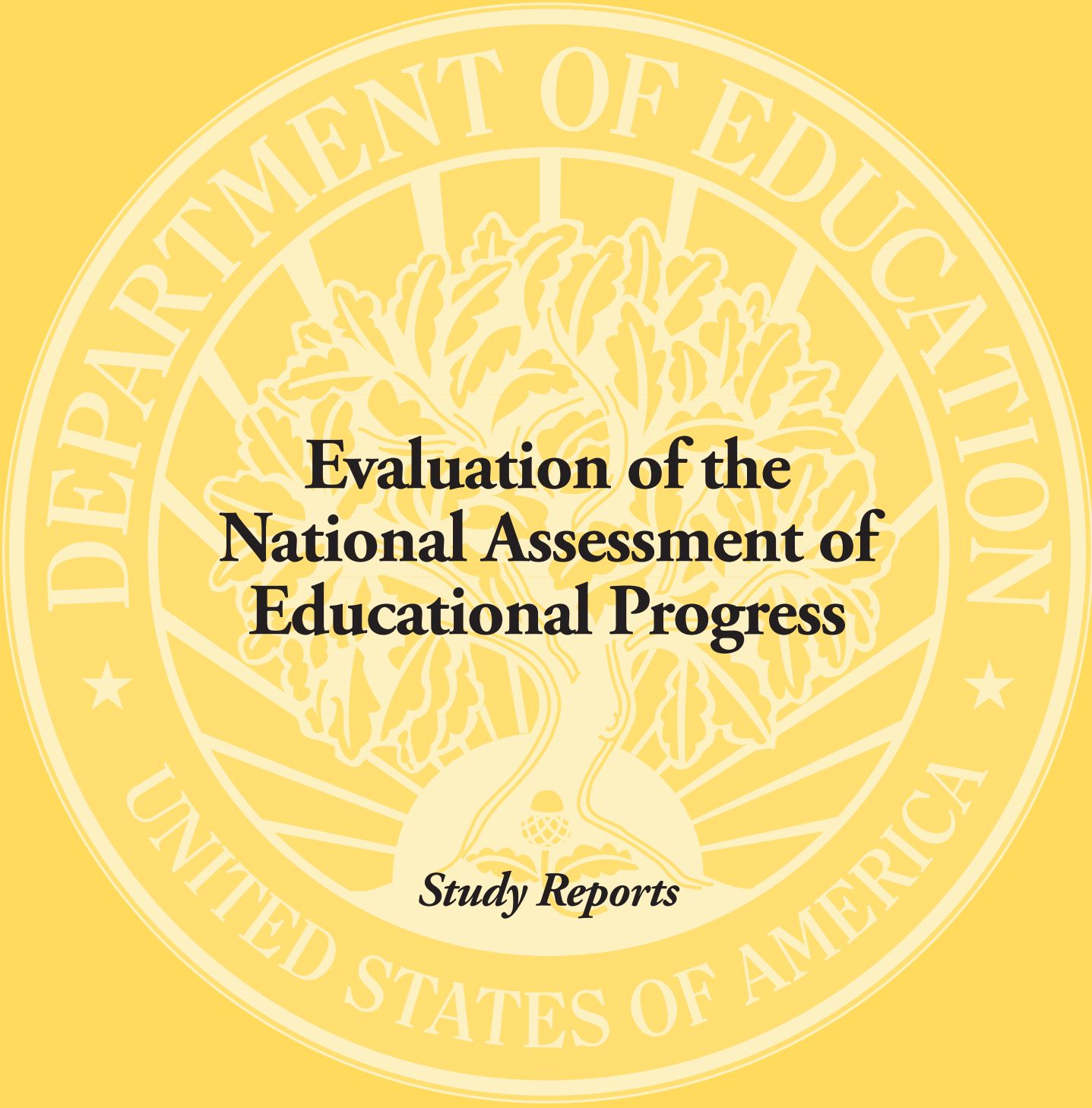


U.S. DEPARTMENT OF EDUCATION



**Evaluation of the
National Assessment of
Educational Progress**

Study Reports

**Evaluation of the
National Assessment of Educational Progress**

Study Reports

Prepared by:

Chad W. Buckendahl
Susan L. Davis
Barbara S. Plake
Buros Institute for Assessment Consultation and Outreach
Buros Center for Testing
University of Nebraska–Lincoln

and

Stephen G. Sireci
Ronald K. Hambleton
April L. Zenisky
Craig S. Wells
Center for Educational Assessment
University of Massachusetts–Amherst

2009

This congressionally mandated report was done under Contract Number ED04CO0159 with the Buros Institute for Assessment Consultation and Outreach, a Division of the Oscar and Luella Buros Center for Testing, University of Nebraska, Lincoln, and the Center for Educational Assessment, University of Massachusetts, Amherst. Jay Noell served as the contracting officer's representative. The views expressed herein do not necessarily represent the positions or policies of the Department of Education. No official endorsement by the U.S. Department of Education of any product, commodity, service or enterprise mentioned in this publication is intended or should be inferred.

U.S. Department of Education

Arne Duncan
Secretary

Office of Planning, Evaluation and Policy Development

Carmel Martin
Assistant Secretary

Policy and Program Studies Service

Alan Ginsburg
Director

September 2009

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the citation should be: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, *Evaluation of the National Assessment of Educational Progress, Study Reports*, Washington, D.C., 2009.

To order copies of this report:

Write to: ED Pubs, Education Publishing Center, U.S. Department of Education, P.O. Box 1398, Jessup, MD 20794-1398.

Or **fax** your request to 301-470-1244.

Or **e-mail** your request to: edpubs@inet.ed.gov.

Or **call** in your request toll-free: 1-877-433-7827 (1-877-4-ED-PUBS). If 877 service is not yet available in your area, call 1-800-872-5327 (1-800-USA-LEARN). Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY), should call 1-877-576-7734.

Or **order online** at: www.edpubs.ed.gov.

This report is also available on the Department's Web site at: www.ed.gov/about/offices/list/opepd/ppss/index.html.

On request, this publication is also available in alternate forms, such as Braille, large print, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-0852 or 202-260-0818.

Contents

Preface.....	v
Acknowledgments.....	vii
Foreword by the Technical Work Group.....	ix
Chapter 1: Audit Study Report.....	1-i
Chapter 2: Evaluation of the Standard Setting on the 2005 Grade 12 National Assessment of Educational Progress Mathematics Test.....	2-i
Chapter 3: How Do Other Countries Measure Up to the Mathematics Achievement Levels on the National Assessment of Educational Progress?.....	3-i
Chapter 4: A Study of the Utility of the National Assessment of Educational Progress...4-i	
Chapter 5: Evaluating Score Equity Across Selected States for the 2005 Grade 8 NAEP Math and Reading Assessments.....	5-i
Chapter 6: Methods for Evaluating the Alignment Between State Curriculum Frameworks and State Assessments: A Literature Review.....	6-i

This page left intentionally blank

Preface

The *Evaluation of the National Assessment of Educational Progress: Study Reports* describes the special studies that comprised the design of the evaluation. In the Final Report, we presented a practical discussion of the evaluation studies to its primary, intended audience, namely policymakers. On this accompanying CD, readers will find additional evidence to support our findings and recommendations in the six study reports. The study reports represent summaries of the data collection, analysis, and findings of the different lines of inquiry that comprised the evaluation design.

In chapter one on this CD, we describe the procedures and results of an audit of the NAEP assessment lifecycle that served as an organizing framework for the evaluation. The purpose of this study was to evaluate the breadth of NAEP's test development, administration, scoring, reporting, and maintenance processes by applying the professionally adopted standards of practice (i.e. *Standards for Educational and Psychological Testing*; AERA, APA, and NCME, 1999). Elements of the audit were designed to respond to each of the four congressional questions.

In chapters two and three on this CD, we describe the two studies that were designed to evaluate an area of congressional interest with respect to NAEP's achievement levels. In the first of these two studies, we evaluated the application of a new methodology for setting achievement levels on the 2005 Grade 12 NAEP Mathematics assessment. In the second study, we evaluated evidence from two international assessments to examine the utility of these external measures of achievement in the context of interpreting NAEP's achievement levels.

In chapter four on this CD, we describe a series of studies that evaluated how stakeholders used and interpreted NAEP results and achievement levels presented in printed and Web-based formats. This area of evaluation is of particular interest given NAEP's increased visibility. Data collection for these evaluation activities included interviews, focus groups, analyses of Web usage data, and studies of how consumers interpreted results reported in print and from the NAEP Web site.

As an important issue of fairness, in the study described in chapter five on this CD, we investigated the consistency across methods for calculating NAEP scale scores across states. Specifically, we evaluated whether the results for selected states would differ if NAEP assessments were statistically placed on the same score scale (i.e., equated) across time using only data from the state, as opposed to data from the entire nation, as is standard operating procedure. Because there are multiple steps involved the process of estimating scale scores, we evaluated whether any of those steps might affect the results for particular states. We also compared item statistics and achievement level results across national and state-specific replications.

In the final study report described in chapter six on this CD, we reviewed alignment methodologies currently used by state assessment programs. Alignment generally refers to the degree of overlap among content standards, curriculum, instruction, and assessments. As a primary source of validity evidence in contemporary educational assessment programs, alignment studies also represent a critical policy consideration when interpreting and using scores. This review provides some context for policymakers as they consider potential uses and interpretations of NAEP results.

This page left intentionally blank

Acknowledgments

This final report of the evaluation of the National Assessment of Educational Progress (NAEP) benefited from the contributions of many people outside and within the U.S. Department of Education. The evaluation team would like to extend its appreciation to these individuals and acknowledge those whose assistance made this final report possible.

First, the evaluation was conducted under the guidance of a Technical Work Group (TWG), whose members' names and affiliations appear in the Foreword. The TWG was co-chaired by Suzanne Lane of the University of Pittsburgh and Bruno Zumbo of the University of British Columbia who both served as liaisons between the evaluators and the full group. Their contributions were invaluable as they provided advice and input on the design, evaluation activities, and reports.

Second, we wish to thank the individuals we worked with in the Department's Policy and Program Studies Service (PPSS). Specifically, Jay Noell provided continuous support and advice throughout the evaluation on how to better characterize our findings in reports that would be meaningful for policymakers. We also want to thank Alan Ginsburg, David Goodwin, and Maggie Cahalan for their valuable contributions during the evaluation.

Third, we appreciate the efforts of the staff of the National Assessment Governing Board (NAGB), particularly Charles Smith, Susan Loomis, Mary Crovo, and Sharif Shakrani, now affiliated with Michigan State University, to provide documentation, clarification, and feedback, on the components of NAEP for which they are responsible. Likewise, we extend our thanks to the Assessment Division of the National Center for Education Statistics (NCES) and specifically to Peggy Carr, Andrew Malizio, Janis Brown, and Andy Kolstad for their assistance during the evaluation. Although many of the documents we requested remained in the internal review process for the duration of the evaluation, draft reports and documents were provided, when possible.

Fourth, we want to thank the organizations and individuals that serve as contractors for the components of NAEP that were included in the evaluation. These organizations in alphabetical order were, ACT, Inc., American Institutes for Research (AIR), Educational Statistics Services Institute (ESSI), Educational Testing Service (ETS), Government Micro Resources, Inc., Hager Sharp, Human Resources Research Organization (HumRRO), the NAEP State Coordinators, Pearson Educational Measurement (PEM), and Westat, Inc.

Finally, because the foundation for this report is based on multiple studies and data collection efforts that comprised the evaluation, there were a number of people who played key roles in the project. We appreciate the efforts of these individuals in contributing to the success of the evaluation. Specifically, we want to thank: Jim Impara, Brett Foley, Teresa Eckhout, Elaine Rodeck, Anja Römhild, Rebecca Norman, Theresa Glanz, Janice Nelsen, and Kurt Geisinger of the Buros Center for Testing at the University of Nebraska, Lincoln; Lisa Keller, Drey Martone, Kelly Smiaroski, Jeffrey Hauger, Su Baldwin, Kyung T. (Chris) Han, Stephen Jirka, Ana Karatonis, Robert Keller, Jill Delton, Christine Lewis, Polly Parker, and Zachary Smith of the Center for

Educational Assessment at the University of Massachusetts, Amherst; Deborah Bandalos of the University of Georgia; Edward Wiley of the University of Colorado; and Barbara Badgett of the University of Nevada, Las Vegas. A special thanks also to Cathy Cohen of C.J. Cohen Associates and Mickey Boisvert of MBDesign for their assistance in providing technical and style editing services through multiple drafts of the report.

Although we have received feedback from the U.S. Department of Education, NAGB, NCES, and the TWG during the evaluation, the judgments expressed in this report are those of the authors. This fulfills the spirit and requirement of the law that this evaluation be independent. The views expressed in this report do not necessarily reflect those of the University of Nebraska, Lincoln, or the University of Massachusetts, Amherst.

Chad W. Buckendahl*
Susan L Davis*
Barbara S. Plake
Stephen G. Sireci
Ronald K. Hambleton
April L. Zenisky
Craig S. Wells

* After October 2007, work on this project by Buckendahl and Davis occurred as employees of Alpine Testing Solutions.

Foreword by the Technical Work Group

The Changing Context of Large-Scale Assessments

The purposes, uses, and consequences of large-scale assessments have changed fundamentally over the past few decades. While the consequences of large-scale assessment results have steadily mounted, the attention paid to making the purposes of and uses of such assessments explicit has not always kept pace. Yet the meanings given to assessment results and the uses to which the results are put are valid only to the degree that supporting evidence exists.

However, if the proposed interpretations and uses of the assessment results are not made explicit during the design and ongoing implementation phases, it lessens the likelihood that appropriate validity evidence will be collected—evidence essential both for supporting the interpretations and uses of the assessment results and for evaluating and monitoring any unintended uses and consequences. Careful delineation of the proposed interpretations and uses of an assessment also draws attention to issues of fairness and equity.

These issues are of particular importance because of the increased use of large-scale assessments to examine and monitor the performance of aggregated subgroups, defined by demographic conditions such as geographic location, race, and ethnicity. When interpretations and uses are clarified and made explicit, fairness and equity issues can be addressed, intended consequences can be evaluated, and unintended, potentially negative consequences can be minimized. It is difficult therefore to overstate the importance of assessment programs being clear and specific about intended interpretations and uses.

What is true for large-scale assessment programs in general is especially true for the National Assessment of Educational Progress (NAEP), given its emerging role as a policy tool to interpret state assessment and accountability systems. While it is the case that there have been numerous validity studies to support many of the interpretations and uses of NAEP results, NAEP has not had the benefit of a comprehensive framework to guide the *systematic* accumulation of evidence in order to substantiate the ways in which its assessment results may be reasonably interpreted and applied. As new uses for NAEP continue to emerge, delineating a validity framework—an organized plan for collecting evidence to support intended uses and interpretations of test scores—must become a priority. The emphasis here is on using the validity framework as an organizing tool, not simply a call for research.

Historical View of NAEP and Its Evolution

The ways in which NAEP results are reported and used have evolved over the nearly 40 year history of the NAEP assessment program. What began as a relatively straightforward, low visibility measure of student achievement at the national level has been transformed to a multilayered measure, extending to states and districts, and increasingly in the public eye. Each change in the structure and reach of the NAEP assessment program has made the process of reporting, interpreting and communicating the results more challenging. A chronology of NAEP's history reveals that many incremental changes were made along the way. Nonetheless, some shifts in practice can be thought of as "turning points," in which key changes in the characteristics and direction of the assessment program surface.

The first administration of NAEP was in 1969. The assessments targeted content and processes characteristic of what the majority of students at a given age would have had an opportunity to study and learn. Results were reported on an item-by-item basis for the nation, regions of the country, and certain demographic groups. The items were easily related to the curriculum and trend data was reported while, at the same time, giving teachers, curricular developers, and school officials information about performance at the national level. NAEP's focus on learning was a hallmark of the program throughout its initial development.

Although the item-by-item results were of considerable interest to curriculum specialists, they received limited attention from policymakers and the general public. Starting with the 1984 NAEP assessment, the reporting shifted from emphasizing item results to emphasizing scale scores, which had a number of advantages. Scale scores were familiar to a public accustomed to college admission scores, facilitated summarizing results for an overall content area, such as mathematics, allowed for comparisons among demographic groups, and expedited monitoring changes in student performance over time. The shift in focus from item-by-item results to overall results in a content area served to heighten the interest of policymakers in NAEP results and NAEP became known as the "Nation's Report Card."

In the early 1990s two additional changes were introduced that made NAEP results even more important to stakeholders: For the first time, results were reported state-by-state and in terms of achievement levels—categories specifying the percentage of students who meet established standards of proficiency (in NAEP these are basic, proficient, and advanced). These changes in reporting had the effect of diminishing the attention given to what students know and can do and its inherent relation to curriculum, and increasing the attention on performances by various subgroups of students, defined by demographic conditions related to geographical, racial, ethnic, sociological, and poverty markers.

The technical and procedural complexity of NAEP deepened in the 1980s and 1990s to accommodate new features of the program and to take advantage of some of the sophisticated developments in assessment methodology. The main NAEP assessment, which is administered to national samples in grades 4, 8, and 12, now uses complex psychometric scaling techniques, marginal estimation procedures, and sampling procedures at the state level. National samples for grades 4 and 8 are used for state-by-state reporting of NAEP results in mathematics, reading, science, and writing.

Most recently, the enactment of the *No Child Left Behind Act (NCLB)* in 2002 required states to participate in NAEP at grades 4 and 8 in reading and mathematics every other year, to administer state assessments in reading and mathematics every year in grades 3–8 and once in high school, and to use the state’s own test results to track school accountability. As NAEP’s assessment arm extended to individual states and to a sampling of urban districts, the interpretation of results has become more challenging—and more contestable—as decision-makers at the national, state and district levels apply the results, sometimes inappropriately, to policies and program planning. Thus, what was once a low-stakes monitor of student achievement has gradually evolved into a high-stakes measure that may be used directly or indirectly for purposes of accountability.

Congressional Mandate for Evaluation of NAEP

In light of NAEP’s rapid ascendancy as a powerful policy lever, Congress’ call for an independent evaluation of NAEP in 2002 was timely. The congressional mandate, broadly stated, directed that the evaluators examine whether the assessment program follows accepted professional standards, with particular emphasis given to the achievement levels, sampling procedures, and fairness issues. Given the complexity of NAEP, planning and conducting an extensive evaluation to examine the major components of NAEP is a considerable undertaking.

The evaluation team initially proposed a comprehensive set of studies to analyze multiple facets of the assessment program. However, not all of the studies were funded, and some that were, had to be narrowed due to imposed budget constraints. Based on discussions between the Technical Working Group and the evaluation team, the evaluation focused on four carefully defined issues: the consistency of NAEP’s overall procedures with professional testing standards, the consistency of NAEP procedures for setting NAEP achievement levels with professional testing standards, the validity of state comparisons using NAEP, and the accessibility and understandability of NAEP reports and results to stakeholders.

Uses and Interpretations of NAEP Results

CURRENT USES, INTERPRETATIONS AND ISSUES

NAEP results are currently used for three major purposes: monitoring trends in student achievement; providing evaluative statements regarding the level of student achievement; and making state-by-state comparisons. To allow for the ongoing examination of trends in student achievement, some design characteristics of NAEP have been maintained. However, supporting additional uses of NAEP—evaluating rather than simply describing student achievement and making state-by-state comparisons—required new methodologies.

Evaluating the level of student achievement required NAEP to create standards of student performance by defining levels of student performance (basic, proficient, and advanced) and establishing cut scores along the score scale. Setting achievement levels requires

evaluative judgments regarding the meaning of different levels of achievement, moving NAEP from making descriptive statements about students' achievements to making evaluative statements about students' achievements compared to standards of student performance (NAEP achievement levels). As the current evaluation points out there has been considerable debate regarding the extent to which the achievement levels being employed with NAEP are too high.

