between assessment programs and school reform efforts.

Marilyn McConachie, Illinois State Board of Education member, summarized these perceptions most succinctly, saying:

> Analysis of public understanding of test results in Illinois parallels national commentary on NAEP and other large-scale testing. Put simply, the public believes too much time is spent on testing and doesn't really understand what students know and are able to do or whether performance is good enough. These beliefs appear to erode support for state testing (and for NAEP).

According to McConachie, only when tests are "demystified" will the public understand what is being tested and why, and only then will the public support the continued gathering of this important information.

Other workshop discussants commented that public release of test items, scoring rubrics, and student work samples could serve to further public understanding of what NAEP tests. Many felt that the public is not generally aware of the difficulty level of the material covered on achievement tests today. For most people, the basis of comparison is their own school experiences, but curricula and expectations for students have changed. Joseph O'Reilly, director of assessment for Mesa, Arizona, public schools offered an example that illustrates this perception:

> The public does not seem to understand the difficulty of the concepts taught today compared to when they were in school, especially in mathematics. For example, basic trigonometric angles are commonplace in algebra today but were not covered until much later in the curriculum, if at all, thirty years ago.

O'Reilly believes that the release of a large collection of items would be very useful in communicating the higher levels of expectations of tests like NAEP. Additionally, because many state testing programs cover content similar to that tested by NAEP, the information learned from NAEP's release of items could also increase understanding of state and local assessments.

STIMULATING DISCUSSION AMONG TEACHERS

Workshop speakers observed that release of a large number of representative items could be used to stimulate discussion among teachers regarding the format and content of test questions. In addition, review of the released items could facilitate discussions about ways to align local curricula and instructional practices with the material covered on the national assessment. Discussants explained that it is often difficult to draw conclusions about their students' NAEP performance because it is not clear whether the material tested on NAEP is covered by their curricula or when it is covered.

The superintendents representing the First in the World Consortium provided examples of how they utilized information from their students' participation in the Third International Mathematics and Science Study (TIMSS) to guide local curricular and instructional changes. The consortium participants did not have access to actual test items but, instead, had information on the specific topics covered and the content areas assessed. Teams of teachers from consortium schools examined the topics covered on the TIMSS assessment. They also considered when and how these topics were presented in their curricula, discussing such issues as: at what grade level the topic is first introduced in their programs; at what grade level mastery of the topic is expected; and how the topic is reinforced over the grades. Analysis of their TIMSS results, particularly of students' strengths and weaknesses, in comparison to teaching and instructional practices, allowed the participating school systems to identify needed changes in their curricula.

This use of TIMSS materials exemplifies one potential use of released NAEP materials. While the First in the World Consortium school systems did not see actual items, they had the benefit of receiving information on the topics and content areas covered by TIMSS items. For school systems and others to realize similar benefits from the released NAEP materials, items would need to be categorized into frameworks and content areas. Otherwise, teachers and other users might categorize items themselves, perhaps incorrectly, in order to make inferences about the relationships between material tested and content covered by their curricula.

Workshop discussants also suggested that it would be useful to associate the released items with the NAEP achievement level category of students expected to answer the question correctly. This matching of items with achievement levels would demonstrate the content and skills students should have mastered at each level, which would facilitate understanding of the assessment. For example, if teachers were able to view items that illustrated what students scoring at the proficient level in fourth-grade mathematics should be able to do, they would be able to adjust their teaching accordingly. Used diagnostically, this information could help students progress from below basic to basic, from basic to proficient, and from proficient to advanced.

ENCOURAGING IMPROVED STATE AND LOCAL TESTING

NAEP often serves as a role model for the development of state and local assessments and the policy governing those assessments. During the committee's earlier workshop on reporting district-level NAEP results, participants commented that NAEP's frameworks, its innovative item design, and its use of achievement-level reporting have greatly influenced assessments around the country (National Research Council, 1999c.) Similar observations were made at the market-basket workshop.

Participants in the market-basket workshop thought that a large-scale release of NAEP items and related test materials could potentially improve state and local assessment programs. NAEP produces high-quality items and test materials. Allowing test developers to view large amounts of NAEP test materials (test questions as well as rubrics for scoring constructed response items) could therefore have a positive effect on the quality of item design for state and local assessments. The release of high-quality NAEP materials could also help revamp classroom-based assessments. Furthermore, in their opinion, policymakers would be able to see the breadth and depth of the content and skills assessed and the grade levels at which students are expected to have mastered certain subject matter—information that could play an important role in redefining curricula. Participants emphasized, however, that for such objectives to be realized, item release would need to be both large and representative of the domain assessed.

6
Using Innovations in Measurement and Reporting: Reporting Percent Correct Scores

A second aspect of the NAEP market basket is reporting results in a metric easily understood by the public. For some time, NAEP has summarized performance as scale scores ranging from 0 to 500. However, it is difficult to attach meaning to scores on this scale. What does a score of 250 mean? What are the skills of a student who scores a 250? In which areas are they competent? In which areas do they need improvement?

Achievement level reporting was introduced in 1990 to enhance the interpretation of performance on NAEP. NAEP's sponsors believe that public understanding could be further improved by releasing a large number of sample items, summarizing performance using percent correct scores, and tying percent correct scores to achievement level descriptions. Since nearly everyone who has passed through the American school system has at one time or another taken a test and received a percent-correct score, most people could be expected to understand scores like 90%, 70%, or 50%. Unlike the NAEP scaled scores, the percent correct metric might have immediate meaning to the public.

PERCENT CORRECT METRIC: NOT AS SIMPLE AS IT SEEMS

At first blush, percent correct scores seem to be a simple, straightforward, and intuitively appealing way to increase public understanding of NAEP results. However, they present complexities of their own. First, NAEP contains a mix of multiple-choice and constructed response items.

In preliminary stages of scoring, multiple-choice items are awarded one point if answered correctly and zero points if answered incorrectly. Answers to constructed response items are also awarded points, but for some constructed response questions, six is the top score, and for others, three is the top score. For a given constructed response item, higher points are awarded to answers that demonstrate more proficiency in the particular area. Furthermore, a specific score cannot be interpreted, even at this preliminary stage, as meaning the same level of proficiency on different items (e.g., a four on one item would not represent the same level of proficiency as a four on another item). This situation becomes more complex at subsequent stages of IRT-based scoring and reporting, and the concept of "percent correct" becomes meaningless. Therefore, in order to come up with a simple sum of the number of correct responses to test items that include constructed response items, one would need to understand the judgment behind "correct answers." What would it mean to get a "correct answer" on a constructed response item? What would be considered a correct answer? Receiving all points? Half of the points? Any score above zero?