Comparing student achievement on NAEP across states is complicated. To appreciate the challenges in making state-by-state comparisons, it is necessary to understand the sampling design adopted by NAEP and its potential impact on the results and their interpretations. In NAEP's multistage cluster sampling procedure, not all students take the assessment, and those students who do take NAEP respond to a subset of the NAEP items in each content area. While this allows for a broad sampling of items from any one content domain, the extent to which subgroups of students are represented adequately in NAEP's state samples is of concern.

As reported in the current evaluation, NAEP's sampling procedures do not ensure adequate representation of various subgroups (including those defined by race and ethnicity) within some states, putting valid interpretations about subgroup performances within a state and across states at risk. Using NAEP to verify state results regarding the achievement of students with disabilities is also problematic because decisions about inclusion and allowable accommodations are made at the state level. Because states vary in their inclusion rates and in their treatment of accommodations for NAEP, the validity of state-by-state comparisons is debatable.

Interpreting NAEP results for grade 12 is very difficult. While states have been required to participate in NAEP at grades 4 and 8 in reading and mathematics every other year under *NCLB*, there is no requirement for grade 12. Consequently, the response rates and participation rates have increased considerably for grades 4 and 8, but not for grade 12. Even if there were a mandate for participation of all students in grade 12, the motivation level of grade 12 students would most likely remain a problem. Concerns with the nonresponse rates and participation rates for grade 12 means any interpretations of the results as an accurate measure of grade 12 student achievement need to be made with caution. These concerns need to be addressed if there are additional uses planned for the grade 12 results, including potential state-by-state comparisons.

A more recent use of NAEP—one that emerged in response to the expressed needs of policymakers and users—is the reporting of district-level results. In 2002, on a trial basis, sampling procedures were modified for several large urban school districts to allow for NAEP results to be reported at the district-level. This additional use of NAEP requires validity evidence to support its use, as does any use of NAEP, as well as consideration of unintended, potentially negative consequences.¹

¹ Although not every unintended consequence can be anticipated, the *Standards* require reasonable effort to prevent negative consequences and to encourage sound interpretations (*Standards*, at 117).

NAEP as a benchmark for state content standards

In an era when concern for accountability is acute, it is inevitable that policymakers will want to use NAEP state results to confirm students' achievement on state tests. However, there is an inherent disconnect between the call for higher-level accountability and the tradition of local control, which has been a hallmark of the nation's public education system and a deeply held value. The tension between the press for higher-level accountability and the prerogatives of local control—for example in determining the scope and sequence of content across the grades—is most apparent in the growing use of NAEP for verifying state assessment results and accountability programs. It is problematic to use NAEP as a benchmark for state assessments due to differences in content standards, population characteristics, standard-setting policies and procedures, and a number of other factors.

In using NAEP to verify a state's assessment results, there is an implicit assumption that the content and skills being assessed by NAEP are similar to the content and skills being assessed by the state assessment. If a state's policymakers perceive that this assumption does not hold, they may alter the state's content standards to be more aligned to the content assessed by NAEP so as to reap the potential benefits of a closer alignment.² The issue at stake is the extent to which state and local content standards and curriculum should be influenced by a national assessment. Such influence may raise concern for local educators, education policymakers, and national content-oriented professional organizations that have always prided themselves with knowing what is best for educating and assessing their students.

NAEP as a benchmark for state assessments

Another issue in using NAEP to verify state assessment results is related to the comparability of achievement levels across NAEP and state assessment programs. It is common to see comparisons of the percentage of students who are at or above the NAEP proficient achievement level and the percentage of students who are at or above the proficient achievement level on state assessments. Although there is considerable variability in the discrepancy between these two percentages across states, with the exception of a few states, NAEP results generally indicate a considerably smaller percentage of students at or above its proficient level compared to state assessment results. Discrepancies between NAEP and state results can be due to a number of factors—differences in the content being assessed, differences in the definition of the achievement levels, and differences in the standard-setting policies and procedures used to establish achievement levels and cut scores. Another factor contributing to these discrepancies is the purposes of these programs. While NAEP has been historically a low stakes assessment for students, schools, and states, state assessments may have higher stakes for schools (i.e., for *NCLB* accountability) and for students (i.e., graduation tests).

We might argue however that the differences in percent proficient or above on NAEP and on some state assessments are so large that they are due to differences primarily in the

² Alignment is illustrated here in one context but can also be used more broadly for describing the degree of concurrence of policies, curriculum, instruction, and assessments within and across grade levels in an education system.

stringency of the NAEP achievement levels rather than due to differences in content coverage. While it is convenient to use the same term, *proficient*, on NAEP and state assessments, it can be misleading because the definition varies across assessment programs. Setting achievement levels and defining the meaning of proficient involves evaluative judgments made within the context in which the assessment is used. Differences in NAEP and state assessment programs, and potential misuses of NAEP in verifying state assessment results, underscore the need for a clear statement of the current and evolving uses, and potential misuses, of NAEP as well as a validity framework to organize the evidence supporting its intended uses.

The utility study in the current evaluation revealed that the differences between NAEP's definition of proficient and individual states' definitions of proficient are not readily transparent to users, leading to potentially inaccurate inferences, comparisons, and related actions. Further, the context of education policy in which achievement levels are set is important to consider when interpreting student results relative to the achievement levels. Evaluations that examine whether NAEP's achievement levels are set too high should take into account the policy context in which NAEP's achievement levels were set relative to the *NCLB* policy environment in which achievement levels were set for state assessments.

A national dialogue regarding priorities in public education and the breadth and depth of local versus state or national authority and control is overdue. Without a frame of reference and explicit delineation of the expectations for degrees of correspondence in both assessed content and achievement levels across states, the use of a national test based on a broadly defined curriculum to verify state assessment results appears to be premature—largely because such interpretations are without a defined reference, making it difficult to gather appropriate evidence to support such interpretations and uses.

Using NAEP in international comparisons

The achievement levels of NAEP have been evaluated by comparing performance of students in the United States and other countries on the Trends in International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA). The current evaluation compared NAEP achievement scores for eighth-grade mathematics with results from TIMSS and PISA. The findings indicated that eighth-grade mathematics students from several other countries performed better than students in the U.S. The proposed validity framework for NAEP needs to address whether international comparisons provide reasonable sources of external validity evidence for NAEP achievement levels. To the extent that they do provide a reasonable basis for comparisons, the framework will need to address how they should be used.

Need for an Organized Validity Framework Given the Complexity and Multiple Uses of NAEP

The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) clearly state the primacy of validity and call for greater attention to continued efforts of validation for all intended interpretations and uses of assessment results. Validation is an ongoing process because it is the interpretation or use of assessment

results that are supported (validated), not the assessment instrument itself. The most important technical characteristics of any assessment are those that address aspects of validity.

Current theory indicates that validation should be comprehensive and explicit, and the higher the stakes the greater the requirement for evidence supporting the proposed interpretations and uses. Thus, as the stakes attached to NAEP results have risen (for example, those implicit in *NCLB*), so has the need for continued validation. Defensibility is not only inherent in the validation process, but has become a legal requirement as well in that case law explicitly recognizes the role of the *Standards* in determining if a particular use of assessment results is defensible.

An organized validity framework takes into account the history of the assessment program, current learning theory, and content-performance expectations from the subject-matter field and related professions. It also addresses contemporary issues in current interpretations and uses of the assessment and anticipates future appropriate and inappropriate uses and consequences of the assessment.

The framework must specify explicitly the interpretations and uses, the assumptions underlying these interpretations and uses, and the kinds of evidence—theoretical, logical, and empirical—that could be brought forth to support these interpretations, uses, and assumptions. A complete treatment of validity would also include the exploration of alternative or competing interpretations or counterarguments. This specification would help the program prioritize validation efforts and resources.

NAEP's design as a cross-sectional survey is effective and cost-efficient for achieving its *original* purposes. However, with each change, policy and legislative customers of NAEP results have been increasingly tempted to use them for new and unanticipated purposes—the attribution of causality in relating background characteristics to achievement, the development of state-by-state comparisons, using national or state results as a benchmark for state assessment programs, and as a measure of the full curriculum in the subject matter domains assessed.

The increased pressure to apply NAEP results in new ways underscores the need for the development of a sound, organized validity framework for the program—one that clearly documents the program's goals and purposes and the appropriate uses of NAEP results along with the uses deemed inappropriate. This would include clear statements of the intended interpretations and uses of NAEP and the types of validity evidence that would support them. An important benefit is that future evaluations of NAEP could then be guided by the validity framework.

Recommendations

The current evaluation identifies a number of worthy recommendations that will enhance and strengthen the NAEP assessment program.

Need for an organized validity framework

As new uses for NAEP continue to emerge, the need for a comprehensive validity framework becomes increasingly critical. The *Standards for Educational and*

Psychological Testing (AERA, APA, and NCME, 1999) provide the foundation for the development of a comprehensive validity framework and a process for identifying the types of evidence that are needed to support the interpretation and use of assessment results. Given the nature of the current and proposed uses and interpretations of NAEP results, multiple levels and sources of evidence are needed in a validity framework for NAEP.

The validity framework should address using NAEP at the national level to measure and monitor student achievement, at the state level to measure student achievement and to make state-by-state comparisons, and at the district level for monitoring student achievement. A validity framework will need to address the multiple levels for which NAEP is used, and the intended uses and interpretations, as well as the potential misuses that can be reasonably anticipated, at each of these levels.

Additional research on achievement levels

The current evaluation examined the application of a new methodology for setting achievement levels on the 2005 grade 12 NAEP mathematics assessment and evaluated the NAEP's achievement levels on the 2003 grade 8 math test using the performance on TIMSS and PISA. It is important to further investigate the stringency of NAEP's achievement levels if they continue to be used as a benchmark in evaluating the results of state assessment programs. NAEP's validity framework will need to address the types of studies that can provide external validity evidence for NAEP achievement levels, including the extent to which international comparisons can provide external validity evidence for NAEP achievement levels.

Additional research

Additional studies are warranted if NAEP is to be used to verify state assessment results. As reported in the current evaluation, there are numerous factors that can jeopardize the validity of interpretations when using NAEP to verify state results. These include differences in content being assessed, differences in standard-setting policies and procedures, differences in the definition of the achievement levels, and differences in the representation of the NAEP state samples. Additional alignment studies that evaluate the congruency between the content assessed by NAEP and state content standards and assessment are crucial. The sampling procedures for NAEP should also be studied. Representation of subgroups across states varies considerably as do the inclusion and exclusion rates for students with disabilities, impacting the validity of the use of NAEP results for state-by-state comparisons and for verifying state assessment results.

The provision of appropriate accommodations for special needs student populations is an area that also needs more study. Additional validity evidence is needed about the accommodations that are used in NAEP for both English language learners and students with disabilities. Furthermore, the criteria for selecting and using accommodations for these students are not defined clearly by NAEP. Only a fraction of these students who are included in the NAEP sample are accommodated. Other studies regarding accommodations for subgroups are also needed, such as an evaluation of the extent to which the accommodations used in NAEP have an impact on the construct being measured, and the implications this may have on interpreting aggregated data.

Given the shifts in demographics, education accountability demands, and the nature of local control of public education, attention to unintended consequences will become even more urgent. Thus the validity framework should not only identify the intended uses and interpretation of NAEP assessment results but also identify potential misuses of NAEP assessment results to help minimize any unintended, potentially negative consequences.

Effective communication strategies to policymakers and relevant stakeholders of NAEP will be essential in promoting valid uses and interpretations of NAEP results. Within this changing landscape, the evolving uses of NAEP need to be considered within a validity framework and future evaluation studies need to be prioritized to support the uses and interpretations of NAEP results in the near future.

Signed,
The Technical Work Group

Jamal Abedi	Cindy Paredes-Ziker
Jeri Benson	Michael Rodriguez
John Dossey	Gregg Schraw
Stephen N. Elliott	Jean Slattery
Michael Kane	Veronica Thomas
Suzanne Lane (co-chair)	Joe Willhoft
Robert Linn	Bruno Zumbo (co-chair)

This page intentionally left blank

Chapter 1:
Audit Study Report

Prepared by:

Barbara S. Plake
Chad W. Buckendahl
Susan L. Davis

Buros Institute for Assessment Consultation and Outreach
Buros Center for Testing
University of Nebraska–Lincoln

This page left intentionally blank

Contents

Figures and Tables.....	1-v
Executive Summary.....	1-1
Summary of Key Findings.....	1-5
Research and Policy Recommendations.....	1-12
Introduction to Audit Study Report.....	1-15
Context for the evaluation.....	1-15
Overview of the audit within the full evaluation design.....	1-19
Focus of the audit report.....	1-20
An Audit of the NAEP Assessment Lifecycle.....	1-21
Background information on the audit.....	1-21
Audit processes and procedures.....	1-29
Conducting site visits.....	1-30
Results from the lifecycle audit.....	1-31
Program Management.....	1-33
Organizational Characteristics.....	1-33
Developing NAEP Assessments.....	1-39
Defining intended uses of NAEP assessments.....	1-39
Developing NAEP Assessment frameworks.....	1-46
Developing Test Items (Questions) and Background Questions.....	1-49
Creating Draft Assessments, Preparing Field Test Designs, and Conducting Field Trials..	1-55
Collecting Data on NAEP Assessments.....	1-57
Constructing Final Assessments.....	1-57
Sampling Schools and Students.....	1-58
Administering NAEP Assessments.....	1-62
Scoring and Analyzing NAEP Assessment Data.....	1-67
Scoring NAEP Assessments.....	1-67
Creating Scales and Links and Analyzing Data.....	1-69
Interpreting and Using NAEP Assessment Scores.....	1-73
Writing, Reviewing, and Disseminating Reports and Data.....	1-73
Setting Achievement Levels.....	1-78
Improving NAEP Assessments.....	1-81
Summary of Key Findings.....	1-89
Key Findings Related to Strengths of the Program.....	1-89
Key Findings Related to Areas for Improvement.....	1-91
Research and Policy Recommendations.....	1-95
References.....	1-99
Appendixes.....	1-103
Appendix A: Glossary of abbreviations and technical terms used in report.....	1-105
Appendix B: Legislation authorizing Evaluation of NAEP.....	1-113
Appendix C: Special Studies in the Evaluation of NAEP.....	1-115
Appendix D: Focus of lifecycle audit.....	1-117
Appendix E: NAEP responsibilities matrix.....	1-121
Appendix F-1: Communication for gathering documents.....	1-123

Appendix F-2: NAEP audit site visit timeline.....	1-125
Appendix G: Site visit reports.....	1-127
Appendix G-1: National Assessment Governing Board (NAGB).....	1-129
Appendix G-2: National Center for Education Statistics (NCES).....	1-147
Appendix G-3: Chief Statistician – NCES.....	1-159
Appendix G-4: Educational Testing Service (ETS).....	1-165
Appendix G-5: American Institutes for Research (D.C.).....	1-175
Appendix G-6: American Institutes for Research (Palo Alto, Calif.).....	1-181
Appendix G-7: Government Micro Resources Inc. (GMRI).....	1-187
Appendix G-8: Human Research Resources Organization (HumRRO).....	1-193
Appendix G-9: Pearson Educational Measurement (PEM).....	1-199
Appendix G-10: Westat.....	1-205
Appendix G-11: NAEP State Coordinators.....	1-255
Appendix G-12: Hager Sharp.....	1-261

Figures and Tables

Figures

Figure 1: NAEP Consortium.....	1-4
Figure 2: The Path to a NAEP Score.....	1-28

Tables

Table 1: Members of the NAEP Consortium and their Roles and Functions.....	1-16
Table 2: Lifecycle audit dimensions and sources of evidence.....	1-23
Table 3: Selected NAEP Validity Research.....	1-44

This page intentionally left blank

Executive Summary

The National Assessment of Educational Progress (NAEP) serves as a broad measure of the level of and change in academic achievement of the nation's elementary and secondary students. The NAEP³ program covers multiple content areas (e.g., reading, mathematics, science) across multiple grade and age levels (e.g., 4th, 8th, 12th) for different populations of interest (e.g., national, state). Although there are a number of NAEP assessments, the core processes for developing, analyzing, and maintaining the assessments are similar. The NAEP assessment system represents the collaborative effort of multiple federal bodies that define policy and oversee operational procedures, and private contractors that implement the operational components.

As part of the *Education Science Reform Act of 2002*, the *NAEP Authorization Act* mandated an evaluation of NAEP and articulated several questions to be addressed in the evaluation (see Appendix B for text of the legislation). These questions were:

1. Whether any authorized NAEP assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);
2. Whether NAEP student achievement levels are reasonable, valid, reliable, and informative to the public;
3. Whether any authorized NAEP assessment is being administered as a random sample and is reporting trends in academic achievement in a valid and reliable manner in the subject areas being assessed; and
4. Whether any of the NAEP test questions are biased; and whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.

In creating the final evaluation design, the evaluation team considered the questions posed by Congress, the magnitude of the NAEP program, the available resources for the evaluation, previous NAEP evaluations, and recommendations from the U.S. Department of Education (ED), National Center for Education Statistics (NCES), National Assessment Governing Board (NAGB), and a Technical Work Group (TWG) of external experts for the evaluation.

The full evaluation design was framed by a psychometric audit of the NAEP lifecycle supplemented by special studies that examined targeted areas of importance. The audit focused on the technical quality of the NAEP program and responded to the breadth of the congressionally mandated questions.⁴ The special studies focused on NAEP achievement levels, consistency in score meaning across various contexts (score equity), the utility of NAEP reports, and methodologies for assessing the alignment of NAEP assessments to state content standards (NAEP-state alignment). These special studies added a depth of analysis to components of these questions based on input from the stakeholder groups noted above.

It is important to note the limited availability of resources and time to conduct this evaluation. We prioritized studies within the evaluation that would be most relevant to ongoing

³ A list of all abbreviations used in this report is provided in Appendix A.

⁴ Qualitative and quantitative data were collected for the audit study primarily between April–October 2005. A component of the design process was to share draft site visit reports with agencies or organizations to review for factual accuracy. Factual statements were reviewed at the time of data collection. For known changes that occurred after the primary data collection period for the audit, we note these as changes throughout the report.

policy discussions about NAEP while responding to the congressional questions. Thus, this is by no means a comprehensive evaluation of the NAEP program. Rather, the purpose of the evaluation is to investigate the operations of NAEP with a focus on several identified areas of importance. This report describes the processes and results of the NAEP assessment lifecycle audit, which begins with identifying the academic content to be tested or assessed and continues through the reporting of students' achievement on the assessment.

A psychometric audit such as this is based primarily on evaluating the quality of available documentation of a testing program's processes and results by applying professionally adopted standards of practice (i.e. *Standards for Educational and Psychological Testing*, American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME), 1999). This study responds to the congressional question in the mandate about how well NAEP meets professionally accepted standards for testing. It also addresses some of the other issues inherent in other congressional questions specified in the mandate as they relate to ongoing assessment development and maintenance (e.g., assessment administration, sampling, test question review for content and bias).

Specifically, the audit was framed around 13 dimensions (identified in italics below) selected by the evaluation team with the assistance of ED. We first considered the *organizational characteristics* of the NAEP program including structure, oversight, staffing, communication, and problem resolution. Our review of the operational procedures began with the processes used to define the *intended scope and uses of NAEP assessments* as the professional standards identify this as the cornerstone for validity of any assessment score. With regard to the development of the NAEP assessments, we considered the procedures used to *develop the NAEP assessment frameworks*, *develop the NAEP items and background questions*, and the pre-administrative tasks of *creating the draft assessments* and *conducting the field tests*.