As an alternative, the percent correct score might be based, not on the number of questions, but on the total number of points. This presents another complexity, however. Simply adding up the number of points would result in awarding more weight to the constructed response questions than to the multiple-choice questions. For example, suppose a constructed response question can receive between one and six points, with a two representing slightly more competence in the area than a one but clearly not enough competence to get a six. Compare a score of two out of six possible points on this item versus a multiple-choice item where the top score for a correct answer is one. A simple adding up of total points would give twice as much weight to the barely correct constructed response item as to an entirely correct multiple-choice item. This might be reasonable if the constructed response questions required a level of skill much higher than the multiple-choice questions, such that a score of two on the former actually represented twice as much skill as a score of one on the latter. Since this is not the case for NAEP questions, some type of weighting scheme is needed. Yet, weighting schemes also introduce complexity to the percent correct metric.

A number of workshop participants addressed the deceptive simplicity of percent correct scores. Several pointed out that the public already has difficulty understanding terms that psychometricians use, such as *national percentile rank* or *grade-level equivalents*. As a result, assessment directors spend a good deal of time trying to ensure that policymakers and the public make the proper inferences from test results. The danger of the percent correct score is that everyone might think they understand it due to their own life experience, when, in fact, they do not.

Still, it should be pointed out that the percent correct metric has much intuitive appeal. If used correctly it might be of great benefit in increasing understanding of NAEP. Moreover, all statistics are susceptible to misuse, percent correct as well as more complex statistics. As Ronald Costello, assistant superintendent public schools in Noblesville, Indiana, observed:

> It doesn't matter what the statistic is, it still will be used for rank ordering when it gets out to the public. There are 269 school districts in Indiana. When test results come out, there's a 1 and a 269. The issue is why are we testing students and what do we want to do with the results.

Costello concluded by saying that more important than the statistic is the use of the results. Attention should be focused on making progress in educational achievement, and the statistic should enable evaluation of the extent to which students have progressed.

DISCONNECT WITH PUBLIC PERCEPTIONS OF "PROFICIENT"

One plan for the NAEP percent correct scores is to report them in association with the NAEP achievement levels. At the workshop, Roy Truby presented a document that showed how this might be accomplished based on results from the 1992 NAEP mathematics assessment (Johnson et al., 1997). An excerpt appears in Table 1. This table displays percent correct results for test takers in grades four, eight, and twelve. Column 2 presents the overall average percent correct for test-takers in each grade. Columns 3-5 show the percent correct scores for each achievement level category associated with the minimum score cutpoint for the category. For example, the cutpoint for the fourth grade *advanced* category (Column 3) would be associated with a score of 80 percent correct. A percent correct score of 33 percent would represent performance at the cutpoint for twelfth grade's *basic* category.

Speakers cautioned that the percent correct scale used in Table 1 is unlike that understood by the public. In their opinion, people typically regard 70% as a passing score; scores around 80% as indicating proficiency; and scores of 90% and above as advanced. What would members of the

TABLE 1 Example of Market Basket Results

(1) Grade	(2) Average Percent Correct Score[a]	Cut Points by Achievement Level (3) Advanced	(4) Proficient	(5) Basic
4	41%	80%	58%	34%
8	42	73	55	37
12	40	75	57	33

[a]In terms of total possible points.
Note: The information in Table 1 is based on simulations from the full NAEP assessment; results for a market basket might differ depending on its composition.

general public think when they saw that the average American student scored less than 50% on the test represented in the table? Would this scheme be an appropriate basis for the public's evaluation of the level of education in schools today? According to one speaker:

> Most test directors would understand why this might be, but no teacher, parent, or member of the public would consider 55% proficient. They would consider that score as representing "clueless" perhaps, and would think even less of the test and the educators that would purport to pass off 55% as proficient.

CONVERSION TO GRADES

While most Americans have at one time or another taken a test and received a percent score, generally that percent score was converted to a letter grade. Although associating percent correct scores with an achievement level might increase public understanding of NAEP, many people would still be tempted to convert the scores to letter grades, and their conversions might not be accurate. Richard Colvin offered his perspective as an education reporter for the *Los Angeles Times*:

> On its own, a percent correct score is only slightly more meaningful than a scale score. The reason is that, in school, percent correct is translated into a grade: 93% or above for an "A," 85% to 93% for a "B," and so forth. If you were to put out a percent correct score for the market basket of items, I assure

> you that journalists will push you to say what letter grade it represents. And, if you aren't willing to do that, journalists will do it for you.

Other participants echoed this concern, noting that the public would need a means for interpreting and evaluating percent correct scores.

ONE STEP FORWARD, TWO STEPS BACK

As described by Andrew Kolstad, senior technical advisor with NCES, in the first decade of NAEP, the percent correct metric was used for reporting results. Use of item response theory (IRT), beginning in the early 1980s, solved many of the interpretation problems that stemmed from the practice of reporting percent correct scores for subsets of items. Therefore, some workshop discussants wondered why NAEP would want to return to the metric used in its early years. David Thissen, professor of psychology at the University of North Carolina, emphasized this pointing out that "NAEP's use of the IRT scale in the past two decades has done a great deal to legitimize such IRT procedures with the result that many other assessments now use IRT scales. . . . [A] potential unintended consequence of NAEP reporting on a percent correct scale might be to drive many other tests, such as state assessments, to imitation."

NAEP uses some of the most sophisticated and high-quality analytic and reporting methods available. If NAEP moves away from such procedures to a simpler percent correct metric, others will surely follow suit. Many discussants maintained that they did not see the benefits of the simpler metric.

DOMAIN REFERENCED REPORTING

During his comments on technical and measurement considerations, Don McLaughlin, chief scientist for the American Institutes of Research, reminded participants that the desired inferences about student achievement are about the content domain, not about the set of questions on a particular test form. The interest is not in the percent of items or points correct on a form. Instead, the interest is in the percent of the domain that children have mastered.