Several steps were examined when considering the procedures used to collect data on the NAEP assessments. The *construction of the final assessments* involves coordination with multiple contractors and relies on strong communication and cooperation among these members of the NAEP alliance. After the final exam forms are created, the *sample of schools and students* is selected and then NAEP contractors work together to *administer the assessment*.

The raw data from the assessments are then transferred to other NAEP contractors who are responsible for the processes used to *score the NAEP assessments*. The scored data is then used to *create the NAEP scales and links and analyze the data*. There is some controversy in the measurement field about procedures used for estimating how a student may have responded to a full-length assessment. Note that students selected for NAEP do not take the full assessment. Instead they take a smaller sample of the full assessment. During the data analyses, estimations are made about the performance on the full assessment based on the examinees' performance on the subset, and other information.

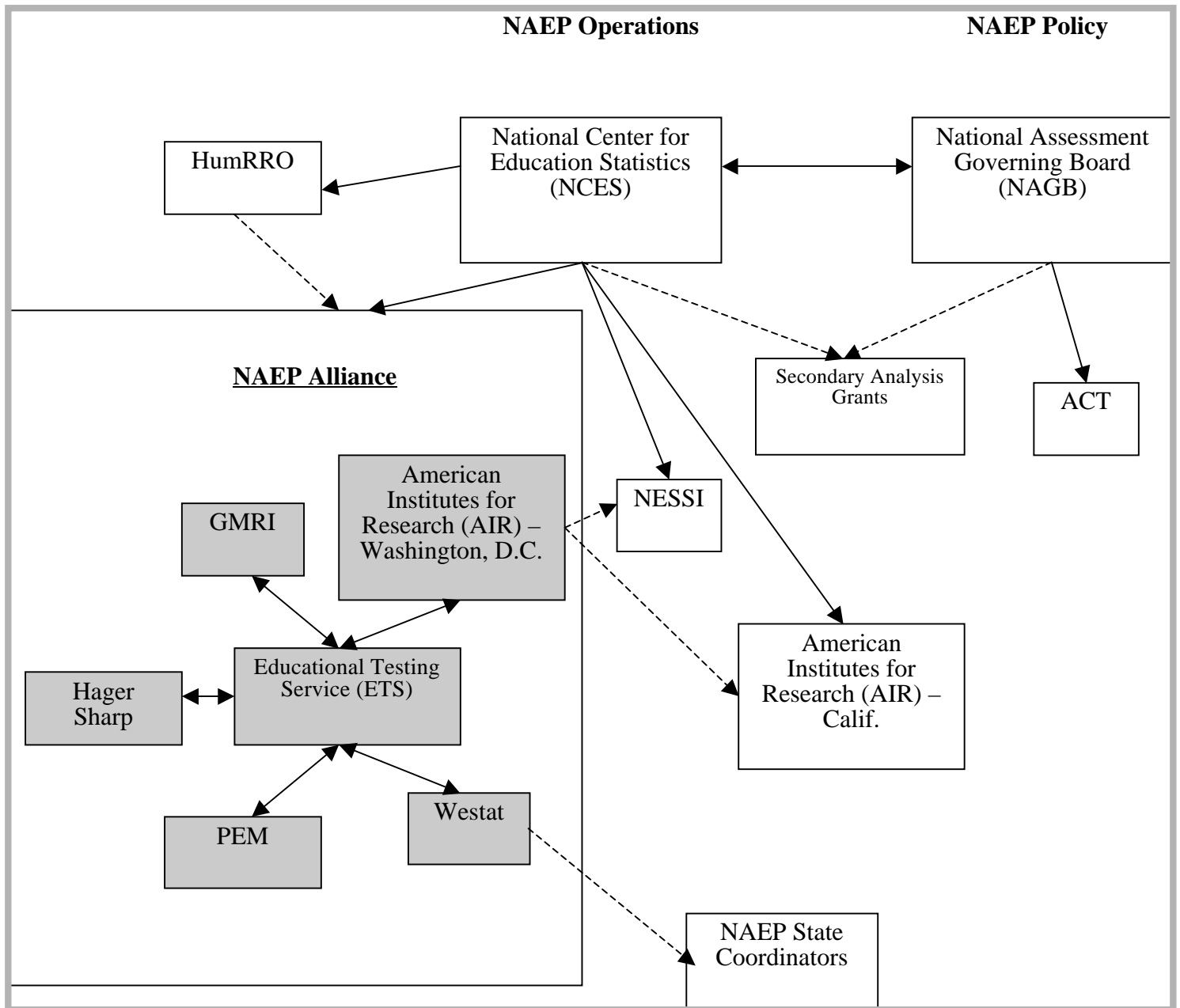
The final NAEP dataset is then prepared for reporting purposes. The reporting of achievement levels is a central feature of the interpretation of NAEP results. The procedures used in setting these achievement-level performance standards (sometimes called "standard setting") are therefore critically important. Therefore, it was an important step in the evaluation project that we review the processes used to *set achievement level standards*. After the achievement levels have been set, the final phase of the NAEP process is to *write, review, issue, and disseminate reports and data*. Finally, we examined strategies in the NAEP program for *renewing and improving the assessment* process through innovations for use in future assessments.

Figure 1 illustrates the NAEP consortium that was the focus of this audit. There are two general areas in which we can identify key players in the program: NAEP policy and NAEP operations. As the policy body for NAEP, NAGB is responsible for setting policy for the NAEP program within the framework established in law by the Congress, overseeing the development of

the assessment content frameworks, approving all questions to be included in an assessment, creating the performance level descriptions for reporting achievement levels, overseeing the setting of achievement level standards, and releasing initial NAEP results. NAGB oversees the work of its contractors, like ACT which assists in the achievement level studies. NAGB also helps to set priorities for the Secondary Analysis Grant (SAG) program administered by NCES.

As the organization responsible for operations, NCES implements the policies articulated by NAGB, produces and administers NAEP assessments, oversees contractual relationships, and reviews and releases technical reports for the program. The contractors that carry out these operational responsibilities include Educational Testing Service (ETS), Westat, Pearson Educational Measurement (PEM), American Institutes for Research (AIR, Washington, D.C., office), Government Micro Resources, Inc. (GMRI), and Hager Sharp. Each has experience in different areas of the program, but collectively the group is called the NAEP Alliance. ETS coordinates activities among the contractors in the Alliance and is responsible for a number of activities included scaling, linking, and data analysis for the programs. Westat is primarily responsible for sampling schools and students, but also has responsibility for the administration of NAEP assessments and supporting NAEP State Coordinators. PEM scores the assessments and transfers results to ETS. AIR-DC develops background questions and items for some assessments. GMRI develops and maintains the Web sites for NAEP and also creates an online management system that assists contractors in their communication with each other. Hager Sharp assists with the dissemination of NAEP results. In addition, NCES has direct contractual relationships with the Human Resources Research Organization (HumRRO), AIR (Palo Alto, Calif., office), and the NAEP-Education Statistics Services Institute (NESSI, formerly ESSI) to provide various services in support of the NAEP program. Specifically, HumRRO conducts quality control activities, AIR-PA is responsible for the NAEP Validity Studies (NVS) panel, and NESSI provides statistical support activities for the program.

Figure 1. NAEP Consortium



Note: The shaded boxes in Figure 1 represent the NAEP Alliance. These are the contractors responsible for the operational activities involved in NAEP assessments. NCES has direct contracts with each organization and oversees the Alliance. The processes and procedures that these organizations use to develop and administer NAEP were the primary focus of the lifecycle audit study.

Procedurally, this lifecycle audit began with a review of documentation on NAEP’s processes and results provided by each of the major organizations involved with NAEP’s policy and operations. We followed this review with site visits to key organizations to interview personnel and clarify or collect additional information that was unclear or absent from the documented materials submitted in advance. Because the NAEP program continues to evolve, the audit took a broad look at current

NAEP practices, processes, and results, in addition to examining the available validity evidence supporting these practices and processes and the interpretations of NAEP results. It was nevertheless limited by the availability of documents or information provided by the agencies and organizations responsible for NAEP.

Summary of Key Findings

Based on the information we were able to gather during our review, it appears that most operational components of the NAEP assessment program were functioning well and were in compliance with sound measurement practices and with the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999). However, there were a few key exceptions. The major exceptions that have the potential to threaten the validity of the program were the absence of a formal validity framework to organize and prioritize evidence to support that validity of score interpretations and use, and the lack of current technical documentation and reports to support the psychometric properties of the NAEP assessment program.

The key findings are organized into two sections. The first presents key findings that were identified as strengths of the program. The second set identifies areas for improvement. Within each section, the findings are organized by importance.

Key Findings Related to Strengths of the Program

Key Finding 1: Main and State NAEP assessments in reading and mathematics are developed, implemented, and maintained in ways that are generally consistent with widely accepted professional assessment standards.

Through this evaluation we were able to explore many aspects of the NAEP program described in the previous section. Except for a few noteworthy exceptions, the methods and procedures used for the Main and State NAEP assessments in reading and mathematics were found to be in compliance with these widely accepted professional assessment standards. This compliance was noted throughout the development, implementation, and maintenance of the program.

The processes used for creating the assessment frameworks are firmly grounded in policy and the review and revision procedures were consistent with sound measurement practices. Further, we found that the methods used by the Alliance contractors to develop and review the NAEP assessment questions are consistent with the *Standards* and follow sound measurement practices. The methods used for field-testing items appear to be technically and psychometrically sound as they involve using embedded field test blocks within the operational administration. This helps to ensure accuracy of the field test data.

We found that systems are in place to support communications and cooperation among the contractors preparing for and conducting the administration. This is an important feature as the administration of the NAEP assessments relies on the coordinated effort of multiple contractors and NAEP state coordinators. We found that the electronic monitoring systems for tracking the materials is a strength of this process as it helps with the administration process and maintaining security of the test materials. Overall, the scoring procedures are generally compliant with the *Standards*; however, there is one exception that is noted in a later finding. In addition, although there is not agreement in the measurement field about which methodologies are the most statistically sound for estimating student performance on a full assessment when they only take a sample of the items, the procedures used for the NAEP assessments are consistent with those used in several other large-scale, international assessments and are generally consistent with the *Standards*. Overall, the psychometric characteristics of the NAEP assessment scores (e.g.,

reliability, standard error) all support the technical quality of the results. The procedures and timelines for the initial release of NAEP results are in compliance with the *Standards* and the NAEP Alliance responded well to the increased pressure to disseminate results and data in a timely and user-friendly fashion.

We found that there are ample opportunities in the NAEP program for gathering information to support renewal and innovations through several research programs that are a part of the NAEP system. The topics of these projects span the NAEP assessment program. However, we are concerned that these opportunities are neither systematic nor integrated—this is detailed in a subsequent finding.

Although the majority of the processes in the NAEP system were found to be compliant with professionally accepted standards, this evaluation of the psychometric (i.e. technical) quality is limited for two reasons. First, the *Standards* clearly specify that evidence of psychometric quality does not exist in a vacuum. Psychometric quality is related specifically to the defined, intended uses and purposes of the assessment. The intended scope and uses of NAEP assessment results are only defined broadly, leaving room for confusion and lack of clarity about which uses and interpretations are intended and which ones are not. Second, our review of technical criteria was limited to the available NAEP technical manuals (e.g., 2003 NAEP Technical Manual) and some of these conclusions were made based on assumptions drawn from dated material about the NAEP program.

Key Finding 2: Methodologies to establish achievement levels were generally consistent with the expectations of the Standards.

The process of setting achievement levels on NAEP assessments has been both highly criticized (e.g., Shepard, Glaser, Linn, and Bohrnstedt, 1993; U.S. General Accounting Office, 1993) and defended (e.g., Hambleton et al. 2000; Loomis and Bourque, 2001). Two prior evaluations described the NAEP standard setting as “fundamentally flawed” (Shepard et al., 1993; Pellegrino, Jones, and Mitchell, 1999); however, some reactions to those evaluations from standard setting researchers were very critical.

These findings are related to the congressional question about the validity and utility of NAEP achievement levels. The *Standards* (AERA, APA, and NCME, 1999) provide guidance on appropriate practice with respect to setting achievement levels (sometimes called standard setting). For example, *Standard 4.19* suggests, “when proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented” (p. 59). Also, *Standard 4.20* indicates, “when feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria” (p. 60). With respect to the judgmental process, *Standard 4.21* suggests that, “. . . the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way” (p. 60). Within the audit, we reviewed information from the previous methodology used by NAGB to establish achievement levels.

Our findings revealed that one of NAGB’s purposes for developing achievement levels was to assist policymakers and other stakeholders in their ability to interpret NAEP scale scores. To facilitate these activities, NAGB also developed Achievement Level Descriptions (ALDs) that provide broad policy definitions of what students should know and be able to do at a given level. These ALDs are then applied to the respective content in more depth during the processes that establish the achievement levels. For these studies, panelists are selected who have content knowledge, some familiarity with the target population of students eligible to take the assessment, and who represent different education stakeholder communities and the public. The results of these activities ultimately represent a policy decision that is within the scope of NAGB’s responsibilities.

As is the case with most policy decisions, there is an element of judgment that goes into the final decision. However, in education these types of value-based decisions are also made at the state level (e.g., levels of student proficiency), in a classroom (e.g., assigning grades of A, B, C, D, F), and with individual students (e.g., what is the best instructional strategy to help this student succeed?). Given the controversy surrounding this topic, a special study within the full evaluation also reviewed a newly employed standard setting method for the 2005 Grade 12 NAEP Mathematics assessment.

Based on the information we gathered during the site visits and through the technical documentation, it appears that the methodologies used to set NAEP achievement levels generally comply with professional technical standards. In particular, there is clear documentation on the rationale and procedures used for setting the achievement levels. The new methodology applied with the Grade 12 Mathematics assessment had features that were designed to aid the panelists in making their judgments in a manner that is consistent with their knowledge and experience.

Key Finding 3: Current structure of NAEP Alliance contracts facilitates cooperation and communication among contractors.

One of the notable strengths of the NAEP program is the organizational and contractual structure of the relationships among those responsible for NAEP assessment operations (i.e., the NAEP Alliance). Under the new procurement model that began in 2002, previous subcontract relationships were changed to direct contractual relationships with NCES. One characteristic was the establishment of a contract for Alliance coordination to facilitate activities among NAEP contractors. Another feature of the contract is the use of built-in incentives for the members of the Alliance to meet mutually beneficial goals and timelines. This facilitates an atmosphere of cooperation as all contractors benefit when the system is working and all lose out on financial incentives if the system strays from critical path timelines and deliverables.

An additional example of contracts that helped to ensure effective and efficient operations under the new procurement model was the establishment of the Quality Assurance contract that was designed to provide external staffing and support for NCES to monitoring the quality of the NAEP Alliance and operations.

A related strength is the observed communication among Alliance contractors. Within the NAEP Alliance one of the strategies to support this strength is a Web-based tool called the Information Management System (IMS). The IMS facilitates communication among contractors regarding progress, timelines, and discussion and resolution of problems. The features of this online tool provide a common language and structure to the Alliance when integrating systems from different organizations. The IMS also allows for greater decentralization of key personnel because it was developed as secure, Web-based solution and provides a forum for contractors to discuss issues or problems that arise.

Key Finding 4: Psychometric characteristics of NAEP assessment scores are consistent with professional standards for testing.

Our review of technical criteria was limited to the available NAEP technical manuals (e.g., 2003 NAEP Technical Manual) and some of these conclusions were made based on assumptions drawn from dated material about the NAEP program. The technical quality included in the available documentation provided strong and supportive evidence of technical quality, especially with regard to estimates of score reliability and standard errors of measurement. These technical characteristics support confidence in the scores. This document also provided information about procedures used to ensure that assessments were fair to protected groups through analysis of differential item

functioning and item reviews for biasing features. We anticipate that when the technical information is available for the current assessment they will report equally strong evidence of psychometric quality and provide even more evidence of how these assessments comply with the *Standards*. This conclusion is drawn, in part, from historical reports that have been released documenting the NAEP program.

Key Findings Related to Areas for Improvement

Key Finding 5: Intended uses of NAEP assessment scores were not clearly defined.

This finding relates to a critical need for all assessment programs: providing a clear definition of the intended and unintended uses of scores from their assessments. The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) has, in the first chapter, specific expectations of test publishers regarding defining intended uses of test scores and the validity evidence needed to support them. For example, *Standard 1.1* notes that a rationale needs to be presented for each recommended interpretation or use of test scores. Because no test is a gold standard (i.e. valid) for all purposes and all situations, *Standard 1.2* specifies that test developers clearly articulate the intended interpretations and uses of test scores. Because the potential for misuse of assessment data and the resulting consequences are critically important for NAEP, unintended uses of scores are also important to clarify for potential stakeholders. *Standard 1.4* indicates that if a test is being used in a way for which it has not been validated, users need to justify the new use and collect new evidence if necessary. This finding responds to the first congressional question that asked whether NAEP assessments were meeting professionally adopted standards.

The current uses of NAEP are broadly defined by legislation leaving the actual uses open to a range of interpretation. Congress and other stakeholders may be using NAEP results for purposes that are not supported by validity evidence. Understanding and clarifying those intended and unintended uses will assist NAGB, NCES, and key stakeholders in developing a validity framework for the program broadly and then prioritizing validity research efforts to target those intended uses that are most critical to the defined uses. It is important to note that validity research opportunities occur multiple times across many of the NAEP contractors, not the least of which is the NVS Panel that operates under a contract with the AIR-CA office. In addition to these efforts, five additional sources of operational validity evidence were cited by NCES: NAEP's Design and Analysis Committee (DAC), Task Order Component (TOC) opportunities, assessment development processes, NESSI, and the NAEP SAG program. Research is also funded through separate programs within HumRRO, AIR-CA office (e.g., state analysis contract), and ETS. Our review of several of these research programs suggests that they have the potential to provide critical information that could support the intended uses of NAEP scores and be used in the continual development and refinement of NAEP. However, because specific, intended uses are not currently defined, there is not a transparent validity framework that organizes and prioritizes studies conducted through these various research efforts. This is a lost opportunity to inform, engage, and provide targeted information pertaining to such a validity structure to communicate the strengths of the program and its uses to policymakers.

Key Finding 6: Lengthy review processes limit the availability and utility of NAEP technical manuals and reports.

The protocol for review and dissemination of NAEP-produced technical manuals and reports that document the program's activities is extensive, and in many ways critical to ensuring

that NAEP publications are accurate both technically and factually. The review process includes multiple reviews by individuals with different areas of expertise—the specific process differs depending on the type of document being prepared for release. However, because of such an extensive and thorough review process, the outcome is that many important NAEP related documents are not available, therefore missing the opportunity to share high quality, technically and factually accurate information about the NAEP program. Given their role as the agency responsible for program operations, there are more reports that go through the review process at NCES; however, NAGB’s review and dissemination practices are also subsumed within this finding. This finding also relates to the first congressional question regarding the program’s adherence to professional testing standards.

To highlight this problem, we note that the most recent released technical manual that could be reviewed for this NAEP evaluation was the 1999 Long Term Trend technical report that was released in April 2005. Although we were provided access to Web-based versions of draft technical reports from 2000–03, it is unreasonable that technical documents for assessments that were administered and results disseminated in the years 2000–05 should still be under review. In addition to the extensive review process, another contributing factor to this delay is that these reports are given a lower prioritization as the focus was primarily on the six-month reporting requirements thus causing many of the delays in the release of technical documentation. Although we understand the burden presented by the six-month reporting requirement, the lack of available technical documentation violates professional expectations. Another illustration of this timeline is NAGB’s initial release of the 2005 NAEP 12th Grade Reading and Mathematics assessment results. These initial releases did not occur until Feb. 22, 2007.