Domain referenced reporting was cited as an alternative to market-basket reporting. Domain referenced reporting is based on large collections of items that probe the domain with more breadth and depth than is possible through a single administrable test form. As described by Darrell Bock, domain referenced reporting involves expressing scale scores in terms of the *expected* percent correct on a larger collection of items representative of the specified domain. The expected percents correct can be calculated for any given scale score using IRT methods and the estimated item parameters of the sample of test questions (see Bock et al., 1997). Bock further explained the concept of domain referenced reporting saying:

> [A] domain sample for mathematics might consist of 240 items by selecting 4 items to represent each of the 60 cells of the domain specification described by [John] Mazzeo. These items could be drawn from previously released items from the NAEP assessment or from state testing programs. Their parameters could be estimated by adding a small number of additional examinees in each school participating in the [NAEP] and administering them special test forms containing small subsets of the domain sample, similar to those proposed for the market basket.
>
> The point is to publish the 240 items in a compendium organized by the content, process, and achievement level categories. . . . For graded open-ended items, the rating categories should also be described and the "satisfactory" and "unsatisfactory" categories identified. The objective of this approach is not only to provide sufficient items from which readers of the assessment report can infer the knowledge and skills involved in mathematics achievement, but also, by publishing the compendium well before the assessment takes place, to encourage its use as a aid to instruction and self-study and as a basis for comment and explication in the media. When the results finally appear, there will then exist a ready and well-informed audience for the assessment report.

Bock went on to offer as an example of such a compendium the procedures used by the Federal Aviation Administration (FAA) to license private pilots. All 915 items that could potentially appear on the exam are published. And all potential pilots receive this compendium so that they may study the necessary material.

7
Simplifying NAEP's Technical Design: The Role of the Short Form

A third component of the NAEP market basket is the concept of the NAEP short form. Under this notion, the short form would be the vehicle for releasing items, providing a set of questions that could serve as an administrable test form. One or more versions of the short form would be released for public use, while other versions of the short form would be kept secure for use in conjunction with national NAEP and, perhaps, with state and local assessment programs.

To guide policy and decision making on the measurement issues pertaining to the short form, NAGB adopted the following principles (National Assessment Governing Board, 1999a):

Principle 1: The NAEP short form shall not violate the Congressional prohibition to produce, report, or maintain individual examinee scores.

Principle 2: The Board shall decide which grades and subjects shall be assessed using a short form.

Principle 3: Development costs, including item development, field testing, scoring, scaling, and linking shall be borne by the NAEP program. The costs associated with use, including administration, scoring, analysis, and reporting shall be borne by the user.

Principle 4: NAEP short forms intended for actual administration should represent the content of corresponding NAEP assessment frameworks as fully as possible. Any departure from this principle must be approved by the Board.

Principle 5: Since it is desirable to report the results of the short form using the achievement levels, the content achievement level descriptions should be considered during the development of the short form.

Principle 6: All versions of the short form should be linked to the extent possible using technically sound statistical procedures.

The proposed short form was the topic of considerable discussion during the workshop. The text below attempts to capture the discussions, highlighting the issues that seemed most important to participants. Addressed first are speakers' comments regarding potential uses of the short form and the data gathered from it. Addressed later are problems associated with the short form.

POTENTIAL USES OF THE SHORT FORM

Benchmarking and Other Comparisons

In September 1999, the Committee on NAEP Reporting Practices held a workshop on reporting district-level NAEP results. One of the clearest messages from participants in the workshop was that states and local jurisdictions want to be able to make comparisons of their achievement test results—comparisons with other jurisdictions and comparisons against national benchmarks. At present, state assessment programs enable within-state comparisons among schools and districts, but they do not allow for comparisons across state boundaries. State NAEP enables comparisons of achievement results from state to state but does not allow for comparisons among districts and schools, since results are not reported at the district and school levels.

District-level workshop participants indicated that comparisons would serve a number of important purposes. For example, comparisons among districts that share common social, economic, and demographic characteristics would help policymakers set reasonable expectations for student achievement. They also would allow districts to identify other districts like them that are performing better, thereby, stimulating discussions about education practices that work well (National Research Council, 1999d).

Workshop participants were also interested in having an external barometer against which to validate results from state and local assessments. Local jurisdictions were attracted to the prospect of being able to compare their students' performance to national benchmarks. They felt that having such information would open up discussions about local standards, curricula, and assessment programs (National Research Council, 1999d).

Participants in the committee's market-basket workshop voiced similar interests and concerns. They liked the idea of having school-level or district-level results that could be compared to NAEP. Many had heard requests from their state legislators for national data to be used as benchmarks in setting goals for improving student achievement. According to Marlene Hartzman, an evaluation specialist with the public schools in Montgomery County, Maryland, "We have more data than we need, but we don't have *what* we need—a national benchmark." They considered the short form to be the mechanism for obtaining benchmarking data, assuming the short form would yield school-level and district-level results.

Speakers commented that benchmarking data would help school administrators assess students' strengths and weaknesses and would enable them to target areas for improvement. Open discussion about weak areas could serve to identify education practices that work. Participants also pointed out that the short form might be a means for encouraging schools to participate in NAEP because it could be used to give schools and districts feedback on their students' NAEP performance, something NAEP does not currently provide.

Embedding the Short Form in Existing Assessments

Prior to the market-basket workshop, discussants were asked to consider the ways they would use the short form, if it were available. Many said that they would want to "embed" the short form in their state or district assessments to obtain results that could be compared with both the local assessments and with NAEP. Marilyn McConachie expanded on this idea for using the short forms, saying, "If these forms could be embedded into state tests, this would help us considerably in two ways: linking to NAEP and providing a strong sample from our state [by supplementing the sample selected for NAEP participation]. Linking to NAEP would help meet the state accountability policy's requirement for national benchmarking."

In preparation for the workshop, Joseph O'Reilly, past president of the National Association of Test Directors, conducted an informal survey of some test directors. Highlighting his findings, O'Reilly stated that:

> Overall, the test directors . . . were almost unanimous in support of a short form or market-basket form of NAEP if it could be incorporated into the state assessment system. I think that test directors are assuming that the proposed short form would be 10-15 items that could be used to scale the rest of the items on a NAEP scale, just as one embeds items on different levels of a test so that you can obtain a common scale across forms or grades.

O'Reilly reported that respondents saw great value in obtaining normative data on NAEP-like tests but were adamant about incorporating items from a short form into existing tests. He found that they wanted the information a short form would provide but would not support additional testing (or additional time for testing) to obtain it. Workshop participants expressed similar viewpoints saying they would consider administering the short form as a separate, common test given to all students but added that they would have to replace one of their regular assessments to do so.

Comparing Local Curricula and Assessments with NAEP

Some discussants noted that, although their schools had participated in NAEP, the results had been of little value because the relationship of NAEP to instructional programs has not yet been established. They emphasized the importance of alignment between curricula and assessment, pointing out that assessment results are of little use if based on material not covered in instructional programs. Ronald Costello described the role of assessment in school reform efforts:

> [A]ssessment is only one aspect of the three parts of what states and school districts are using testing to accomplish. The other two are standards and accountability, and we are only beginning to justify . . . the time taken away from instruction [to] serve those ends. Unless it can be connected to state and local curriculum and instructional practices, there will be little value [in continued participation] in NAEP. It doesn't matter how good the assessments . . . [are if] they can't be connected to standards and accountability in our states and school districts. For the market basket to have value at the state and local level, it must add value to what we do in schools to improve student learning. . . . [We must] be able to use the information in the change process.

Participants thought that the short form would provide relevant information for school systems considering changes. The released short form would permit educators and policymakers to see first hand the material included on the test. Their review of the released material would promote discussions about what is tested and how it compares with the skills and material covered by their own curriculum. The secure short form would yield data that could further these discussions. Educators could examine student data (even if it were aggregated to the school or district level) and evaluate performance in relation to their local practices. They could engage in discussions about their curricula, instructional practices, and sequencing of instructional material, and could contemplate changes that might be needed.

Stimulating Discussions with Teachers

As described earlier in this report (see Introduction and Chapter 4), members of the First in the World Consortium participated in TIMSS and used the results to learn how the consortium schools fared against world-class standards and to examine and revise curriculum, instructional strategies, and assessment practices. According to First in the World representatives, Paul Kimmelman and Dave Kroeze, a key component of the consortium's efforts was the establishment of learning networks that allowed teachers and administrators to participate in the reform discussions and improvement efforts (Hawkes et al., 1997).

The consortium established a research agenda covering four broad areas: student performance; curriculum and instruction; instructional practices; and teacher characteristics. An essential aspect of the research agenda was the involvement of teachers and administrators. Teachers and administrators reviewed their coverage of relevant content, amounts of instruction in specific areas, and the depth of understanding expected of students. They also studied teachers' attitudes and beliefs about instruction, the amount and type of homework assigned, and the extent to which teachers were using computers and calculators.

Teams of teachers from the participating schools examined students' performance on the topics covered on the TIMSS assessment. While they did not have access to the actual test items, they did have data on performance in specific topic and content areas. They used the results to evaluate their students' performance on the topics compared to students in other nations and considered when and how the topics were covered in their curricula. Kimmelman and Kroeze felt that discussions with and among teachers represented some of the most valuable outcomes of the consortium's participation in TIMSS.

Out-of-Cycle NAEP Testing

While the initial plan is to pilot test a short form for fourth grade mathematics, if the pilot is successful, NAGB may extend use of short forms to other grades and other subject areas. Because NAEP does not currently administer every subject to every grade every year, workshop participants suggested that short forms would help fill in the gaps; that is, they could be used to survey students in grades, subjects, and times that are not part of the regular NAEP schedule. At present, for example, the fourth grade mathematics assessment occurs every four years as part of state NAEP. If available, a short form in fourth grade mathematics could be given in the "off-years," thereby enabling compilation of yearly trend data. If short forms were produced for subjects not tested as part of state and local assessments, then states and districts could use the short forms to expand their assessment programs.

PROBLEMS WITH THE SHORT FORMS

Scoring: Faster, Easier, Better?

One advantage cited for the short forms was that they could be faster and less expensive to score than traditional NAEP assessments, providing score distributions without NAEP's usual complex statistical methodologies. Although it is not clear what NAGB's policy would eventually be with regard to how scores on the short form would be derived, workshop participants discussed scoring advantages—and disadvantages—at length. The text below highlights their comments.

NAEP uses a complex statistical process for scoring to compensate for the fact that no one student takes the full assessment. Since no one student responds to a sufficient number of items in a given content area to produce reliable estimates of performance, ability estimates are not computed for individuals. Instead, a conditioning process is used to generate the likely ability distributions for each individual based (or "conditioned") on background characteristics and responses to cognitive items. Five ability estimates (or "plausible values") are drawn from the distributions as estimates of the individual's proficiency level. These "draws" are aggregated over individuals to produce estimates of group-level distributions of performance.

There is a fundamental difference between NAEP's process and the process used by tests that are designed to produce individual scores: tests that produce individual scores do not condition. In fact, test users would most likely reject conditioning were it used to derive individual scores. As a result of conditioning, an individual's performance is adjusted according to the way others like him or her perform, others who share common characteristics such as gender, race, ethnicity, and socioeconomic status. This process is justifiable when the purpose of an assessment is to estimate group-level performance, but it is not typically used for tests that generate individual scores. In addition, test results based on conditioning are not comparable to unconditioned results.

One option for using the short form would make it available to states, districts, and schools to administer as they see fit. Scores on the short form would be generated for individuals then aggregated to provide group-level data in the percent correct metric. While this would circumvent NAEP's complex statistical procedures, it also means that short-form scores would not be conditioned. Hence, short-form results would not be comparable to the regular NAEP-scale results.

During their opening presentations, John Mazzeo and Robert Mislevy described methods that could be used to achieve NAEP-comparable results from the short form. These methods would use complex and lengthy statistical procedures. Given some of the complexities involved in producing scores that would be directly comparable to NAEP, several speakers questioned the extent to which it is critical to place the market-basket results and NAEP on exactly the same scale. In the words of one speaker, "Would it not be sufficient to provide results that are only *somewhat NAEP-like*?"

Reliability and Generalizability

The items selected for the short form are intended to represent NAEP's fourth grade mathematics frameworks. One might expect, therefore, the scores on the short form would be generalizable not only to the set of questions at hand but also to the content domain from which the items were drawn. At the workshop, Darrell Bock of the University of Chicago expressed concern about the reliability and generalizability of scores based on the short form. One of the two pilot test versions of the short form contains 31 items and the other 33. Bock estimated that the reliability of a professionally developed 31-item test would likely fall in the low .80 range, a value judged to be too low when tests are being used to make decisions about individuals. He stressed that the more germane concern is not about reliability, but about generalizability. Can 31 items adequately represent the content domain? He reminded participants that fourth grade mathematics crosses content strand with process category with item type. The result is a matrix with about 60 cells. While it is possible to represent these cells under the current matrix sampling approach for NAEP, how well, he asked, could a 30-some item test represent these cells?

In preparation for the workshop, Patricia Kenney, co-director of the NCTM NAEP Interpretive Reports Project, considered the feasibility of creating market-basket forms that matched the grade four NAEP mathematics assessment on the basis of content strand coverage, ability category, and item type. Based on material in John Mazzeo's paper, Kenney reported that the market-basket forms appear to represent the frameworks in terms of the content strand. But, she called attention to the fact that the grade four mathematics framework covers 56 topics and subtopics. Like Bock, Kinney questioned the extent to which 30 items would be able to represent the frameworks at the topic or subtopic level.