As with the finding of the lack of clearly defined intended uses of NAEP assessment scores, the *Standards* (AERA, APA, and NCME, 1999) expect testing programs to provide documentation for their program(s). For example, *Standard 6.1* suggests that test documents (e.g., test manuals, technical manuals, user’s guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use. In addition, *Standard 6.3* indicates that this documentation include the rationale for the test, recommended uses, support for such uses, and information that assists in score interpretations. Although some lag time may be expected due to a comprehensive review process, the current timeline for the release of technical documentation extends beyond what a large-scale testing program should tolerate, and is in violation of the *Standards*.

Key Finding 7: NCES’s Assessment Division is understaffed to respond to current demands of the NAEP assessment program.

The NAEP assessment program relies on a series of interactions among the numerous organizations and agencies involved in the development, administration, and dissemination of NAEP assessments and results (See Figure 1). NCES’s Assessment Division staff members play a number of roles in the lifecycle. Most important, they oversee the work and deliverables that the Alliance contractors produce. The contracting officer’s representatives (COR) at NCES are also responsible for facilitating communication among the NAEP contractors and those external to NAEP (e.g., secretary of education, policymakers, evaluation team) and assisting in resolving any issues that arise. The Assessment Division of NCES has 20 full-time employees. This is a small staff when compared with other divisions within NCES that have similar budgets but 80 or more full-time employees. Currently, the Assessment Division staff members oversee the work of approximately 1,300 permanent and temporary employees working for various NAEP contractors. Although more than half of these employees are involved primarily in the administration of NAEP assessments, this number is large considering the number of staff within the Assessment Division

and responsibilities they have in terms of overseeing quality control procedures for these contractors. Although not directly related to any one congressional question, the capacity for organizations within the NAEP Consortium to respond to the needs of the program is related to all of the questions mandated in the evaluation legislation.

In addition, as the operations agency for NAEP, the NCES Assessment staff members are responsible for responding to requests for information from multiple stakeholders and responding to questions or inquiries about NAEP results or the proper interpretation of these results. NCES also needs to maintain a close relationship with NAGB to provide input and respond to policies that impact the program's operational activities. The Assessment Division staff members also assume responsibility for reviewing and disseminating technical reports that document program activities (See also Key Finding #6). After noting the many responsibilities of the Assessment Division's staff, it was apparent to the evaluation team that this part of the NAEP consortium is dangerously understaffed to respond to these increasing multiple program needs.

Key Finding 8: Some current uses of NAEP assessments may not be accounted for in the current sampling plan.

Sampling procedures represent an important component in the NAEP assessment program that has a long tradition of driving advances in survey technology. Many of the survey and weighting procedures now used are adequate and consistent with generally accepted methods in sampling. However, the intended uses of NAEP assessments influence how the sampling design is developed and implemented. For example, collecting representative data for the nation requires a different sampling frame than collecting representative data for a state or an urban school district. The sampling frame also extends to student groups. Although this makes intuitive sense, as the intended uses and the policy contexts for NAEP assessment scores are clarified, further evaluation of current sampling practices is necessary. Some of these policy considerations that are unique to NAEP sampling methods are described here and directly relate to the congressional question regarding whether NAEP assessments were conducted as a random sample.

First, appropriate accommodations for the NAEP assessment are expected to be provided to sampled students who require them. Two subgroups of students are most affected by this: students with disabilities (SWD) and English language learners (ELL). On the surface, more widespread awareness and use of testing accommodations would appear to lead to an increase in the overall percentage of students included in the assessment as well as the consistency across states in student inclusion rates. However, different inclusion rates and cross-state consistency remain a problem. States differ in their rates of exclusion and also in the accommodations they provide to special needs students who are not excluded. Thus, even included students may have incomparable test experiences in different states. Differential exclusion rates threaten any state-by-state comparisons.

Second, factors that reduce the initial sample, specifically school and student nonresponse and refusal to participate, represent a significant potential threat to the validity of NAEP assessment scores. Although not directly addressed in the legislation, the *No Child Left Behind (NCLB)* legislation has raised the visibility for NAEP and discussions about potential uses of this data such as comparisons across states have been occurring. At the same time, *NCLB* has changed the context in which NAEP operates and may indirectly change the nature of student and school nonresponse in NAEP assessments.

Third, state samples must be adequate in size and representation to provide reliable estimation of performance. Estimation at the state level has traditionally required sample sizes of about 2,500 students from approximately 100 schools per subject area assessment. Because the specific intended uses of NAEP assessments are not clearly defined (See Finding #5), policymakers' interest in NAEP scores often does not stop at the national or state level for all

students. For example, reporting is also required for historically prioritized student subgroups (e.g., ethnicity, lunch program status, language proficiency, and student disability). NAEP has traditionally taken steps to oversample students in some key subgroups (e.g., sampling schools with larger representation of blacks and Hispanics at double the rate of other schools). Today many states are seeing significant demographic changes; furthermore, demographic characteristics differ substantially from state to state. At the same time, some of the most significant data problems faced by NAEP involve missing Title I data and the representation of these students, uncertain National School Lunch Program data, and problems with some schools' identifications of racial or ethnic status. All of these issues can affect sampling via less accurate sampling frames and the incomparability of results over time.

Fourth, several schools and districts are sampled with certainty or near certainty across multiple NAEP assessments. As such, what appears to be a random sample in a given year may be more systematic when considered across multiple NAEP administrations. Even though the student sample in certainty schools is refreshed annually, students in these schools may share characteristics that are not shared with students in non-certainty schools. Although this may not yet lead to measurement concerns, as the level of certainty in the sample increases, the data may be increasingly viewed as similar to census (entire population) rather than sampled information. As school professionals become familiar with the NAEP assessment, scores of their students may improve in ways that may not be shared with students in districts for which NAEP is a more novel experience. On the other hand, districts repeatedly selected for NAEP participation may experience some fatigue with and resistance to the assessment, adding another potential threat to the validity of these results.

Key Finding 9: Procedures for scoring constructed-response questions are not fully consistent with best practice.

This finding focuses on procedures employed in scoring constructed-response questions for NAEP assessments and relates to the congressional question about whether NAEP assessments adhere to professional standards. Two issues emerged through our evaluation efforts. The first issue relates to protocols for what happens when a student paper is selected for double scoring to estimate inter-rater agreement reliability. In these instances, the score assigned by the second rater is not used, even when it deviates from the score assigned by the first rater. Only the score assigned by the first rater is used in scoring. Given the subjective nature of the scoring guidelines for these item types, we noted two concerns with this practice. First, some raters score at a pace that is more rapid than others when scoring student responses. For these situations, the more rapid raters' scores will be "counted" more often as the operational score. Second, if the scores assigned by the two raters differ, it indicates some potential inaccuracy or at least, uncertainty about our confidence in the resultant score assigned to the performance. Note that if the intended uses of NAEP assessment scores expand in scope beyond the current low-stakes assessment system that does not directly impact individuals, schools, or most districts, these scoring practices would become more critical to our confidence in the resulting scores and decisions.

The second issue within this finding relates to practices for scoring validity papers. Validity papers represent student performances with "known" scores that are included in the scoring process to monitor the consistency and accuracy of raters' performance throughout the scoring process as a quality control strategy. Previously, validity papers were scored as an "event", so that raters knew when a paper would be used as a validity check. This strategy has the potential to influence raters' performance if they know which student performances are being used to monitor the quality of their scoring. During the evaluation, the NAEP Alliance was pilot testing a new strategy for embedding these validity papers so that they would not be scored as an event.

Research and Policy Recommendations

The NAEP assessment lifecycle audit was intended as a broad look at a multifaceted testing program to evaluate important steps in the development, maintenance, and improvement of NAEP processes. Although some select topic areas were evaluated in-depth through special studies within the overall evaluation (See Appendix C), there are aspects of NAEP that could not be investigated in this evaluation because of limited resources but would benefit from additional study. Some of these (e.g., unclear definition of intended uses of NAEP, limited availability of NAEP technical documentation) have been highlighted in the findings noted above. In this section, we have included specific recommendations for the NAEP program that flow from the findings described above and briefly note some areas for additional research that were beyond the scope of this evaluation, yet important to the NAEP program.

Recommendation 1: We recommend that the NAEP program develop a transparent, organized validity framework beginning with a clear definition of the intended and unintended uses of NAEP assessment scores (*Standard 1.2*). The specification of the intended uses and the development of an organized validity framework should be a joint responsibility of NAGB, NCES, and additional stakeholders (e.g., educators, policymakers). As indicated by *Standard 1.1*, a rationale, supporting research and documentation should be provided to justify the intended use(s) of any test score. Review of previous or ongoing NAEP research as is described in the body of the report will likely provide support for the intended uses; however, it is expected that reviewing this body of work will reveal some overlap as well as areas in which sufficient work has yet to be conducted. The validity framework can build on existing research and be organized in a way that supports validity issues in development, program maintenance, and future directions of the program.

Given the importance of a highly visible national assessment program, it is essential that a validity framework be created to coordinate a program of validity research on NAEP, aimed at informing the validity of score interpretation and use. This should be a highlighted component of NAEP; particularly as its perceived role has evolved in the wake of *NCLB*.

Recommendation 2: We recommend that NAGB continue to explore achievement level methodologies as applied to NAEP and consider employing multiple methods with future studies to better inform the policy decision and communicate the policy nature of the decision. The interpretability of NAEP scale scores through the use of achievement levels was an initiative identified by NAGB to aid the public and policymakers. As setting achievement levels is ultimately a policy decision, it is within NAGB's scope to define, establish, and interpret these scores. It is generally accepted among measurement professionals that different methods for setting achievement levels typically produce different results (Jaeger, 1989). Thus, the selection of any one methodology to gather judgments, whether on test characteristics (e.g., Angoff, Bookmark, Mapmark) or examinee characteristics (e.g., borderline group, contrasting groups), only provides one source of evidence for the resultant policy decision. Thus, we further recommend that NAGB consider additional sources of external validity evidence that would be informative to the final policy decision. Some of these sources at the high school level may include results from additional methods, ACT or SAT scores, state university entrance levels, and transcript studies that evaluate course performance. By triangulating these sources of evidence, the cut scores and the resultant impact would strengthen the validity argument.

Recommendation 3: We recommend that NAGB's and NCES's current review and release processes for technical manuals and reports be revised to streamline these efforts while still ensuring high quality and accuracy of NAEP reports. For example, technical information for the aspects of NAEP that have not changed (e.g., test development, scaling procedures) should be

publicly available, and information for the most recent tests should be released simultaneously with the test results. This approach would not require reproduction of voluminous technical manuals that repeat much of what is contained in earlier reports but would rather reference the existing reports and present only information related to the most recent assessments. Although some efforts in this direction have been made as the NAEP technical manuals are transitioning to a Web-based medium, this transition was incomplete during the course of this evaluation.

Recommendation 4: We recommend that the current staffing capacity for NCES's role in NAEP be increased to respond to the increased magnitude of the program. Current NCES staffing levels are inadequate to respond to the operational demands placed on NAEP. To respond to operational needs, some of the activities that may otherwise be conducted within NCES are outsourced to contractors to sustain the program.

Recommendation 5: We identified three areas in which additional inquiry is needed in response to the changing policy context of NAEP assessments that have implications for changes in the methods used for sampling. First, we recommend further study that addresses the impact of differential exclusion and accommodation of special needs students (SWD and ELL) across states. Strategies for estimating the impact of exclusion—including full population estimation (a statistical method for predicting scores in the full population of students) work done at AIR-CA—appear promising as ways to improve the comparability of State NAEP scores. These and additional strategies should be further explored as well.

Second, we recommend exploration of several questions regarding nonresponse and refusal to participate in NAEP in the current context. Some of these research questions may include: a) What is the impact of nonresponse on NAEP estimates? b) How do the current methods of replacement affect the results? and c) How do these participation rates impact the 12th grade assessments?

Third, we recommend further exploration of whether NAEP samples are sufficient to support robust estimation of subgroup performance within states or other intended populations. This area of study is important because some of these inferences regarding subgroups were not necessarily intended at the time these sample sizes were determined. The ability of state samples to provide accurate, valid estimates of subgroup performance in the face of challenges and demographic changes in states and nationally needs to be examined. Related to this recommendation is the need for additional analyses to estimate the impact of repeated administration in units often (or always) selected for NAEP.

Recommendation 6: We recommend that policies and practices related to scoring constructed-response questions, particularly as they relate to the use of the scores assigned by second or subsequent rater, be studied. We also recommend that the NAEP program develop strategies that improve the current practices related to embedded validity papers to monitor the accuracy of raters' performance during the operational scoring procedures. These improvements will help ensure that the validity data derived from these papers more accurately represent the validity of the rating process.

Recommendation 7: We recommend that future contracts for NAEP involving multiple contractors build on the positive experiences learned in the use of the Alliance, Alliance Coordination, and Quality Assurance contracts. The continuation of incentives for cooperative, positive outcomes in an Alliance-like contract is also recommended because it appears to be effective in facilitating collaboration among the members by helping distribute responsibilities for the success of the program to all contractors within the Alliance.

Additional Research: One additional area of research that has the potential to greatly influence policy considerations is what could be characterized as “alignment.” As used here, alignment refers to the overlap among the NAEP assessment content frameworks and state academic content standards for elementary and secondary education; state assessments and NAEP assessments; and state assessments and NAEP assessment frameworks. Because NAEP is often used by the public as a basis for comparing results from state assessments, whether defined as an intended use or not, further exploration of this area is necessary to properly understand the limitations of such interpretations.

Introduction to Audit Study Report

The National Assessment of Educational Progress (NAEP) annually assesses samples of 4th-, 8th-, and 12th-grade students from public and private schools across the country. Depending upon the year, students may be assessed in reading, mathematics, science, writing, U.S. history, civics, geography, or the arts. The results of these assessments are reported at different levels of specificity. For example, Main NAEP includes assessments across a number of subject levels at grades 4, 8, and 12 and reports results on a national level. State NAEP assessments are administered at grades 4 and 8, but are limited to reading, mathematics, science, and writing. These results are reported on the state level. Trend NAEP is administered to 9-, 13-, and 17-year old students in reading and mathematics.

These large scale assessments are administered to a sample of students from across the country that are defined by the scope of the score. For example, in 2005, the Main NAEP assessment was administered to over 300,000 students in reading and over 300,000 students in mathematics as well as for the State NAEP assessment in these subject areas. From these assessments, reports are produced for different groups of stakeholders—over 150 national and state reports and dozens of informative documents were developed from the 2005 mathematics and reading data.

Context for the evaluation

The NAEP program provides information on the educational achievement level of students nationwide. Unlike many large-scale testing programs such as those administered by states, the reporting of NAEP results is not at the individual student level. Instead, NAEP results summarize the achievement of students at a higher aggregate level, such as states and the nation. Also, unlike most large scale testing programs administered in the states, not all students take the assessment. Instead, a complex sampling procedure is used to ensure that the results are generalizable to the student population at the appropriate grade level. Such features of the NAEP assessment program, including no individual reporting of student results and a sample administration, pose some special challenges to both the agencies who are responsible for the NAEP program and to the people who use and interpret NAEP results.

The NAEP program is the outcome of many cooperative organizations and agencies. Figure 1 (see Executive Summary) provides a comprehensive overview of NAEP's programmatic structure. There are a number of key agencies and contractors who together have the responsibility for the NAEP assessment program. The primary agencies are the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES). NAGB⁵ has major responsibility for policy level decisions about NAEP, including the specifications for the test framework and overseeing the release of NAEP results. NCES has primary responsibility for the development, delivery, administration, scoring and reporting of the assessment results and to ensuring that the assessments continue to evolve with current technical advances in the testing industry. NCES achieves this outcome through its contracts with the NAEP "Alliance", a number of contractors who work in tandem to produce the NAEP assessments. Principal contractors within the Alliance are Educational Testing Service (ETS), Pearson Educational Measurement (PEM), ACT, American Institutes for Research (Washington, D.C., office, AIR-DC), Westat, Government Micro Resources Incorporated (GMRI) and Hager Sharp. As will become clearer when the results of the audit are presented later in this report, each of these contractors has well identified roles in the NAEP assessment program. In addition to the NAEP Alliance, Human Resources Research Organization (HumRRO) has a special contract outside of the Alliance to provide quality control

⁵ A glossary of all abbreviations used in this report can be found in Appendix A-1.

oversight. NAEP State Coordinators also provide important service to the NAEP assessment program through their on-site state level access to NAEP programs and procedures. The American Institutes for Research (Palo Alto, Calif., office, AIR-CA) also contributes to the NAEP system as the organization responsible for some of the validity research related to the program. Taken together, for this report, these agencies and contractors are called the NAEP Consortium. More details about the individual roles and responsibilities for the members of the NAEP Consortium are shown in Table 1.

Table 1: Members of the NAEP Consortium and their Roles and Functions

Organization	Role and Function
<i>National Assessment Governing Board (NAGB)</i>	This independent federal body is appointed by the secretary of education to set policy for the NAEP program. NAGB is responsible for the development of the assessment frameworks, approval of all questions included in an assessment, creation of the achievement level descriptions, setting achievement level standards, and disseminating the initial release of NAEP results.
<i>National Center for Education Statistics (NCES)</i>	This agency is a division of the Institute of Education Sciences (IES) in the U.S. Department of Education, implements the policies articulated by NAGB and is responsible for the full production and administration of NAEP. NCES is also responsible for the contractual relationships with the members of the NAEP Alliance and additional contractors (e.g., Hager Sharp, HumRRO, NESSI), and reviews and releases all technical reports generated by members of the NAEP Alliance.
<i>NAEP Alliance</i>	This a term used to describe the organization of contractors selected by NCES whose responsibilities include the development of the test and background questions, creating the assessments, administering and scoring of the assessments, scoring, data analyses, and disseminating results.
<i>Educational Testing Service (ETS)</i>	This Princeton, N.J., organization provides a range of test development, research, and support services in education, admissions, and credentialing; and coordinates the NAEP Alliance contractors, develops test questions for some content areas, creates scale scores, conducts data analyses, and prepares reports of the results.
<i>American Institutes for Research (AIR)</i>	This Washington, D.C. (AIR-DC), and Palo Alto, Calif. (AIR-CA), organization’s offices provide research in education, human development, and health and serve different roles in NAEP. Their D.C. office develops test items or questions for some content areas as well as background questions; their California office conducts state analyses and coordinates the NAEP Validity Studies Panel.

Continues next page

Table 1. Members of the NAEP Consortium and their Roles and Functions (Continued)

Organization	Role and Function
<i>NAEP-Educational Statistics Services Institute (NESSI)</i>	A part of American Institutes for Research, NESSI, formerly known as ESSI, provides technical support services (e.g., item review, report review) for operational components of NAEP.
<i>Pearson Educational Measurement (PEM)</i>	This Iowa City, Iowa, organization is a division of a multinational company that publishes books, develops testing programs, and offers test scoring services. PEM prepares NAEP test booklets for administration, ships test booklets to administration sites, and monitors inventory control of all assessment materials; scores constructed response items; and prepares score records and database for transmittal to ETS for creating scale scores.
<i>ACT, Inc.</i>	This Iowa City, Iowa, organization develops tests and conducts research for a range of admissions, placement, and workforce development programs. One of their tasks within NAEP, under subcontract with NAGB, has been to conduct the standard-setting process for achievement levels. These studies were accomplished for the 12th-grade mathematics assessment in this contract period. ACT is also one of the organizations awarded a contract with NAGB to develop assessment frameworks.
<i>Westat</i>	This Rockville, Md., organization specializes in sampling, surveys, and research methodology, develops the sampling plan for the administration of NAEP and oversees the administrations in the field. Westat also provides a support system for the network of NAEP state coordinators.
<i>Government Micro Resources, Inc. (GMRI)</i>	This Manassas, Va., organization provides information technology solutions and services for a range of government agencies and supports the communication systems for members of the Alliance, including creating and maintaining an information sharing Web site for the Alliance. GMRI also provides technology solutions for the Web-based reports, releases, and tools. The company was acquired in October 2006 by PC Mall Gov.
<i>Hager Sharp</i>	This Washington, D.C., organization specializes in communications for education, government, health, and safety organizations. They serve as an external contractor to NCES to support and enhance the messaging and imaging of the NAEP program.
<i>Human Resources Research Organization (HumRRO)</i>	This Alexandria, Va., organization provides diverse research and evaluation services in education, credentialing, and employment; and serves as an external contractor to NCES to assist with quality control across the NAEP Alliance.

Continues next page

Table 1. Members of the NAEP Consortium and their Roles and Functions (Continued)

<p><i>NAEP State Coordinators</i></p>	<p>These individuals are hired and paid by each state’s department of education to assist with recruitment and administration of NAEP within states and provide guidance to their constituencies on the interpretation and use of NAEP results. These states then contract with NCES to receive funds that pay for the positions and training.</p>
---------------------------------------	--

Both NCES and NAGB, as the agencies primarily responsible for NAEP, have consistently reported that the primary purpose for the program is to measure student achievement and change at the national level. These purposes are found in documents on their respective Web sites:

NAEP has two major goals: to measure student achievement in the context of instructional experiences and to track change in achievement of fourth-, eighth-, and twelfth-graders over time in selected content domains. (<http://nces.ed.gov/nationsreportcard/faq>).

The primary purpose of NAEP is to report to the American public on academic achievement and its change over time. (Background Information Framework for the NAEP, <http://www.nagb.org/pubs/backinfoframew.pdf>).

These goals are broadly stated, leaving their potential definition and operational scope subject to interpretation by policymakers and stakeholders. Flexibility in the interpretation of the program’s purpose has the potential to influence the validity evidence necessary to support those interpretations. Evidence to support valid score interpretations may be collected from both judgmental and empirical sources. Judgmental sources may include recommendations from advisory committees, consensus decisions by representative panels, or position papers from individuals or organizations. Including information from these varied sources reminds us that there cannot be absolute rules for acceptability of procedures or results. Expert judgment is needed to consider the context of multiple interpretations in combination with the other available evidence and to appropriately weight evidence in the decision-making process. Because NAEP is considered by many as the most comprehensive analysis of the condition of education in the United States, it is imperative that the information provided by the program support its intended purposes.

NAEP provides a unique source of information to policymakers about the level and change in the educational achievement of American students at select grade or age levels across a variety of content areas. Broader uses of NAEP assessment data and the validity evidence to support those uses serve as a primary context for the evaluation. Changes in education policy at the national level has increased the visibility of NAEP and requires that we consider current validity evidence if current uses have expanded beyond historical purposes. Because validity is the primary concern for any testing program, we focused our evaluation on the available evidence for the NAEP program. It is also important to note that the characterization of validity itself has also evolved over the course of NAEP’s grammatic history and has changed since the previous evaluation.

The contemporary approach to assessing the validity of tests and assessments is defined and explained in the *Standards for Educational and Psychological Testing* jointly issued by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (AERA, APA, and NCME, 1999). As stated in the *Standards*—

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity, therefore, is the most fundamental consideration in developing and evaluating tests (p. 9).

It is important to emphasize here that validity is not something that a test or assessment has or does not have. Validity is a matter of degree relative to the interpretations made of the test scores. In conducting this audit our focus was on the types and quality of evidence supporting the interpretation of NAEP test scores as defined by the test's intended uses. More broadly, our entire effort can be seen as an evaluation of the validity evidence supporting the intended uses of NAEP scores. Many of the special studies that were conducted within the evaluation have developed new validity evidence. Brief descriptions of these are provided in Appendix C.

Overview of the audit within the full evaluation design

Within the *NAEP Authorization Act of the Educational Science Reform Act of 2002*, Congress mandated this NAEP evaluation and articulated several specific questions to be addressed in the evaluation. (See Appendix B for the text of the legislation):

1. Whether any authorized NAEP assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);
2. Whether NAEP student achievement levels are reasonable, valid, reliable, and informative to the public;
3. Whether any authorized NAEP assessment is being administered as a random sample and is reporting trends in academic achievement in a valid and reliable manner in the subject areas being assessed; and
4. Whether any of the NAEP test questions are biased; and whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.

In creating the final evaluation design, the evaluation team considered the questions posed by Congress, the magnitude of the NAEP program, the available resources for the evaluation, previous NAEP evaluations, and recommendations from the U.S. Department of Education (ED), NCES, NAGB, and the Technical Work Group (TWG) for the evaluation. The full evaluation design is framed by a psychometric audit of the NAEP lifecycle supplemented by special studies designed to examine targeted areas of importance. The audit focused on the technical quality of the NAEP program. The special studies are focused on NAEP achievement levels, consistency in score meaning across various contexts (score equity), the utility of NAEP reports, and methodologies for assessing the alignment of NAEP assessments to state content standards (NAEP-state alignment). Given the available resources and time available for the evaluation, this is by no means a comprehensive evaluation of the NAEP program. Rather, the purpose of the evaluation is to investigate the operations of NAEP with a focus on several identified areas of importance. Within NAEP, the evaluation focused on Reading and Mathematics assessments for the Main and State NAEP programs.

Focus of the audit report

The purpose of this report is to describe the background of the lifecycle audit, the procedures used to collect information, and findings from the different assessment lifecycle dimensions. The four questions that were identified in the congressional mandate served as a foundation for the audit. Further, to more fully address the lifecycle of NAEP, two approaches were identified. First, using the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999) as the guide, key dimensions were identified to direct the various components of the audit. These components were supplemented by also gathering some additional information communications and problem solving among the contractors and agencies. These features are also important to a program like NAEP. Second, a flow chart was developed to aid in the understanding of the developmental path to produce scores in the NAEP assessment program. These two approaches to representing the dimensions of the NAEP assessment program are discussed more fully later in this document.

An Audit of the NAEP Assessment Lifecycle

Within this section of the report there are several parts. Part 1 presents background information on the audit process, documenting both the Buros Center for Testing's history in conducting psychometric audits and the rationale for the structure of the NAEP audit. The outcome of Part 1 is a matrix that characterizes the evaluation team's understanding of the shared responsibilities for NAEP among the agencies and contractors involved in the NAEP Consortium as communicated by NCES, NAGB, and contractors for the program.

Part 2 reviews the procedures that were followed in conducting the audit, including a) the acquisition of documents in preparation for conducting site visits with the various agencies and contractors, b) the communication procedures established for interacting with contracting officers' representatives (CORs) for NCES and each of the contractors and agencies, and c) issues that were considered when scheduling the site visits. A COR serves as the liaison between a government agency and a contractor. His or her responsibilities include monitoring and reviewing the contractor's work, evaluating the contractor's compliance with their contract, and facilitating any communication between the government and that contractor. Within the scope of the evaluation, the audit team interacted with CORs from ED and NCES who were responsible for work conducted by numerous NAEP contractors.

Part 3 describes the procedures for the site visits, including preparatory communications, materials, and processes. In addition, Part 3 contains information about the preparation and vetting of the site visit reports that were generated following each site visit. Parts 4 and 5 provide information about the results and key findings from the audit.

Background information on the audit

The Buros Center for Testing, through its Institute for Assessment Consultation and Outreach (BIACO) division, has an established program for conducting psychometric audits of testing programs and practices. This audit program began in 2000 with the creation of a set of audit standards that were derived from the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999). These audit standards, designed for testing programs that produce noncommercially available tests, served as a starting point for identifying the dimensions (components of the NAEP assessment system) that would be considered in the NAEP lifecycle audit. Although most of the dimensions identified for consideration in the NAEP audit were derived from these audit standards, they were either adapted or augmented to be appropriate for the unique structure of the NAEP program. In addition, discussions with ED, the Technical Working Group (TWG), and other members of the NAEP evaluation team guided the revision and adaptation of the dimensions to be used in conducting the audit.

Through this work, a total of 13 dimensions were identified to serve as the foundation for information gathering about the psychometric quality and integrity of NAEP results. In many ways, these dimensions show a continuous flow of processes and procedures that support the NAEP assessment program. The audit began with a look at a dimensions characterized as the *organizational characteristics* of the NAEP assessment program and the contractors and agencies that have vital roles in NAEP. The purpose of starting the audit with a consideration of the organizational characteristics is to ensure the organizational structure and capacity is consistent with good test development practices. Within this dimension, factors such as staff qualifications, internal and external systems for communications, clarity of roles and responsibilities both within and across agencies, and attention to problem documentation and resolution were key elements. The NAEP assessment program encompasses a complex network of contractual relationships. For this reason, it was viewed as important to consider organizational features that would contribute to or

work against the smooth and effective transitions necessary to complete a NAEP assessment from conceptualization to reporting to renewal.

The majority of the remaining dimensions for the audit follow the normal sequence of processing for an assessment program. First and foremost, the purpose(s) for the assessment must be clearly stated. In the audit this dimension is phrased as *defining intended uses of NAEP assessments*. Addressed in this dimension of the audit are elements such as the identification of a validity framework, various sources of evidence to support the validity of uses of NAEP test results, and the need for clearly articulated statements of intended and unintended interpretations and uses of NAEP test results.

Following in the typical test development sequence, the next dimension addresses the procedures used to *developing NAEP assessment frameworks*. The assessment framework sets out the components of the domain that will be measured through the assessment. The basis or foundation for these components of the domain need to be clearly identified and the relevant knowledge, skills, and processing levels should be articulated in the framework. Documentation is needed to support both the decisions made about the domain components to include and their relative weightings in the assessments and the processes followed to ensure that these decisions have support in the respective professional community. This assessment framework guides the development of test questions.

Once the assessment frameworks have been agreed upon through the review and vetting process, test developers begin the task of preparing test questions that will be used in comprising the test. In this stage, *developing test items (questions) and background questions*, information is gathered about the procedures used to select or commission item writers (including their qualifications), the training that these item writers are given, and the criteria that are applied to the developed items to ensure content accuracy and technical quality. Because the NAEP assessment has both cognitive and background questions, this dimension applies to both of these parts of the NAEP assessment.

As part of the verification of the content accuracy and technical quality of test questions, the next step in the test development process is to prepare for and conduct field trials of the items. The next dimension in the audit is directed at evaluation of this step in the test development process; *creating draft assessments, preparing field test designs, and conducting field trials*.

Once the assessment has been revised based on the results of the field tests, the next steps involve several processes including (1) using field test data to *set achievement levels* for NAEP, (2) using field test data to *constructing final assessments*, and (3) *sampling schools and students*. Each of these processes serves as separate dimensions in the audit design.

The next dimension, *administering NAEP assessments*, is the culmination of many processes. Administrators needed to be identified and trained and therefore training materials and procedures need to be developed and disseminated. Procedures need to be established for administering the assessment with accommodations for students with special needs. Mechanisms need to be put into place to control the flow of materials to and from the field. All of these features are considered in the audit for this dimension.

After the assessments are administered and returned for processing, scoring of student answers occurs. Because the NAEP assessments are composed of selected-response items (multiple choice) and constructed-response items, different scoring activities must be undertaken. Complex inventory control must be in place to ensure proper receipt of student booklets and proper process monitoring to ensure student response records are maintained across the different scoring processes. The outcome of this step is a response record for each individual taking the NAEP assessment; this response record is then stored in a database that is used to create the NAEP scales and linkages across NAEP assessments. In the audit dimension, the scoring step is called *scoring NAEP assessments*.

Following on the heels of the scoring and database preparation is the creation of scale scores and the links that serve to connect the assessment across time in order to examine score changes and to monitor the long-term trend. It is in this audit dimension, *creating scales and links and analyzing data*, that much of the traditional psychometric quality data for reliability and item analysis information is gathered. Studies of differential item functioning, or the tendency for an item to show different characteristics when administered to equally able students with differing demographics (such as ethnicity or gender) are conducted.

Once the results have been completed and verified through the complex processes used for scoring, scaling, and linking, the results are ready for dissemination to the public and policymakers. Much attention is given in the NAEP assessment program to provide test results that are useful and readily available to interested users. This dimension is called *writing, reviewing, and disseminating reports and data*.

Because NAEP is a long-standing and evolving assessment program, the next dimension draws attention to the need for continual monitoring and efforts to *improve NAEP assessments*. The focus of this dimension is on looking both backward, to ensure that the documentation needed to support decisions about the program are in place, and forward to enable the program to stay vital and ever growing as the science of assessment and the uses of NAEP results evolve.

More information on the focus of the final audit design, including details of each of the 13 dimensions is included in Table 2. Following the development of these dimensions, a decision was made about which of these dimensions were relevant for the various agencies and contractors in the NAEP consortium. Again, these decisions were informed through discussions with ED, the TWG, and other members of the NAEP evaluation team. As we learned more about the NAEP program, we updated our dimensions to reflect the unique characteristics of this program. The dimensions and agencies and contractors were then organized into a matrix that crossed the dimensions with perceived responsibilities of the members of the NAEP consortium. When an agency or contractor had a primary role for a particular dimension, an asterisk was indicated. The preliminary list of dimensions and the responsibility matrix were used in planning for and conducting the audit. During the audit, we made revisions to the preliminary responsibility matrix as the audit team members learned more about the roles and responsibilities of the members of the NAEP consortium. The final NAEP responsibilities matrix is shown in Appendix E.

Table 2. Lifecycle audit dimensions and sources of evidence

<p>1. Organizational Characteristics [NCES, NAGB, ETS, Westat, AIR, PEM, HumRRO, Hager Sharp, GMRI, ACT, state coordinators]</p> <ul style="list-style-type: none"> • Qualifications of staff • Structure of organization • Communications <ul style="list-style-type: none"> ○ Within staff ○ Among contractors • Mechanisms for problem identification and resolution • Clarity of roles and functions • Deadlines/timelines • Potential conflicts of interest with other programs and/or products within the organization • Security procedures

Continues next page

Table 2. Lifecycle audit dimensions and sources of evidence (Continued)

<p>2. Defining intended uses of NAEP assessments [NCES, NAGB, ETS, AIR-CA, PEM, state coordinators]</p> <ul style="list-style-type: none"> • Validity framework for gathering supporting evidence for intended uses • Technical reports from contractors • Connections between validation efforts and intended uses of scores • Clear articulation of unintended/inappropriate score uses and interpretations
<p>3. Developing NAEP assessment frameworks [NAGB]</p> <ul style="list-style-type: none"> • Procedures for framework development <ul style="list-style-type: none"> ○ Framework design ○ Identification of subject matter experts ○ Timeline for development ○ Review process
<p>4. Developing test items (questions) and background questions [NCES, NAGB, ETS, AIR-DC]</p> <ul style="list-style-type: none"> • Item writing procedures • Security procedures for item development • Selection and training of item writers • Procedures for review and revisions • Documentation of item development • Documentation of item banks <ul style="list-style-type: none"> ○ Inventory ○ Prioritization for item development • Schedule for new item development • Examination of background questions <ul style="list-style-type: none"> ○ Inventory ○ Alignment to policy and data analysis needs
<p>5. Creating draft assessments, preparing field test designs, and conducting field trials [ETS, Westat]</p> <ul style="list-style-type: none"> • Strategies used to pilot new test items/tasks • Security procedures for conducting field trials • Form assembly; number of items per test; length of sections • Logistics for pilot administration <ul style="list-style-type: none"> ○ Administrative procedures ○ How and when administered ○ Criteria for site selection ○ Administrator manual ○ Quality control/audit of administration procedures ○ Examinee accommodation procedures • Scoring procedures (preparation of rubrics; piloting of scoring) • Analyses/criteria for evaluating pilot results and actions taken <ul style="list-style-type: none"> a. Item revision/deletion b. Item calibration

Continues next page

Table 2. Lifecycle audit dimensions and sources of evidence (Continued)

<p>6. Setting achievement levels [NAGB, ACT]</p> <ul style="list-style-type: none"> • Rationale for standard setting procedure • Identification of panelists • Procedures used for setting achievement levels • Use of feedback data (internal/external) • Procedural validity evidence • Internal validity evidence (e.g., consistency estimates) • External validity evidence (e.g., comparisons to TIMSS, PISA) • Security procedures for standard setting activities
<p>7. Constructing final assessments (content, design, and production) [PEM, ETS]</p> <ul style="list-style-type: none"> • Form assembly; number of items per test; length of sections <ul style="list-style-type: none"> ○ Distribution of content areas across sample ○ Strategies for weighting of items • Number of alternative forms • Content distribution across forms (e.g., matrix sampling) • Specifications and quality control for printing • Specifications for packaging, spiraling, and distribution • Security procedures for handling and storage of assessment materials
<p>8. Sampling Schools and Students [ETS, Westat]</p> <ul style="list-style-type: none"> • Sampling design <ul style="list-style-type: none"> ○ Sufficiency for Main and State NAEP scores ○ Strategies for weighting of individuals ○ Representation of sub-populations • Results <ul style="list-style-type: none"> ○ Response/participation rates overall and by groups ○ School and student replacement rates ○ Quality indicators for population estimates (distributions and standard errors of total and groups) ○ Imputation for missing data ○ Representation of school districts and schools within districts sampled
<p>9. Administering NAEP assessments [Westat, PEM, state coordinators]</p> <ul style="list-style-type: none"> • Selection and training of test administrators/monitors • Rates of exclusions, ineligibles, accommodations, and exceptions • Logistics for Administration <ul style="list-style-type: none"> ○ Administrative procedures ○ How and when administered ○ Administrator manual ○ Quality control/audit of administration procedures ○ Examinee accommodation procedures • Security procedures for administration of the assessment

Continues next page

Table 2. Lifecycle audit dimensions and sources of evidence (Continued)

<p>10. Scoring NAEP assessments [ETS, PEM]</p> <ul style="list-style-type: none"> • Selection, training, and quality of scorers • Quality of scoring (e.g., inter-rater consistency, quality checks) • Monitor and quality control of data entry and machine scoring • Collection and storing of examinee data • Security procedures for collection and storage of examinee data
<p>11. Creating scales and links and analyzing data [ETS, AIR-CA, HumRRO]</p> <ul style="list-style-type: none"> • Equivalence of score meaning/equating strategies <ul style="list-style-type: none"> ○ Evidence of content equivalence ○ Equating procedures and results • Psychometric Properties <ul style="list-style-type: none"> ○ Reliability/precision <ul style="list-style-type: none"> ⇒ Scores (Main and State NAEP) ⇒ Decision consistency ⇒ Standard errors at cutpoints ⇒ Diagnostic sections/subscales/strands ⇒ Information functions ⇒ Standard errors ○ Item analyses <ul style="list-style-type: none"> ⇒ Procedures for item analyses ⇒ Summary statistics, distributions of item parameters ⇒ DIF analyses ⇒ Item exposure/scale drift analyses ○ Scoring <ul style="list-style-type: none"> ⇒ Missing data/omit rates ⇒ Procedures for estimating item and post-stratified student weights
<p>12. Writing, reviewing and disseminating reports and data [NCES, NAGB, ETS, Hager Sharp, GMRI, state coordinators]</p> <ul style="list-style-type: none"> • Report development process • Stakeholder appropriateness/utility • Distribution to appropriate audiences • Procedures for timely reporting of results • Use of appropriate data • Web site evaluation <ul style="list-style-type: none"> ○ Hit rates ○ Record of downloads ○ Responsiveness to requests ○ Quality of interactive tools ○ Satisfaction of consumers/stakeholders

Continues next page

Table 2. Lifecycle audit dimensions and sources of evidence (Continued)

13. Improving NAEP assessments [NCES, NAGB, ETS, AIR-CA, HumRRO, PEM, Westat, state coordinators] <ul style="list-style-type: none">• Use of quality control team members /advisory groups• Technical reports from contractors• Innovativeness of procedures used• Quality control results

As another mechanism for presenting the processes that comprise the NAEP assessment program, a flow diagram was constructed that shows the path of procedures for the creation and reporting a NAEP test results, called the “The Path to a NAEP Score” (Figure 2). Many similarities can be seen in the developmental flow shown in this diagram and the sequence of test development events that are characterized by the audit dimensions. Each component in the diagram represents a specific activity in the NAEP program along with notation of the responsible organizations.