Kenney pointed out an additional potential problem. In NAEP, items can be administered at more than one grade level. For example, an algebra item might be given at grades four and eight to facilitate measuring growth between grade levels. Because NAEP results are not reported at the student level, students are not disadvantaged by including topics that they may not have studied. The problem with these "grade overlap" items, however, is that they might become misinterpreted as NAEP recommendations. For instance, suppose an algebra item appeared in the fourth grade mathematics form. Would this imply that schools should teach algebra to fourth graders?

Kenney's most overarching concern was with "retrofitting," that is, manipulating the features of an existing system to adjust for new purposes and uses. In the case of mathematics, neither the existing NAEP framework nor the item pool was developed with market-basket reporting in mind. Therefore, Kenney wondered if the existing materials would support such procedures. She reminded participants that the mathematics test development committee "was not able to assemble a collection of items that they felt were all exemplary of the framework and met all statistical, content, and process, and format specifications" (Mazzeo, 2000:11). Kenney repeated Mazzeo's concerns about retrofitting saying, "In my own view, it is often more difficult and risky to attempt to retrofit a reporting and data collection system to an existing assessment that was not designed for such purposes than it is to build such an assessment system from scratch" (Mazzeo, 2000:27).

Level of Disaggregation

Workshop participants discussed the utility of disaggregated data, and questions arose as to the level of disaggregation that would be permitted. Carroll Thomas, superintendent of schools for Beaumont, Texas, emphasized the importance of having data for various population groups (e.g., data separated by racial/ethnic, gender, or other background characteristics). He pointed out that the conclusions one draws based on seeing total-group results could be very different from the conclusions one might draw based on results for various population groups. Thomas said he believes that decisions about changes in educational practices should be based on examining disaggregations of group-level data. Some participants asked if results would be reported by background characteristics. Others inquired whether results would be made available by school or only by district. Generally, participants believed that disaggregation will enhance the utility of results.

Other participants spoke about having individual results. NAGB's approved policy with regard to the short form explicitly prohibits reporting individual scores. While individual results would be generated initially, they would need to be aggregated for reporting purposes. The prohibition against producing individual results based on the short form stimulated considerable discussion. The short form could be administered to all children in a specific grade in a manner closely resembling other testing in schools—testing that results in individual score reports. How would one account for not having individual scores? Assessment directors and policymakers at the workshop maintained that the situation would indeed be difficult to explain to interested parties, particularly parents. Many felt the temptation to generate individual scores would be great, and difficult to resist, despite NAGB's prohibition. Further, in a scenario in which the short form was embedded into an existing assessment, participants wondered if the items would contribute to the individual scores generated by the existing assessment.

Under the current NAGB plan, two short forms would be produced, one for public release and the other kept secure and retained for use by states and districts. While school administrators might be able to control the generation of individual results from the secure form, the released form would be publicly available for any, and all, uses. In commenting on this, Richard Colvin of the *Los Angeles Times* described probable uses of the released test and suggested ways to handle derivation of individual scores:

> No matter what caveats you offer, people will take the test and will calculate a percent correct score for their performance. I can guarantee that the *[LA] Times* would post such a test on its web site. . . . Americans are used to taking tests in magazines and comparing their performance to a scale. Despite your caveats, schools will have their students take it, and they will calculate a percent correct score. You won't be able to stop that so you need to figure out what to say about it. It would be better if there were a way to have a conversion scale of some sort. Another idea would be to set up a web site of your own where people could take the tests on-line. Then, perhaps there'd be a way to actually produce a score, based on which questions were answered correctly.

Embedding the Short Form in State and Local Assessments

The most common projected use of the short form cited by policymakers and directors of assessment was embedding it in state and local assessments. This potential use prompted considerable discussion. Scott Trimble of Kentucky's Department of Education pointed out that a given state's curriculum might not be completely congruent with the NAEP frameworks. It might be the case, for instance, that the state's curriculum includes areas not tested by NAEP, in which case the state assessment would have to cover the areas not covered on the short form. Or it might be that the NAEP form tests areas not covered by the curriculum, in which case, students would not have been taught the skills and knowledge being tested. Testing students on material they have not been taught results in less-than-useful measures of achievement.

Some workshop discussants raised the question of how students' motivation to do well might factor into performance on the short form. State and local assessments tend to be high-stakes exams that carry consequences for those who do not perform well; thus, motivation to do well is high. At present, NAEP is not a high-stakes test. Administering a NAEP short form as part of a high-stakes assessment program would change the context in important and relevant ways. While there are still unanswered questions about the effects of motivation on assessment results, the introduction of higher stakes could render results from the short form incomparable to national or state NAEP results and call into question the types of inferences that might be made.

Testing burden was also a concern to participants. Many judged that it was unlikely they could introduce additional assessments into their states and districts nor could they sacrifice more instructional time for testing. Thus, NAEP items, in essence, would need to do "double duty." That is, to prevent test administration from taking any more time than it already does, NAEP items would need to count toward the score on the short form and also replace items currently on state and local tests that measure similar skills and content.

Such uses of the NAEP items bring to the forefront issues about linking state and local assessments to NAEP. Several discussants referenced reports from two earlier NRC committees, the Committee on the Equivalency and Linkage of Educational Tests and the Committee on Embedding Common Test Items in State and District Assessments (National Research Council, 1999d; 1999a). Both committees studied issues associated with linking state and local assessments to NAEP. And, after in-depth exploration, both committees concluded that many problems surround attempts to link assessments not initially designed for the purposes of linking.

8
Summing Up: Issues to Consider and Address

One of the objectives for convening the market-basket workshop was to hear individuals representing a variety of perspectives respond to the plans for the NAEP market basket. The earlier chapters of this report summarize the remarks, organizing speakers' comments into four broad categories: (1) the parallels between the NAEP market basket and the CPI; (2) the use of a representative set of items as a means for facilitating public understanding of the material tested by NAEP; (3) the use of percent-correct scores as the summary indicator to communicate performance on the set of released items; and (4) the use of a short form to enable comparisons of performance on state and local tests with performance on NAEP. In listening to the speakers' reactions and the discussion among workshop participants, the committee identified a number of themes that emerged from the interactions at the workshop. While no attempt was made to establish consensus on these themes, they are discussed in this chapter to further assist NAEP's sponsors in their decision making regarding implementation of the market basket.