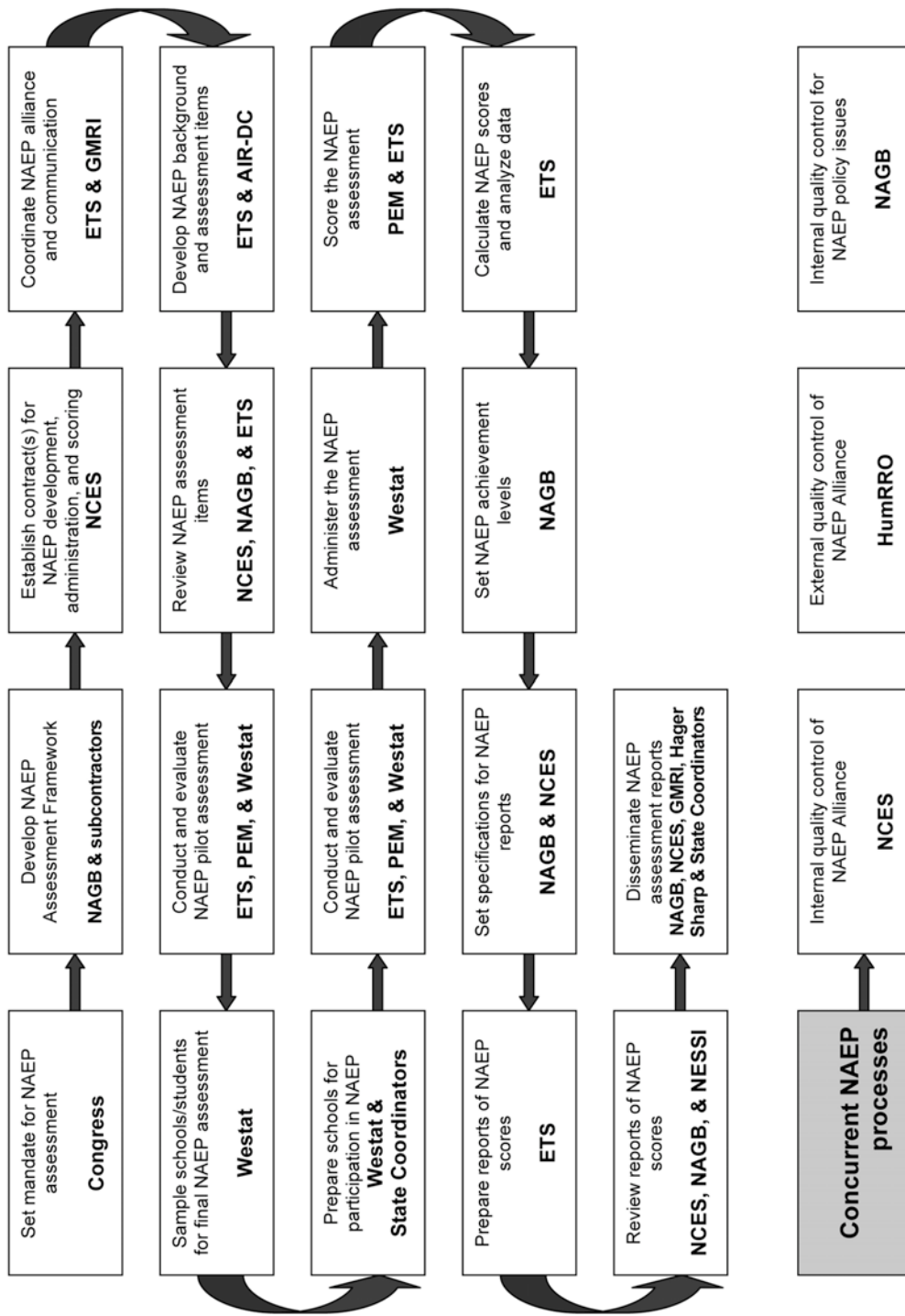


Figure 2 – The Path to a NAEP Score

Audit processes and procedures

Following the development of the preliminary list of audit dimensions and the responsibility matrix, efforts began to obtain documents that would provide relevant information about how the various agencies and contractors were fulfilling their responsibilities in the NAEP program. A broad-based approach was used to gather these documents. The first step in this process included preliminary meetings with NCES and NAGB to ascertain the scope of available information. The evaluation team then conducted independent Web searches and literature reviews. After a preliminary review of NAEP materials, we requested specific relevant information from each of the agencies and contractors in the NAEP consortium. As agreed upon in planning meetings, these requests were directed to the CORs for the respective agency or contractor. This was accomplished by sending the CORs an initial e-mail communication with a general overview of the audit process, the full audit dimensions and responsibility matrix, and information specific to that agency or contractor about which of the audit dimensions had been identified as relevant to their work. An example of this initial e-mail communication is shown in Appendix F. This process continued with follow-up e-mails, telephone conversations, and intervention by CORs to streamline acquisition when necessary. Even with this broad-based approach, we experienced some difficulties in obtaining materials to inform the audit based on at least three factors: characteristics of the evaluation team, characteristics of the agencies and organizations involved in the NAEP consortium, and the current reporting requirements.

The nature of this evaluation, as defined by the congressional mandate, dictated the evaluators of NAEP needed to be independent of the system. Therefore, the contractors selected to conduct the evaluation were only somewhat familiar with the NAEP system and very unfamiliar with the interworkings of the NAEP consortium. In addition, the evaluation team was unfamiliar with the procurement process used to obtain many of the documents produced within the NAEP consortium. As a result, many of the requests for documentation were broad, leaving the COR and the respective contractors seemingly unsure in some instances about what would constitute appropriate evidence that supported their work for each of the relevant dimensions. A common frustration experienced by the evaluation team was an expectation by NAGB, NCES, and some contractors that requests for materials specify the title of the document or report. However, in many instances the evaluators were unable to specify this information because it was not possible for them to know in advance an exact title of a report or document that had not been released. To our knowledge, most of the materials requested were eventually received from NAGB, NCES, and contractors; however, our findings are limited to materials to which we had access during the evaluation. We were also challenged in our efforts to obtain statements of work that were to guide the activities of the contractors. This made it difficult to evaluate the actions and products of a contractor when we, given our limited knowledge of the NAEP system, were not initially aware of all of the activities these contractors had committed to accomplish in their contract agreements. It is important to note that this was not the case for all agencies and contractors but contributed to increased search and acquisition efforts on the NAEP evaluation teams' part and led to several frustrated discussions with NCES, CORs, and ED.

Also contributing to this frustration were the characteristics of the NAEP consortium. First, because many of the documents and reports that were requested were still in the review process, NCES policies about document release prohibited or restricted access to many documents that had been submitted to NCES but had not yet been completely through the review process or released. Second, difficulties in obtaining materials could also be attributed to confusion on the part of some of these agencies and contractors about what the evaluators were asking them to provide. Through in-person, electronic, and telephone communications with NCES and the CORs, we attempted to explain the role of the documentation in the audit process and the kinds of evidence we were seeking to obtain.

A third factor that appeared to contribute to the difficulty in obtaining materials was the six-month timeline for the NAEP consortium to disseminate results. This timeline required the consortium

to prioritize operational responsibilities ahead of the requests of the independent, external evaluators. Whenever possible, we attempted to be accommodating to the workloads and priorities of the NAEP consortium staff; however, we were also restricted by our timeline for the audit.

Once documents were received they were cataloged and reviewed by members of the audit evaluation team. Summaries of the documents were prepared and reviewed by the audit team leaders to inform decisions about the accuracy of the responsibility matrix and the prioritization of site visits with the respective agencies and contractors. Based on this review of documents, the following agencies and contractors were identified for site visits: NCES, NAGB, AIR (Washington, D.C. and Palo Alto, Calif. offices), HumRRO, Westat, and ETS. These prioritized decisions about which agencies and contractors to visit were shared with the TWG at their August 2005 meeting. At that meeting, the TWG encouraged the evaluation team to add, if feasible, at least PEM and GMRI to the list of contractors who were visited for the audit. Moreover, they recommended that members of the audit team observe upcoming NAEP related meetings that would allow access to NAEP state coordinators and allow for some evidence gathering about the work by Hager Sharp. In response to the TWG's request, the evaluation team added site visits to PEM and GMRI and also had an opportunity to observe activities conducted by Hager Sharp.

Conducting site visits

Once a decision was made about which agencies and contractors to visit, communications were initiated with the relevant COR. In that communication, efforts were made to identify dates and times that would be most acceptable to the agency and contractor based on their respective responsibilities and deliverables in the NAEP process, for the site visit. Because the evaluation team recognized the six-month timeline for reporting NAEP results, every effort was made to reduce the potential intrusiveness of the site visits. Negotiations were sometimes made directly with personnel at the agency or contractor, but only after access and contact information was given by the COR. Through these discussions, dates were set well in advance of the site visit to allow for the agency or contractor time to prepare for the site visit. Agendas were negotiated in advance of the meeting and in some cases additional materials were sent to the evaluators as background information for the site visit. In every case, at least two members of the audit team met with staff from the agency or contractor. In some cases, the COR or contract officers from ED also attended, but this did not occur often.

Most site visits were either one or two full days. At the site visits, staff from the agency or contractor made presentations or led discussions related to their respective roles and responsibilities regarding the NAEP program. Members of the audit team asked questions to clarify information provided and, in most cases, asked that additional documents be sent for review. Following the site visit, a draft report of the findings and recommendations from the site visit was prepared by the audit team leader and shared with all members of the audit team for input. Once the draft report was reviewed and edited by the audit team members, it was sent to the respective agency or contractor with a request for a review for factual accuracy. The agency or contractor was typically given two weeks to review the draft report. Most agencies or contractors provided comments or feedback on the draft site visit report. Following the review for factual accuracy and once any needed revisions were completed, the audit report was distributed to the agency or contractor, the agency's or contractor's COR, and senior staff at NCES. For NAGB, the draft site visit report was shared with the Executive Director and Deputy Executive Director of NAGB. The timeline in Appendix F documents the sequence of the site visits that were conducted—identifying the dates for the visit and the members of the audit team who participated in the site visit.

Results from the lifecycle audit

In the following sections, results of the lifecycle audit are presented. These results are organized by audit dimension. For each audit dimension, we first provide a brief overview of the key elements of the dimension. Next, relevant standards are presented from the *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 1999) providing another context for interpreting the meaning of the audit dimension under consideration. Once this orientation to the dimension is complete, the specific components of the audit dimension are noted, followed by an identification of the NAEP organizations, agencies, and contractors whose work contributed to the fulfillment of that dimension. The audit dimensions are organized into subsections by specific components of the NAEP system including:

Program Management

1. Organizational Characteristics: Communications and problem resolution

Developing NAEP assessments

2. Defining intended uses of NAEP assessments
3. Developing assessment framework
4. Developing test items (questions) and background questions
5. Creating draft assessments, preparing field test designs, and conducting field trials

Collecting data on NAEP assessments

6. Constructing final NAEP assessments
7. Sampling schools and students
8. Administering NAEP assessments

Scoring and analyzing NAEP assessment data

9. Scoring NAEP assessments
10. Creating scales and links and analyzing data

Interpreting and using NAEP assessment scores

11. Writing, reviewing, and disseminating reports and data
12. Setting achievement levels

Improving NAEP assessments

13. Improving NAEP assessments

This page intentionally left blank

Program Management

Organizational Characteristics

Communication

To achieve all of the steps identified above in the test development, administration, and reporting procedures, attention needs to be paid to the communications among the multiple agencies and contractors in the NAEP consortium. This is especially important for a testing program that involves several contractors who rely on the others for the testing program parts. As seen in Figure 2, several of the steps in the NAEP assessment system involve multiple contractors and agencies. Therefore, communication among organizations involved in this system is vital for the successful completion of each step and the progression of one step to the next.

This section differs somewhat from the others in terms of how it relates to the *Standards*. The critical elements here are unique to the NAEP program considering the magnitude of the program and the number of persons involved in the NAEP system. Although not directly related to any testing standards, the communication structure ensures the separate components of the process (described above) work well together and the process as a whole runs smoothly. Given these factors, it is critical that the systems for communication are very clear and organized as communication is necessary for problem resolution within the system. Systems that support communications were examined both within and among contractors and agencies in the NAEP consortium. In addition, mechanisms that were in place to identify and resolve problems were considered. All members of the NAEP Consortium were included in this audit dimension. In this section, we will discuss the communication structure among the NAEP organizations as well as problem identification and resolution. As some of the communication elements are key parts of the NAEP cycle, other details about the communication system within difference contractors will be discussed in later sections of this report.

The NAEP system, as illustrated in Figures 1 and 2, and Table 1, involves a number of organizations and hundreds of individuals to implement the program. Referring back to Figure 1, although this is an oversimplification of the communication lines within the NAEP system, it is useful in explaining the system for the purposes of this review. On the right side of Figure 1 is NAGB, which represents the policy component of NAEP. NAGB oversees contractors for certain functions such as setting achievement levels (ACT), and these contracts are managed and run by NAGB staff. Each NAGB staff member takes responsibility for contracts that are within the purview of their respective subcommittee. NAGB is in direct communication with NCES, the operations side of NAEP. NCES and NAGB hold two joint meetings prior to each NAGB Board meeting to discuss the meeting agenda and materials needed. The first is six months prior and the second is approximately three weeks prior. NCES staff members also attend NAGB meetings.

NCES, as the head of the operations side of NAEP, is in direct communication with all the contractors involved in the NAEP operations. The majority of these contractors are within the NAEP Alliance. This includes ETS, AIR-DC, GMRI, Hager Sharp, PEM and Westat. The Alliance contractors use the Integrated Management System (IMS) for virtual discussions, sharing of materials, and review of materials. The IMS system appears to offer NCES the ability to monitor discussion and work among the contractors within the Alliance. Each of the Alliance contractors also has a COR within the assessment division at NCES. The NCES CORs are in contact almost daily with their respective contractors for a variety of purposes (contractors must consult NCES before making any major design decisions). More formal teleconferences between the contractors and NCES are held approximately every two weeks.

ETS has the responsibility for coordinating the NAEP Alliance. Prior to the most recent contract procurement model, ETS was the prime contractor for the NAEP assessment but worked with other principal contractors (except Westat) through subcontracts. Under the current model, members of the

Alliance have separate contracts with NCES; their work is coordinated through NCES (which oversees all the contractors directly), and ETS which has a separate contract with NCES for Alliance coordination. ETS sees its Alliance coordinator role as one of “air traffic controller,” ensuring that the project stays on the “critical path,” toward fulfilling overall NAEP outcomes and expectations (especially the six-month reporting timeline for reading and mathematics assessment results). In addition, their role is as a conduit to ensure that potential problems are brought to the attention of NCES and to focus the Alliance on quality control improvements (which overlaps somewhat with the external roles and responsibilities of HumRRO).

ETS accomplishes its Alliance coordination responsibilities through a variety of communication strategies, including regular meetings with contractors, holding an annual NAEP Design Summit, conducting regular conference calls with Alliance partners and NCES, and the use of the IMS that allows for easy sharing of documents between contractors. The IMS also has varying levels of accessibility depending on the sensitivity of the material that is posted; it permits posting of logs of problems with documentation of resolutions. ETS has found that serving as the Alliance coordinator has its challenges because the company has no real authority over the contractors, but it is held accountable for ensuring compliance across the Alliance partners and that NAEP goals are achieved. Strategies used to coordinate functioning of the Alliance have been dynamic over the years of the contract, with changes made in response to experience with communication procedures and recommendations by Alliance members.

There are several components of the contractual agreements with the contractors that enhance effective communication, including bonuses paid for meeting key deadlines. Some of the contractors indicated there was a “one for all and all for one” spirit maintained in the Alliance for cooperative efforts. The communication system also allows members of the Alliance to work together in responding to changes in policy or procedure. For example, a decision was made by the Office of Management and Budget (OMB) regarding reporting categories for race and ethnicity and a problem with allowing the surrogate socioeconomic status (SES) variable (free or reduced-price lunch status) to be considered as a school-level instead of a student-specific variable. Both of these changes could have serious ramifications across Alliance partner roles. Specifically, information about SES is currently used by ETS in the conditioning process for scaling, Westat uses SES in designing its sampling plans, AIR-DC includes SES questions in its background questions, and SES is used as a reporting variable of NAEP results. NAGB asked for advice on this issue and major contractors from the NAEP Alliance met to discuss the possible ramifications of this change. Special studies were designed and through a recent NAGB decision, study designs are being further developed.

In addition to the NAEP Alliance, NCES works with additional contractors to complete the NAEP process. This includes the NAEP state coordinators, HumRRO, and AIR-CA. The NAEP state coordinators communicate with NCES when needed and also through scheduled meetings and trainings via the internet. NCES uses commercial software for Web conferencing with NAEP state coordinators; each week, there are three training sessions that state coordinators can attend. These are recorded and can be re-played at a later date. In addition, the NAEP state coordinators are brought together twice a year for group meetings. Communications are also fostered through HumRRO’s role in various meetings, including attendance and preparation of NCES-specific notes. These meetings include NAGB, NAEP Validity Studies Panel, the Design and Analysis Committee (DAC), and NCES/Contractor (including Annual Design Summit).

As shown in Figure 2, all NAEP contractors within the Alliance are important to the operational components of the NAEP program. Therefore, there is also a communication path among contractors within the Alliance and those outside the Alliance. For example, there appears to be some interaction with HumRRO and ETS regarding efforts to renew and improve NAEP. In addition, Westat operates a support center for the NAEP state coordinators. This effort began as a broader vision to have people in the states help recruit schools for participation, communicate NAEP information, conduct state data

analyses, and write or disseminate reports. Although the NAEP state coordinators are employees of their respective state departments of education with funds from NCES, they are supported for their activities through this contract with Westat. Westat provides professional development/training workshops on relevant topics, some of which are requested by the NAEP state coordinators. Many of these training sessions are offered via online meeting software to help control costs for participation. Another key element of the State Support Center is a secure Web site (NAEP Network) that serves as a link between the states and operations.

Another communication line is the path between NAGB and the NAEP Alliance. As the Alliance is under contract with NCES, NCES is responsible for communications between the contractors and NAGB staff. However, it was noted during site visits with several of the contractors that NAGB does occasionally directly contact the NAEP Alliance contractors. These communications can lead to some confusion if there are contradictions between direction suggested by NCES and NAGB to individual contractors.

Problem Identification and Resolution

Communication structures are important to any large organization and are vital specifically for problem identification and resolution. With a program of this size, problems are inevitable within organizations as well as among them. Our review of this area focused on issues that arose between organizations and the mechanisms that were in place to identify and resolve such issues.