LIMITATIONS OF THE CPI METAPHOR

During his market-basket workshop presentation, Richard Colvin of the *Los Angeles Times* questioned the analogies being drawn between the CPI and the NAEP market basket. He maintained that metaphors can be a very effective means for conveying the meaning of complex phenomena.

The image of a shopper filling up a shopping cart works very well for the CPI. Since the CPI has to do with how much products and services cost, the market basket is an appropriate metaphor. Colvin wondered, however, if it was the right metaphor to associate with student achievement:

> NAEP already holds claim to the best metaphor—"The Nation's Report Card." What would the report based on a representative set of items be called? This is not a trivial issue. The name must relate to something the public already understands. And, it must also be seen as complementing, not conflicting with, the NAEP as a whole. We in the press are going to call it something. We might call it a "Quiz," a "Final," or a sample of "What Students Need to Know." You need to choose what you think it ought to be called. And, if *you* don't, *we* will.

Colvin concluded by suggesting that a reference point is needed for the body of knowledge and skills that is to be part of the market basket. "If the market basket is analogous to what is covered by the entire NAEP, then what does NAEP represent?" he asked.

Although the concept of a market basket of goods and services is readily grasped, the construction and measurement of the CPI market basket is clearly a very subtle and complex undertaking. In addition, the current conception of the NAEP market basket differs from the CPI operation in several potentially important ways. First, the CPI market basket was conceived as a purely descriptive measure and is built using extensive data on actual consumer purchases. In contrast, the NAEP market basket would not passively measure what is actually being learned by school students. Rather, construction of the NAEP market basket would require normative judgments about the content of the domain and selection of representative items.

Second, regional and area differences in purchasing habits are reflected in the CPI local area indexes and limit comparability across geographical areas. For NAEP, the same collection of items will be used to summarize performance across geographical areas. In fact, this is one of the stated goals for the market basket, to facilitate reporting of regional-level results and making comparisons to national results. However, for the NAEP market basket, there will be no attempt to reflect regional differences in curriculum. Thus, students may be tested on concepts and skills that have not been covered by their instructional programs.

Finally, production of the CPI involves the development and execution of sample surveys designed specifically for pricing the CPI market basket. Computation of the CPI is not accomplished by simply embedding market-basket questions in an existing consumer survey. The CPI experience suggests that subtle and difficult measurement issues may await efforts to incorporate the market-basket concept into the existing structure of NAEP.

RELEASING A LARGE REPRESENTATIVE SET OF ITEMS

Many workshop participants commented on the utility of public release of a large representative set of NAEP items. They thought that such a release could potentially impact education reform by allowing teachers, administrators, curriculum specialists, and assessment directors to use the items in discussions about their instructional practices, curricular changes, and state and local assessments. Representatives from the First in the World Consortium offered examples of the ways in which released material might be used to further education reform efforts.

Discussants also suggested that such a release would be useful for increasing public awareness about the content and skills tested on NAEP. While this is certainly a well-intended objective, consideration should be given to the extent to which the public would take advantage of a large release of items and the inferences they might make. Would parents and others be willing to spend time reviewing large numbers of items? What would they think about the material they were seeing? Would NAEP's sponsors offer guidance to help them understand the content and skills the items are intended to assess? Simply placing a large number of items in the hands of the public would not necessarily enhance understanding. It will be important to consider the mechanisms that will be used to communicate with the public about the content and skills covered by the items. It may be enlightening to consider other testing programs' experiences with disclosing test forms and providing practice tests.

PERCENT CORRECT: NOT AS SIMPLE AS IT SOUNDS

A clear message from the workshop discussants was the deceptive complexity of the percent correct metric. One factor contributing to its complexity is the denominator of the percent-correct ratio; that is, percent correct of what? Total questions on the test? Total points on the test? Total content in the domain?

A second concern voiced by participants was the meaning attached to the percent correct score. Speakers cited a disconnect between the public perception of what constitutes a passing score and the actual percent correct scores that would be associated with the basic, proficient, and advanced achievement levels. They argued that the public is accustomed to seeing a letter grade attached to percent correct, and the temptation for the media and the general public to translate percent-correct scores to grades would be overwhelming.

A third issue raised was the comparability of percent correct scores with NAEP scores. NAEP currently reports results as scaled scores derived from IRT-based latent trait estimates. While the statistical machinery exists to transform the latent trait scale to a percent correct scale, the procedures are very complex and time consuming. Would the move toward percent correct scores be worth it, given the difficult procedures involved and that it might not lead to the desired improvements in understanding of NAEP results?

LINKING SHORT FORM RESULTS TO NAEP AND TO STATE AND LOCAL ASSESSMENTS

The most common use cited for the short form was embedding it into state and local assessments, thus providing states and localities with a "link" to NAEP. A previous NRC committee studied this subject in depth by examining several scenarios for embedding, including the embedding of representative blocks of NAEP material (National Research Council, 1999a). The committee maintained that while using representative blocks of material would help increase the comparability of scores across states, many issues would remain unresolved. They identified a number of factors that would bear on the comparability of scores, including NAEP's use of conditioning to estimate performance, likely misalignment of local curricula with NAEP, the contextual circumstances of testing within a given state or district, students' and administrators' motivation levels, administrative conditions, time of testing, and differing criteria for excluding students from participation (e.g., disabilities or limited English proficiency). These factors led the NRC's Committee on Embedding Common Test Items in State and District Assessments to conclude that:

> Embedding part of a national assessment in state assessments will not provide valid, reliable, and comparable national scores for individual students as long as there are: (1) substantial differences in content, format, or administration between the embedded material and the national test that it represents: or (2) substantial differences in context or administration between the state and

> national testing programs that change the ways in which students respond to the embedded items (National Research Council, 1999a:3).

This finding closely parallels an earlier conclusion reached by another NRC committee, the Committee on Equivalency and Linkage of Education Tests, which stated:

> Under limited conditions it may be possible to calculate a linkage between two tests, but multiple factors affect the validity of inferences drawn from the linked scores. These factors include the content, format, and margins of error of the tests; the intended and actual uses of the tests; and the consequences attached to the results of the tests. When tests differ on any of these factors, some limited interpretations of the linked results may be defensible while others would not be (National Research Council, 1999b:5).

As thinking about the design and intended uses of the short form proceeds, it is important to keep in mind the findings from these two committees.

OTHER ISSUES TO CONSIDER AND RESOLVE

Workshop participants brought up a number of other issues related to the development of the NAEP market basket. These issues bear on practical matters related to developing the market basket as well as unintended consequences that may be associated with its implementation.

Self-Elimination

One of the stated goals for NAEP's short forms is to make assessments available in subjects and grades not assessed every year. However, it is possible that NAEP could end up being a victim of its own success. If plans for the short form were successful, states and districts would have a test to administer in NAEP off-cycle years and could have easily derived scores comparable to the NAEP scale. Why, then, would they need to participate in NAEP? If they could do this in off-cycle years, why not do it every year?

Costs

NAEP has invested a considerable amount of time and money in the development of two short forms. Future work will be needed to score the short forms, to devise a mechanism for comparing percent correct scores with main NAEP scores, and to develop reporting procedures. If the short forms proceed to the stage of operational use, continued development of additional forms will be needed. But to what extent will this process result in more useful, more understandable results? To what extent will the market basket produce the desired outcomes? At a time when only limited funding is being made available for educational purposes, is this the best use of funds? The costs and benefits of the market basket should be carefully considered.

Retrofitting the Design

Originally, NAEP was developed as a survey of what American school children know and can do. The frameworks cover broad content areas. The content areas are combined with other item characteristics (such as item type, item difficulty, and cognitive process) to form a test blueprint matrix. For mathematics, this matrix has some 60 cells. Currently, no one student takes sufficient items to represent the matrix fully. Instead, a matrix sampling procedure is used to assign items to blocks, blocks to forms, and forms to students. A single student takes three blocks of items.

This sort of test assembly is very different from that typically used in tests developed for educational purposes, where a test form that has the proper mix of content and item type to represent the test specifications is the end result, and a given student takes the entire test. Construction of the short form would require this other type of development. NAEP's current frameworks and existing item pools were not created with this type of development in mind. Limitations on the amount of time schools have for testing places restrictions on the number of items that can be administered. And NAEP's current frameworks and existing item pools, which are very broad, may not be able to be represented with the type of test that can be administered in a 45-minute session. In fact, as noted by John Mazzeo, test specifications had to be generated in order to assemble the short forms, and these specifications were based on an examination of the characteristics of the item pool and what it would support.

One key issue to emerge from the workshop is the need for explicit consideration of the ramifications of building a new system by manipulating the features of a pre-existing system. During their presentations, both John Mazzeo and Patricia Kenney expressed concern about the difficulties associated with trying to retrofit a pre-existing reporting and data collection system to new purposes and needs, particularly when the pre-existing system was not originally designed for such purposes and needs. It is important to keep these cautionary words in mind.

Changing NAEP's Purpose

During his presentation at the workshop, David Thissen of the University of North Carolina-Chapel Hill used Holland's characterization of "testing as measurement" versus "testing as a contest" to describe different purposes for testing. When thinking of testing as measurement, the goal is to make the appropriate inferences; that is, to measure performance as *accurately* as possible. When thinking of testing as a contest, the goal is get the *highest scores* possible. According to Thissen, NAEP's current procedures treat testing as measurement, seeking to obtain the most accurate estimates of student performance. Implementation of procedures that involve the short form will move NAEP into the category of testing as a contest.

Testing as a contest is high-stakes testing. NAEP traditionally has been a low-stakes test, since decisions about schools, teachers, and individuals have not been based on test results. As reporting moves to smaller units, the stakes increase, as does pressure and motivation to do well. Motivation to do well will, undoubtedly, affect performance. Thus, it is not clear how comparable data from national NAEP (taken under low-stakes conditions) will be with local results based on the short form (taken under high-stakes conditions). This undermines one of the main goals articulated for the short form—to facilitate comparisons with national benchmarks. Again, NAEP's sponsors should consider these potential consequences as policy and decision making about the market basket proceeds.

References

Beaton, A.E. and R. Zwick

1992 Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*. 17(2):95-109.

Bock, R.D.

1993 Domain-Referenced Reporting in Large-Scale Educational Assessments. Paper commissioned by the National Academy of Education for the Capstone Report of the NAE Technical Review Panel on State/NAEP Assessment.

Bock, R.D., D. Thissen, and M.F. Zimowski

1997 IRT Estimation of Domain Scores. *Journal of Educational Measurement* 34:197-211.

Forsyth, F., R. Hambleton, R. Linn, R. Mislevy, and W. Yen

1996 Design and Feasibility Team: Report to the National Assessment Governing Board, Report of July 1.

Hawkes, M., P. Kimmelman, and D. Kroeze

1997 Becoming "First in the World" in Math and Science. *Phi Delta Kappan*, September.

Johnson, E., S. Lazer, and C. O'Sullivan

1997 NAEP reconfigured: An integrated redesign of the National Assessment of Educational Progress. Washington, DC: National Center for Education Statistics.

Lord, Frederick M.

1980 *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mazzeo, J.

2000 NAEP's Year-2000 Market-Basket Study: What Do We Expect to Learn? Paper prepared for National Research Council Workshop on Market-Basket Reporting.

National Assessment Governing Board

1996 Redesigning the National Assessment of Educational Progress. Policy Statement.

1997 Resolution on Market Basket Reporting, August 2.

1999a Policy Guidance on the NAEP Short Form: Summary of Principles Related to Measurement Issues.

1999b The National Assessment of Educational Progress: Design 2000-2010.

National Education Goals Panel

1994 The National Education Goals Report: Building a Nation of Leaders. Washington, DC: Government Printing Office.

National Research Council

1999a *Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests.* Committee on Embedding Common Test Items in State and District Assessments. D.M. Koretz, M.W. Bertenthal, and B.F. Green, eds. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

1999b *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress.* Committee on the Evaluation of National and State Assessments of Progress. J.W. Pellegrino, L.R. Jones, and K.M. Mitchell, eds. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

1999c *Reporting District-Level NAEP Data: Summary of a Workshop.* Committee on NAEP Reporting Practices: Investigating District-Level and Market-Basket Reporting. P.J. DeVito and J.A. Koenig, eds. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

1999d *Uncommon Measures: Equivalence and Linkage Among Educational Tests.* Committee on the Equivalency and Linkage of Educational Tests. M.J. Feuer, P.W. Holland, B.F. Green, M.W. Bertenthal, and F.C. Hemphill, eds. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

APPENDIX A
Workshop Agenda and Participants

AGENDA

Workshop on Market-Based Reporting

February 7-8, 2000

Monday, February 7

Open Session

10:00-10:15	**Opening Remarks** Pat DeVito, Chair
10:15-12:15	**Goals, Purposes, Uses, Plans, and Options for the NAEP Market Basket** Facilitators: Audrey Qualls and Douglas Herrmann

Topics:
(1) What are the primary objectives for market-basket administration, market-basket reporting, and the short form?
(2) Who are the proposed users for market-basket materials and the short form?
(3) What types of inferences are expected to be supported by the short form and market-basket results?
(4) What is the status of research and development work on the market-basket and short form?
(5) What are the Board's plans for pursuing work on the market basket/short form—with regard to the 2000 assessment and subsequent work?