Our review of problems and issues began with the management and technical review of NAEP conducted by KMPG in 1996. The KMPG study (1996) indicated that NAGB occasionally infringed on the operational side of NAEP. By legislation, NAGB is responsible for oversight of policy. However, because some NAGB staff members have psychometric expertise, there are instances when NAGB becomes involved in the operations of NAEP when these responsibilities go beyond their scope. This can be viewed as particularly problematic when decisions made by NAGB board members who may not be qualified to render such judgments override the decisions of content or measurement specialists. In turn, decisions or policies made by NAGB in these instances often overlap with existing NCES policies. Given the increased importance of NAEP and the additional responsibilities of each organization, overlap and differences of opinion in interpreting NAGB's and NCES's responsibilities are inevitable. For example, NAGB established a policy for participation rates when policies already existed for these data in NCES's Statistical Standards. Other examples included the specifics mandated by NAGB for the execution of the fall pilot study and requests for projects or changes to frameworks that are outside the bounds of NAEP's limited budget (e.g., addition of a vocabulary scale to reading, foreign language assessment). Through the audit site visits, we also found evidence of how changes in legislation have created tension and confusion between NCES and NAGB. Recent legislation (P.L. 107-279) has changed the policy for preparation and dissemination of NAEP reports. The new legislation appears to expand NAGB's role into areas that were historically within the purview of NCES leading to some confusion about roles and responsibilities. This confusion has likely caused some differences of opinion between NCES and NAGB regarding the interpretation of this legislation.

The outcomes of these tensions between NAGB and NCES are different interpretations as to how the results can and should be used, some duplication of efforts, and in some instances disagreements over responsibilities. Reduction of these tensions would most likely result in better communications between these organizations and more effective functioning of both.

We also noted concerns about communications between NAGB and organizations within the NAEP Alliance. In the previous section we alluded to tensions that can arise as a result of NAGB making requests of Alliance contractors that are under contract with NCES. In addition, policy decisions made by NAGB sometimes create problems with timelines and procedures of Alliance projects. For example, with the arts assessments, delays in making decisions about new item development and the

possible inclusion of performance assessments created some pressures within AIR-DC's item development efforts. Further, NAGB's decisions had implications for the configuration of blocks for assessment design and administration, which impacted other Alliance contractors. Although it is clear that there are communications between NCES and NAGB staff members regarding implications of NAGB policy decisions, and instances when NAGB has sought advice from NCES about pending NAGB policy decisions, these policy decisions nonetheless seem to put stress on Alliance partners in their ability to comply with their expected roles and functions.

Finally, the cooperative design of the Alliance contract has likely contributed to successful completion of many NAEP projects, but also some issues. The NAEP Alliance contract has made it difficult to adhere to an agreed upon schedule among the contractors because there are a number of dependent components that require certain activities to occur before others. If there is a delay in one of these activities, it automatically challenges subsequent activities to meet original timelines. For example, delays in the Common Core of Data (CCD) pushed the 2006 sampling activities two months later than is typical. Although it is beyond the control of Westat, it has the potential to impact how quickly data can be handed off to PEM to create the shipping materials needed for the administration.

These three examples of types of problems were those that were identified during the audit review. There are likely other sources and types of problems or issues that arise in many areas of the NAEP program. Such problems are inevitable whenever there are so many moving parts in a system such as NAEP and so many agencies or organizations involved in the process. The important aspect here is that there are systems in place to identify and resolve such problems through communication. Within the NAEP system there are several such systems. Such processes help to establish an environment that supports good quality control procedures and has the potential to be proactive in identification of potential problems and facilitate early resolution.

As one source of external quality control for NCES, HumRRO facilitates two specific communication forums for problem identification and resolution known as the Quality Assurance Council (QAC) and the Quality Control Team (QCT). These groups were formed in December 2003 in response to identified needs to enhance cross-Alliance communications regarding quality control issues. The QAC consists of representative from NCES, the NAEP Alliance, and HumRRO. The purpose of QAC is to facilitate the discussion of quality matters, develop broad quality control policies and standards, and to promote a cross-organizational atmosphere. The QCT also consists of representatives from each of the Alliance members and HumRRO. This team implements standards and policies articulated by QAC; coordinates quality control activities across the Alliance; develops tools and methods to address quality control issues; and informs QAC of critical quality control issues. The QAC meets quarterly and the QCT holds biweekly conference calls. There is a mechanism for documenting issues identified through these communications on a secure private Web site that is only accessible to QAC and QCT members. NCES does not have access to this Web site because it was decided that this arrangement would support free and open discussion of problems and issues. HumRRO maintains minutes of these meetings and all issues are logged in the Process Improvement Log (PIL). Unresolved issues remain open on the PIL until resolution is obtained.

In addition, HumRRO's responsibilities include two other roles that offer problem prevention. First, HumRRO conducted interviews with Alliance members and others to document problems that occurred in the past and identify how these problems either were resolved or what steps should be taken to ensure they would not recur (the HumRRO *Past Problems* report). Second, each contractor in the Alliance prepares a Quality Control (QC) plan on an annual basis. These QC plans are reviewed by HumRRO to ensure that appropriate QC plans and documentation are in place.

In addition to the services facilitated by HumRRO, there are regularly scheduled meetings between NAGB and NCES; and NCES and the contractors as described in the communication section. When a difference of opinion arises between NCES staff and NAGB staff, the issue is first discussed between the two organizations. If this does not produce a viable solution or resolve the issue, assistance

may be sought from the NCES commissioner.

Conclusions: Organizational Characteristics

Given the multiple organizations involved in the NAEP Consortium, it is important that a communication and quality control infrastructure support the ongoing activities of the program. We observed that such an infrastructure has been created, facilitated by technology innovations to support communications and quality assurance. The multiple communication systems within the NAEP program help facilitate organization in the system as well as problem identification and resolution. This information about problems that occur and solutions to such problems would be utilized better if there were a feedback loop of information gained through the examination of the Quality Control plans, recommendations from the site visits, and the QCT problem identification logs. Such information could be used for continuous system improvement. One improvement, though, is the need for a better communications flow from NAGB through NCES to Alliance contractors—this might enhance a mutual understanding of how some policy decisions affect operational timelines and personnel resources.

Finally, there continue to be differences of opinion regarding the roles and responsibilities of NAGB and NCES. This results in part from the clarity of the legislation but also from the differential interpretation of the NAEP legislation (modified in 2002). One possible option to help to resolve these disagreements would be to seek clarification from Congress on these issues.

This page intentionally left blank

Developing NAEP Assessments

Defining intended uses of NAEP assessments

The first and most important step in the sequence of events for any assessment development effort is the definition of the specific, intended purpose(s) or uses of the results. An assessment itself is neither valid nor invalid; the degree of its validity can only be examined in light of the intended uses and interpretations of the results. Therefore, it is critical that the intended purposes of NAEP results be specifically identified and that guidance be provided for gathering evidence to support the validity of the scores for these uses. To aid stakeholders in understanding the appropriate and intended uses of NAEP test results, it is also desirable to anticipate and identify inappropriate and unintended uses and interpretations of NAEP results.

These are the relevant professional *Standards* (AERA, APA, and NCME, 1999):

Standard 1.1: A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use of interpretation.

Standard 1.2: The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described.

Standard 1.3: If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.

Standard 1.4: If a test is used in a way that has not been validated, it is incumbent on the user to justify the use, collecting new evidence if necessary.

Standard 1.24: When unintended consequences results from test use, an attempt should be made to investigate whether such consequences arise from the test's insensitivity to characteristics other than those it is intended to assess or the test's failure to fully represent the intended construct.

Standard 6.3: The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standard 15.1: When the same test is designed or used to serve multiple purposes, evidence of technical quality for each purpose should be provided.

For this dimension of the audit, information was sought on several topics. First, evidence was sought regarding the intended purpose(s) of NAEP assessments and the intended interpretations of test scores. As noted in the *Standards*, an important component to this aspect of test development is clear articulation of unintended and inappropriate uses of NAEP results. Second, information was gathered about the validation efforts by the contractors to support the intended uses of the scores. This evidence could come from research studies initiated by the contractors and in technical reports and documents

prepared as deliverables by the contractors. Involved in this dimension of the NAEP assessment program are NCES, NAGB, ETS, AIR-CA, PEM, and the NAEP State Coordinators.

Intended uses

As noted in the introduction, both NAGB and NCES provide statements regarding the intended use interpretation of NAEP scores on their Web sites. These specifications of purpose come from the legislation mandating the assessment and the scope of the NAEP program. As the policy body overseeing NAEP, NAGB's job is to provide information for the development of public policy and to implement established policies but not to create public policy. Therefore, with respect to defining the intended uses of NAEP assessments, NAGB is responsible for interpreting the legislation. For example, NAGB was recently given the responsibility of releasing NAEP results. From P.L. 107-279, NAGB's duties include:

7. Develop guidelines for reporting and disseminating results; 9. Take appropriate actions needed to improve the form, content, use, and reporting of results of any assessment authorized by section 303 consistent with the provisions of this section and section 303; and 10. Plan and execute the initial public release of National Assessment of Educational Progress reports (Section 302, 5).

NAGB's responsibility in this situation is to articulate how NAEP data should and should not be reported. NAGB avoids telling states directly how to interpret NAEP results in relation to state test data; however, states are free to make their own comparisons. NAGB's responsibility is to ensure that NAEP reports include caveats that such comparisons are difficult to make because NAEP is a survey (not a census) testing program and the NAEP assessment frameworks are built differently than the state frameworks, often characterized as content and process standards.

With the *No Child Left Behind (NCLB, 2002)* legislation, there has been increased interest in NAEP assessment frameworks across the country. The Board cannot advocate use of the NAEP frameworks by states (NAGB, 2002d); however, they make the frameworks available to any states that request them. In the introduction to the current NAEP Mathematics Framework (NAGB, 2004d), it states:

Of critical importance is the fact that this document does not attempt to answer the question: What mathematics should be taught (or how)? This is an assessment framework, not a curriculum framework. It was developed with the understanding that some concepts, skills, and activities in school mathematics are not suitable to be assessed on NAEP, even though they may be important components of a school curriculum. ¶3

In this sense, because these assessment frameworks may not align with curricula at state or local levels, NAGB has to react to how states' might use their assessment frameworks to comply with their mission and scope of work. States have demonstrated varying levels of using NAEP in their state assessment and accountability systems. NAGB is also responsible for initiating efforts to expand the scope of NAEP. For example, problems have been noted with 12th grade NAEP. A commission was charged to examine these problems. Several meetings and papers resulted from this issue and the commission prepared a report that included five recommendations (National Commission, 2004). The issue of whether high school graduates are prepared for college, the workplace, and the military is currently being advanced by NAGB. The prioritization of this issue is tied to concerns across the country that the nation should be producing qualified students. Measuring preparedness will mean changes for NAEP frameworks. A second important issue with the 12th grade assessment is that of

student motivation and effort. Students are often well aware that there are no stakes for poor performance and are likely more focused on other issues in their academic career at the time. NCES wrote a 30-page response which included several foreseeable challenges related to this proposed change for NAEP. Finally, it is also important to note that the 2005 NAEP 12th Grade Mathematics results were not released until Feb. 22, 2007.

Unintended uses of NAEP data

Because the intended uses of NAEP are not clearly defined, it is difficult to ascertain what might be considered an unintended use. For example, NAGB has indicated that State NAEP scores should not be used to directly compare state performance (Shakrani, 2005) suggesting that this was an unintended use. More recently, a NAGB member (Jeb Bush) and the mayor of New York City (Michael Bloomberg) suggested that NAEP should be used to confirm or evaluate state's performance on their state assessment and accountability systems (Bush and Bloomberg, 2006). This suggestion also indicates that using NAEP scores for this purpose is currently an unintended use. More information about this topic emerged from our discussion with NAEP state coordinators who work closely at the state level with disseminating and interpreting results.

NAEP state coordinators cited several common misuses of NAEP data they had observed among various stakeholders. First, NAEP assessments are often used to compare performance across states without considering the necessary precautions before doing so. Second, many states also use NAEP data to confirm trends found in state assessment data, which may be problematic when it involves direct comparisons of achievement levels. Third, many stakeholders misinterpret change in NAEP scores, as they are unaware of the meaning of a small shift in the NAEP scale. State coordinators reported several strategies used to discourage problematic misuses. First, many of the state coordinators hold meetings throughout the year across the state within regions, counties, districts, and schools to discuss current NAEP activities (e.g., what tests are going to be given and reported that year) and to familiarize individuals with NAEP tools and resources. Such meetings are also held at universities with preservice teachers. Second, coordinators stay in continual contact with school administrators via newsletters, e-mail, and phone calls to keep them up to date on NAEP activities. This also serves to familiarize stakeholders with their State NAEP coordinator in case they have any questions on how to interpret NAEP data. Third, the NAEP state coordinators and public information officers monitor the press after a NAEP release as many reports within their state include misinterpretation of NAEP results. By closely monitoring what is being reported about NAEP, the coordinators can refute incorrect interpretations and be prepared to address questions related to these interpretations.

Validity evidence to support use of NAEP scores

Several organizations cited work they were involved in as providing validity evidence for the NAEP program. From the beginning of the NAEP assessment process, NAGB noted that the validity of inferences on NAEP scores is built on the NAEP assessment frameworks. Although they serve as the foundation for NAEP development and reporting, the creation of these frameworks alone does not ensure appropriate interpretations of NAEP results. During the development process, the frameworks are reviewed by a panel of experts who look at the frameworks in late draft form. There are also additional formal and informal reviews during the framework development process. After the frameworks are developed, items are created to match the frameworks; however, there do not appear to be any alignment studies conducted independent of the item development contractor. The closest independent review is conducted by NAGB Board members when they review the match between the framework and item pools as reported by the contractors and send an observer to item development meetings. It was unclear whether Board members would meet the general qualifications for serving as subject matter experts for

these reviews. NCES indicated that it has a minimal role in defining the intended scope of NAEP assessments and that NAGB is responsible for creating the frameworks and content specifications. NCES is invited to attend these planning meetings. NCES is responsible for translating the frameworks and content specifications into the operational NAEP assessments.

In addition to the initial development of the NAEP assessments, NCES noted six sources of validity evidence within the NAEP system that can be used to support the inferences made from NAEP data.

1. NAEP Validity Studies (NVS) Panel: The NVS is facilitated through the AIR-CA office. Research by this group has resulted in work that has been presented at conferences and published by the NVS on comparing state assessment and NAEP data, on the inclusion and exclusion policies, and accommodations. The NVS is an independent advisory group to NCES and may be viewed as an extension of the trial state assessment evaluation. The NVS is broadly representative of the NAEP research community and has a strong overlap with researchers who were part of NAEP's Trial State Assessment evaluation. Because NVS is independent of NCES, reviews of the study designs and final reports are conducted by panel members before AIR-CA publishes them. However, the determination of which studies are funded appears to be greatly influenced by the director of the assessment division of NCES. AIR-CA staff members indicated that NCES encourages them to present research at professional conferences and publish in the professional literature. The NVS prioritized several validity issues in *An Agenda for NAEP Validity Research* (AIR, 2002). The studies identified through this prioritization and rated as "essential" or "high" to "essential" were ones that addressed issues related to NAEP's capacity to evaluate state results, alignment with state standards, estimation of domain sampling error and accommodations. Areas rated as "high" included topics such as contaminations, representation of SD and LEP student, construct definition of what is being measured, and issues related to scoring and population bias. Topics that were not rated highly would have addressed interpretations of test results, comparisons of assessments to curricula, and controls and supports for secondary analysis. Although representing a number of important topics related to the validity of NAEP results, the NVS research agenda is nonetheless narrowly focused and does not address many critically important topics that warrant research in order to support intended uses of NAEP results.

2. NAEP Design and Analysis Committee (DAC): The DAC does not necessarily conduct or set an agenda for validity research in NAEP; however, in its advisory capacity to NAEP, its work relates to validity issues. The DAC deals with real time problems and monitors ETS's assessment development and maintenance activities. The DAC focuses primarily on methodologies and statistical quality, and provides technical advice.

3. Task Order Component (TOC): This is a subset of the NAEP Alliance contract and involves specific research studies requested by NCES or NAGB and may include quick turnaround projects that are requested by NCES throughout the duration of the contract. This is an innovative approach to anticipating the need to conduct studies that may not be within the original scope of work, but that may be necessary during the course of the contract.

4. Assessment Development: Much of the work conducted and documented by ETS during the development of the assessments can be viewed as contributing to validity evidence (e.g., attribute study—how much of an item is related to an irrelevant construct). These procedures, methodologies, and results are included in technical reports; however, the most recent publicly released technical report is from the NAEP 1999 Long Term Trend study (Allen, McClellan, and Stoeckel, 2005) and may not reflect current procedures. A Web site is currently under development that would present the technical report online. The lack of availability of recent technical manuals interfered with the evaluation team's

ability to learn about many of the key features of the NAEP assessment program, particularly those related to technical quality that would support intended uses of NAEP results. This delay does not appear to be due to the Alliance contractors (e.g., ETS, Westat, AIR) as they are required to submit their technical documentation per contract timelines.

Many of the research projects conducted by the ETS NAEP research division are directed at improving connections between validation efforts and intended uses of results. Validity studies are included in the NAEP program of research. A long list of research studies aimed at enhancing the validity of item development, test administration, test scoring, data analysis, and score reporting were described during the site visits. The design of the reports and the messaging from Hager Sharp were noted as ways that ETS works to improve the appropriateness of interpretations of score reports.

5. *NAEP-Education Statistics Services Institute (NESSI)*: As a subsidiary of AIR-DC, this group may conduct special studies related to validity as part of their broader responsibilities under contract with NCES. For example, one study focused on researchers' reliance on the assumption of a normal distribution of scores. NESSI also assists with different quality control components of the program (e.g., reviewing reports for compliance with NCES Statistical Standards). Note that the name of this agency changed during the course of the evaluation and was formerly known as ESSI.

6. *Secondary Analysis Grants (SAG)*: Although independent of the operational elements of the NAEP Alliance, work from these grant projects may contribute to the validity framework of NAEP. For example, some work on accommodations has come from this program that has helped inform NAEP policy. However, because these are run as a grant program, there is often little input or control over the final products of this work. A recent revision in the proposal review process has appeared to focus the priorities of the program and incorporated an external, independent process for proposal review and selection. NAGB is responsible for setting the priorities of the secondary analysis grant program. However, reviews of the proposals are conducted by an external peer review panel organized by the Department of Education's Institute of Education Sciences (IES; 2005). NAGB is not responsible for ensuring a match between the noted priorities and completed work of the secondary analysis grants.

In addition to the six sources noted above, NCES also reviews work by contractors to consider any validity implications (e.g., AIR's work on Full Population Estimates—estimates of performance for all students, not just those selected in the sample—that arose from the state analysis project). The issue of perceived competition between contractors was discussed during our site visit with NCES. NCES feels that even though there is some overlap in work conducted by contractors, the resulting competition can be beneficial for NAEP (e.g., ETS released software used to conduct their analyses because AIR distributed a similar version). Some competition is fostered by NCES to get the best work possible and these contractors are encouraged to take this work to the professional community through conference presentations and professional journals.

This multifaceted effort results in a substantial amount of research on the NAEP program and the methods used at each step in the NAEP program (See Figure 2). As an illustration of the research efforts, we have compiled a list of selected NAEP research studies that were conducted 2003–06 and are related to different aspects of the NAEP assessment program. These research studies, listed in Table 3, include both proposed and completed research, and when appropriate, the responsible agency or organization is noted. Taken together, they provide an illustration of the types of research that could serve as the foundation for the current validity framework for NAEP.