Speakers:
Roy Truby, National Assessment Governing Board
Andrew Kolstad, National Center of Education Statistics
Robert Mislevy, Educational Testing Service
John Mazzeo, Educational Testing Service

Questions and Answers

12:15-12:45 *Lunch in the meeting room*

12:45-2:45 **The Policy Perspective**
Moderators/Discussion Facilitators: Linda Bryant and Lou Fabrizio

Topics:
(1) What information needs might be served by market-basket reporting? Who would use the results? How would they be used?
(2) What information needs might be served by the market basket/short form? Who might use it and how?
(3) What if district-level results were available or could be generated from the short form? Who might use the results and how?
(4) What are the implications of market-basket reporting for other national, state, and local assessment programs?

Speakers:
Wayne Martin, Council of Chief State School Officers
Marilyn McConachie, Vice Chair, Illinois State Board of Education
Carrol Thomas, Superintendent of Schools, Beaumont, TX
Marlene Hartzman, Office of Accountability, Montgomery County, MD

Discussion and Synthesis of Ideas

2:45-3:30 **The First in the World Experience with TIMSS**
Moderators: Melody Carswell and Maryellen Donahue

Speakers:
Paul Kimmelman, Superintendent, West Northfield School District 31, IL
David Kroeze, Superintendent, Northbrook School District 27, IL

Tuesday, February 8

Open Session

8:00-8:30 *Continental breakfast*

8:30-10:00 **The Perspective of Users and Practitioners**
Moderators/Discussion Facilitators: LeAnn Gamache and Maryellen Donahue

Topics:
(1) What information needs might be served by market-basket reporting and the market basket/short form? Who would use the results? How would they be used?
(2) Will the content and skill coverage be adequate?
(3) What if district-level results were available or could be generated from the short form? Who might use the results and how?
(4) What are the implications of the market basket for state and local assessment programs?
(5) What are the implications of the market basket for state and local curriculum and instructional practices?

Speakers:
Ronald Costello, Assistant Superintendent, Noblesville, IN
Patricia Kenney, University of Pittsburgh
Joe O'Reilly, Unified School District, Mesa, AZ
Scott Trimble, Office of Assessment, Kentucky Dept. of Education

Discussion and Synthesis of Ideas

10:00-10:30 **A Reporter's Perspective**
Facilitators: Douglas Herrmann and LeAnn Gamache

Topics:
(1) What information needs might be served by market-basket reporting and the market basket/short form? Who would use the results? How would they be used?
(2) What information does the public want to know about student achievement? To what extent will the proposed market basket fulfill these needs?
(3) How might market-basket results be interpreted by the press and the public? What cautions and guidance should be considered?

Speaker:
Richard Colvin, *Los Angeles Times*

Discussion and Synthesis of Ideas

10:30-10:45 *Break*

10:45-12:30 **The Measurement Perspective**
Moderators/Discussion Facilitators: Mark Reckase and Duane Steffey

Topics:
(1) What issues should be considered in trying to implement the use of market-basket forms and/or market-basket reporting?

(2) To what extent will the current plans for the market basket (and/or short form) produce results that will fulfill the intended purposes?
(3) What types of inferences would be supported by market-basket results?
(4) Will market-basket reporting pose any threats to the validity of inferences from national and state NAEP?
(5) Specific comments regarding procedures for assembling the market basket, deriving scores, reporting performance information, and making comparisons with national/state NAEP results.

Speakers:
Darrell Bock, University of Chicago
Don McLaughlin, American Institutes for Research
David Thissen, UNC Chapel Hill

Discussion and Synthesis of Ideas

12:30-1:30 *Lunch in meeting room*

1:30-2:00 **The Milwaukee Experience with District-Level Results**
Facilitators: Linda Bryant and Lou Fabrizio

Speaker:
Paul Cieslak, Milwaukee Public Schools

2:00-2:45 **Comparisons with the Consumer Price Index**
Facilitators: Duane Steffey

Speaker:
Kenneth Stewart, Bureau of Labor Statistics

2:45 Workshop Adjourns

PARTICIPANTS

R. Darrell Bock, University of Chicago
Mary Lyn Bourque, National Assessment Governing Board
Linda Bryant, Westwood Elementary School, Pittsburg
Peggy Carr, National Center for Educational Statistics
Paul Cieslak, Milwaukee Public Schools
Richard Lee Colvin, *Los Angeles Times*
Ronald Costello, Noblesville Schools, IN
Patricia Dabbs, National Center for Educational Statistics
Pasquale DeVito, Rhode Island Department of Education
Maryellen Donahue, Boston Public Schools
Lou Fabrizio, North Carolina Department of Public Instruction
LeAnn Gamache, Littleton Public Schools, CO
Arnold Goldstein, National Center for Educational Statistics
Steve Gorman, National Center for Educational Statistics
Marlene Hartzman, Montgomery County Public Schools
Douglas Herrmann, Indiana State University
Carol Johnson, National Center for Educational Statistics
Patricia Kenney, University of Pittsburgh
Paul L. Kimmelman, West Northfield School District, IL
Kaeli Knowles, National Research Council
Judith Koenig, National Research Council
Andrew Kolstad, National Center for Educational Statistics
David Kroeze, Northbrook School District 27, IL
Wayne Martin, Council of Chief State School Officers
John Mazzeo, Educational Testing Service
Marilyn McConachie, Illinois State Board of Education
Donald McLaughlin, American Institutes for Research
Robert Mislevy, Educational Testing Service
Karen Mitchell, National Research Council
Joseph O'Reilly, Mesa Unified School District, AZ
Audrey Qualls, Iowa Testing Program
Mark Reckase, Michigan State University
Alex Sedlacek, National Center for Educational Statistics
Larry Snowhite, MCA Enterprises, Inc.
Holly Spurlock, National Center for Educational Statistics
Duane Steffey, San Diego State University
Ken Stewart, Bureau of Labor Statistics

Alan Thiemann, Association of Test Publishers
David Thissen, University of North Carolina, Chapel Hill
Carrol Thomas, Beaumont Independent School District, TX
C. Scott Trimble, Kentucky Department of Education
Roy Truby, National Assessment Governing Board
Lauress Wise, Human Resources Research Organization