Table 3. Selected NAEP Validity Research

<p>Developing NAEP Assessment Frameworks</p> <ul style="list-style-type: none"> - A content comparison of the NAEP and PIRLS fourth-grade reading assessments (NCES, 2003a) - The impact of changes implemented in 2003 NAEP—Study 2. (ETS, Jenkins et al., 2004)
<p>Developing Test Items (Questions) and Background Questions</p> <ul style="list-style-type: none"> - Considerations in the use of constructed (open-ended) response items in NAEP (proposal, ETS, 2003) - Impact of changes implemented in the 2003 NAEP (ETS, 2004)
<p>Constructing Final Assessments</p> <ul style="list-style-type: none"> - Sparse block-matching designs in NAEP (proposal, ETS, 2004)
<p>Sampling Schools and Students</p> <ul style="list-style-type: none"> - The effects of finite sampling on state assessment sample requirements (NVS, Chromy, 2003) - Using state assessments to impute achievement of students absent from NAEP: An empirical study in four states (NVS, McLaughlin et al., 2005) - Use of sampling weights in multilevel models fit to NAEP data (proposal, Stokes, L, no date) - Development of analytic strategies to account for student nonparticipation in NAEP (proposal, ETS, no date) - Development of analytic strategies to account for student nonparticipation in NAEP—Extension of examine exclusion (proposal, ETS, 2004)
<p>Administering NAEP Assessments</p> <ul style="list-style-type: none"> - SD/LEP inclusions/exclusions in NAEP: Research design and instrument development study (proposal, ETS, 2004) - Cognitive laboratories to evaluate NAEP instructions (proposal, ETS, 2005)
<p>Scoring NAEP Assessments</p> <ul style="list-style-type: none"> - Reporting the results of the National Assessment of Educational Progress (NVS, Jaeger, 2003)
<p>Creating Scales and Links and Analyzing Data</p> <ul style="list-style-type: none"> - Using state assessment to assign booklets to NAEP students to minimize measurement error: An empirical study in four states (NVS, McLaughlin et al., 2005) - Differential item functioning analyses for students with test accommodations on NAEP test items (proposal, Kamata, 2003) - A study of equating in NAEP. (NVS, Hedges and Vevea, 1997) - Application of small area estimation methods to NAEP (grant proposal, AIR, 2001) - Skill profiles for groups of students at a given NAEP scale level: Development and demonstration (proposal, ETS, 2003) - Monitoring students with disabilities using NAEP data (proposal, Cornell University Program on Employment and Disability, 2003)

Continued next page

Table 3. Selected NAEP Validity Research (Continued)

<p>Interpreting NAEP Scores</p> <ul style="list-style-type: none"> - Test-based accountability and student achievement: An investigation of differential performance trends on NAEP and state assessments. (SAG, Jacob, 2003) - Including special-need students in the NAEP 1998 Reading Assessment Part II (ETS, 2004) - Statistical power analysis and empirical results for NAEP combined national and state samples (ETS, 2003) [Also informs the sampling section.] - Reading test design, validity, and fairness: A reanalysis of data from the 2000 NAEP Fourth Grade Reading Assessment (proposal, ETS, 2002) - Test-based accountability and student achievement: An investigation of differential performance trends on NAEP and state assessments (proposal, NAEP secondary analysis program, Jacob, 2003) - Federal sample sizes for confirmation of state tests in the <i>No Child Left Behind Act</i> (NVS, Mosquin and Chomy, 2004) - Using state assessments to impute achievement of students absent from NAEP: An empirical study in four states (NVS, McLaughlin, Scarlosa, Stancavage, and Blankenship, 2005) - Sensitivity of NAEP to the effect of reform-based teaching and learning in middle school mathematics (NVS, Shepard, McLaughlin, and Stancavage, 2005) - State implementation of NCLB policies and interpretation of NAEP performance on English Language Learners (NVS, Duran, 2005) - Linking the NAEP database with other state or federal databases: School level correlates of achievement 2000 revised synthesis plan (NVS, deMello and McLaughlin, 2005) - Inclusion of accommodations for students with disabilities (NVS, Harr, Perez, McLaughlin, and Blankenship, 2005) - A closer look at mathematics achievement and instructional practices: Examinations of race, SES, and gender in a decade of NAEP data (Lubienski and Shelley, no date)
<p>Writing, Reviewing, and Disseminating Reports and Data</p> <ul style="list-style-type: none"> - High school exit examinations and NAEP long-term trends in reading, mathematics, and science, 1970–2004. (proposal, Warren, 2004). - NCES' NAEP report formats (Goldstein, 2005) - A tool for improving precision of reporting in secondary analysis of national and state level NAEP (proposal, Von Davier and Yamamoto, no date)
<p>Improving NAEP Assessments</p> <ul style="list-style-type: none"> - Estimating relationships in NAEP: A comparison of IV and traditional methods. (proposal SAG, Chaplin, 2003). - NAEP quality assurance checks of the 2002 reading assessment results for Delaware (NCES, 2003b) - Working group on alternative estimation methodologies (Mazzeo and Drescher, no date) - Maximum estimation in NAEP: Current operational procedures and AM (Mazzeo, Donoghue, and Johnson, 2003) - Analyzing state NAEP data to address educational policy (Grissmer, 2001)

There does not appear to be an organization or agency responsible for evaluating consequential validity, although the *Standards* suggest this need. However, at the NAGB Board meeting in May, 2005 (and previous meetings) there were some discussions within the reporting and dissemination subcommittees regarding information that Board members could have to respond to media requests after the initial release of data. It would appear that these materials serve as a factor in encouraging appropriate interpretation of NAEP data in addition to discouraging inappropriate uses. These validity standards are also relevant to research or dissemination efforts within the states. One of the state coordinator goals is to promote the intended use of NAEP. Several coordinators have approached this goal by trying to promote awareness of NAEP within the state. This is accomplished by educating administrators and teachers about NAEP and including a link to the NAEP Web site from the state education Web sites. This goal also includes ensuring the proper use or interpretation of NAEP results. The state coordinators noted the intended use of NAEP data and results was to evaluate progress of students in this country.

Conclusions: Defining Intended Uses of NAEP assessments

In some respects, the intended scope and use of NAEP results are dictated by statute. However, there appear to be instances when unintended or inappropriate uses have occurred. Although NAGB does not have the power to enforce proper use of NAEP data and results, the policy body is encouraged to follow the recommendations of the *Standards* (1.3, 6.3) and preempt improper uses by documenting foreseeable interpretations that are unsupported by the available validity evidence or that violate the intended use of NAEP scores. Unlike most testing programs, data for NAEP assessments are based on a sample of schools and students rather than a census. Thus, district-level, school-level, or student-level data cannot be computed and reported due to insufficient information.

Standard 1.1 highlights the importance of providing evidence that supports any intended uses of test scores. Evidence to support validity of score interpretations abounds across the contractors; several members of the NAEP Alliance have a special studies program to provide such evidence. The NVS is a good example of how programs of research are undertaken to address validity questions and issues. The secondary analysis program encourages researchers outside of the NAEP Consortium to contribute research to support and explore dimensions of validity. These efforts, however, are hindered in their effectiveness due to the lack of an overarching validity framework with prioritization of research questions.

Given the magnitude and importance of the NAEP program, it is critical that validity research be driven by an organized blueprint designed to reflect critical questions within the program and that the results of such research be integrated into the NAEP system to provide for continual improvement of the assessment program. The existing independent research activities by members of the Alliance, NVS, and the SAG programs would be more effective were they coordinated and complementary to a strong validity program designed to address key validity questions about intended scope and uses of NAEP results.

Developing NAEP Assessment frameworks

Following the statement of purpose(s) of the test and the intended interpretations of test scores, the next step in the test development process is the articulation of the test framework including the content, skill, and processes of the construct to be measured. The test framework serves as a guide for all phases of test development. The basis for the framework can be either theoretical or based on existing statements or studies of the important knowledge and skills to be measured by the test. Once the overall framework for the test has been delineated, the next step is to translate the framework into specific

content specifications. These content specifications indicate the format of the items or tasks. All subsequent test development efforts are dictated by the content specifications.

Relevant *Standards*

Standard 3.2.: The purpose(s) of the test, definitions of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of the items to the dimensions of the domain they are intended to represent.

Standard 3.3: The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.

Standard 3.5: When appropriate, relevant experts external to the testing program should review the test specifications. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of the expert judges should be documented.

Standard 3.11: Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.

Standard 13.3: When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and target domains should be described in sufficient detail so that their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent.

Two major components were considered in this dimension of the audit: the procedures used for framework development and the process by which the test specifications were derived from the framework into the test's table of specifications (TOS). When considering the procedures used for framework development, relevant factors included the basis for the framework design and related organizational standards, the procedures used to form the framework development committee members (sometimes called subject matter experts), the timeline for development of the framework, and the procedures for review of the framework. For many testing programs, there is a distinction between the broader content specifications and the resultant table of specifications. In NAEP, the assessment frameworks are the table of specifications. For this dimension in the audit, NAGB was identified as having a key role in achieving this step in the NAEP assessment process as it has the responsibility for developing the assessment frameworks through collaboration with contractors.

According to NAGB policy, contractors for content framework development are selected based on a competitive process facilitated by NAGB (NAGB, 2002b). The evaluation team for proposals that are received for this development includes NCES, Board members, and outside individuals. The Board helps in developing the statement of work (SOW) for the request for proposals (RFP) and a subset of these individuals (who help develop the SOW) help in reviewing proposals. A designated staff member is involved in both the NAGB meetings and contractor meetings. During the process of framework

development (approximately 18 months) the Board has several opportunities to review the work of the contractors and then the framework goes for Board approval. After approval, approximately 20–25 percent of the framework committee must serve on the NCES standing committee for the item development process.

The Framework Development policy (NAGB, 2002b) describes who is involved in the process and documents the need to have content experts, educators, members of the public, and policy makers on the panel. There is an international perspective to these frameworks as many individuals on NAEP framework committees have also served on international assessment committees (e.g., Progress in International Reading Literacy Study—PIRLS, Trends in International Mathematics and Science Study—TIMSS, Programme for International Student Assessment—PISA). NAEP was able to borrow from these frameworks and subsequent research has examined the overlap between these frameworks. Because the typical NAEP framework panel consists of approximately 20 percent teachers it appears that most committee members do not have classroom teaching experience. It was unclear whether the criteria for panel membership included content knowledge or familiarity with the target population of students.

The NAGB Framework Development Policy (NAGB, 2002b) specifies the following seven guiding principles by which these frameworks should be developed.

Principle 1: The Governing Board is responsible for developing an assessment framework for each NAEP subject area. The framework shall define the scope of the domain to be measured by delineating the knowledge and skills to be tested at each grade, the format of the NAEP assessment, and preliminary achievement level descriptions.

Principle 2: The Governing Board shall develop an assessment framework through a comprehensive, inclusive, and deliberative process that involves the active participation of teachers, curriculum specialists, local school administrators, parents, and members of the public.

Principle 3: The framework development process shall take into account state and local curricula and assessments, widely accepted professional standards, exemplary research, international standards and assessments, and other pertinent factors and information.

Principle 4: The Governing Board, through its Assessment Development Committee, shall closely monitor all steps in the framework development process. The result of this process shall be recommendations for Board action in the form of three key documents: the assessment framework; assessment and item specifications; and background variables that relate to the subject being assessed.

Principle 5: Through the framework development process, preliminary achievement level descriptions shall be created for each grade being tested. These preliminary descriptions shall be an important consideration in the item development process and will be used to begin the achievement level setting process.

Principle 6: The specifications document shall be developed during the framework process for use by NCES and the test development contractor as the blueprint for constructing the NAEP assessment and items in a given subject area.

Principle 7: NAEP assessment frameworks and test specifications generally shall remain stable for at least ten years. (p. 3-4)

Often, the frameworks make use of standards from national learned societies; however, the frameworks do not necessarily follow these standards. When possible, they are included as one piece of information to be considered. Given the lag time between framework development and administration of the operational NAEP assessment, the framework development process requires forward thinking (e.g., where do we want to be in X number of years when this assessment becomes operational?) and the need to reflect best practice. The panel is not dominated by one type of panel member (e.g., policymakers, teachers).

Frameworks are reviewed whenever there is a major change in the direction of state or international assessments. The decision to change a framework is weighed between the desire to maintain a trend in the assessment and wanting to keep the assessment current. For example, in a survey of state policymakers concerning the 2005 NAEP mathematics assessment it was apparent that an update was needed in fourth- and eighth-grade mathematics but the desire was to maintain trend. The geography framework will be ready for an update in 2010 and the subgroup will revisit the framework but again, there is the desire to maintain trend. Although in other testing arenas (e.g., licensure, certification) content may be revisited more often as professions evolve, reforms within K–12 educational systems may not occur as quickly because of the systemic changes that are needed and the time needed to observe the impact.

Conclusions: Developing NAEP Assessment Frameworks

This dimension is fundamentally an activity conducted by NAGB and is firmly grounded in policy. The systems for review and revisions of the developing framework are generally consistent with sound measurement principles. Two improvements are suggested. First, some of the review processes appear to occur with reviewers who may not meet generally accepted requirements for content expertise. Second, studies that independently evaluate the alignment of the NAEP assessment frameworks with learned society standards (e.g., National Council of Teachers of Mathematics) and state content standards would provide needed validity evidence for uses of NAEP scores that have been proposed.

Developing Test Items (Questions) and Background Questions

Once the assessment framework has been defined, the next step in the test development process is to develop items or tasks that measure these frameworks. The test developer must provide information about the procedures used for item development; in some cases the test developer will use in-house item writing specialists or train external item writers. In either case, information should be provided on the procedures used for developing the items and the criteria used for evaluating the acceptability of the items produced. If external item writers are employed, documentation should be provided on their qualifications. For educational tests, in particular, evidence is needed to ensure that the items do in fact align with the assessment frameworks; often this is accomplished through the use of external alignment studies that examine the match of different dimensions (content, cognitive demand) of the items or tasks to the intended component of the assessment frameworks.

Typically, test developers construct a pool of items that is larger than the number needed for test development purposes. Items in the pool are evaluated for content accuracy and technical quality through item reviews and pilot testing. In addition to a review for content accuracy, clarity, and lack of ambiguity, items are also often reviewed for cultural sensitivity and gender issues. The procedures for item review, criteria used to evaluate the acceptability of the items, and steps used for item revision should all be documented.

For an ongoing testing program, such as NAEP, test developers often use a long-standing, but refreshed, item bank. In such a program it is important that the status of the item bank be routinely evaluated to ensure that the items maintain their technical and content integrity over time. By

periodically evaluating the status of the item bank, areas in which targeted item development is needed can be revealed and prioritized for future item development efforts. A schedule should be articulated for item development activities.

In NAEP assessments two major categories of questions are presented, those that address the cognitive domain and those that seek to measure background information about the examinee and school personnel. In both cases, a framework is used to guide item development. All the components identified above apply both to the cognitive and background questions contained in a NAEP assessment.

Relevant *Standards*

Standard 3.6: The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that the intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Standard 3.7: The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for classification and the appropriateness and accuracy of the classification should be documented.

Standard 13.3: When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and target domains should be described in sufficient detail so their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent.

Based on the elements of this dimension, the audit focused on the procedures for item development for both the cognitive and background questions. Central to the audit was information on the identification of item writers and their qualifications, the evidence gathered to support the match of the developed questions to the assessment frameworks, and the components critical for item review. Four agencies were identified as having key roles in this dimension: NAGB, NCES, AIR-DC and ETS.

Test Item (Question) Development

The NAGB NAEP Item Development and Review policy (NAGB, 2002c) lists the following principles as guiding the item development and review process:

Principle 1: NAEP test questions selected for a given content area shall be representative of the content domain to which inferences will be made and shall match the NAEP assessment framework and specifications for a particular assessment.

Principle 2: The achievement level descriptions for basic, proficient, and advanced performance shall be an important consideration in all phases of NAEP development and review.

Principle 3: The Governing Board shall have final authority over all NAEP test questions. This authority includes, but is not limited to, the development of items, establishing the criteria for reviewing items, and the process for review.

Principle 4: The Governing Board shall review all NAEP test questions that are to be administered in conjunction with a pilot test, field test, operational assessment, or special study administered as part of NAEP.

Principle 5: NAEP test questions will be accurate in their presentation and free from error. Scoring criteria will be accurate, clear, and explicit.

Principle 6: All NAEP test questions will be free from racial, cultural, gender, or regional bias, and must be secular, neutral, and non-ideological. NAEP will not evaluate or assess personal or family beliefs, feelings, and attitudes, or publicly disclose personally identifiable information. (p. 3)

These principles are detailed in specific procedures required to satisfy each policy requirement. After the items are created, a clearance package is created that shows the item or content match and the intended cognitive level. This information is then shared with the NAGB Board.

NCES is responsible for overseeing the item development process and ensuring that it follows the specific frameworks created by NAGB. Specifically, the process is overseen by standing committees made up of roughly 12–20 content specialists from the national, university, state, and local levels. Typically, one-fourth to one-third of the members of the standing committees will also be members of corresponding framework committees. The standing committees meet between two and four times per year.

In the first phase of cognitive item development, pilot items are written by different contractors based on content area: ETS and AIR-DC. ETS is responsible for writing items for the reading, math, and science assessments and ultimately for all items that appear on the NAEP assessments which include those written by AIR-DC. AIR-DC is responsible for developing items for the writing and social science assessments and background questions. AIR-DC hires content specialists and trains them on item writing procedures and their work is supervised by AIR-DC staff. ETS uses mostly in-house item writers for Reading but has a fairly substantial pool of external items writers for Mathematics. They use external item writers for some other content areas.

NCES oversees both contractors and helps with the training of the item writers to ensure the items conform to specifications and fit the frameworks specified by NAGB. Roughly twice as many pilot items are written as will be included on the final NAEP assessment to account for attrition that may occur during the piloting process.

AIR-DC brought some new expertise and procedures to the long-standing item development procedures that were used historically by ETS (who had the only item development contract prior to the new Alliance procurement model). AIR-DC directed efforts to improve the evidence of alignment of extant and newly developed cognitive test questions to the respective frameworks. Their efforts to examine item characteristics that provide better differentiated scales have been translated into item development training procedures. AIR-DC is in the process of bringing items from ETS's database into AIR-DC's Item Tracking System (ITS). The ITS has features that enable password- and privilege-dependent access to item writing, item review with comment tracking, item status checking, item statistics database generation, and eventual simulated test creation procedures to monitor compliance with test specifications.

There appears to be an issue regarding the transfer to ETS of NAEP items that have completed the full developmental and review cycle at AIR-DC. ETS, whose editors and item developers may

decide to make additional changes to the items after NAGB review, does not always articulate these changes to AIR-DC so the ITS can be brought up to date with changes subsequent to the hand off. However, it is unclear which operational contractor is the “responsible party” for the final survival and quality status of the items. The assessment items must be positively reviewed by NAGB before they are deemed acceptable for use in a NAEP assessment. Further, once the items are used in the field, either in a pilot, field, or operational administration, item statistics are computed to document the technical quality of the items. Some quality indicators of AIR-DC’s item development efforts may be distorted if these AIR-DC finalized items receive additional edits and revisions from ETS staff (which may or may not have been deemed acceptable by AIR-DC test developers as they are not consulted following ETS’s editorial decisions).

Test Item (Question) Review

ETS assumes responsibility for all items that appear in operational assessments and therefore uses their own item review processes for the items that are developed by AIR-DC. The items are then passed to NCES and the standing committee for review. Each item (with related scoring guides, when appropriate) is individually examined for match to the NAGB framework, appropriateness of the difficulty level, clarity of the question and response options, and appropriateness of scoring. Items may be rewritten by the group during the review process to achieve greater agreement among the reviewers. The items are returned to the contractors for revision, and then sent back to the steering committee for further review. A larger goal of this process is to ensure that the frameworks are being properly interpreted by the contractors (i.e., did the contractors do their job in writing items to match the NAGB framework). Also at this point, the standing committee may determine that the frameworks need additional clarification.

After the standing committee has completed their review of the items, NCES conducts a state item review. NCES pays for two representatives from each state to participate in the review (states may send more representatives at their own expense). The state representatives may be curriculum specialists, state testing coordinators, or teachers. While the feedback from these representatives may not directly affect which questions will ultimately appear on NAEP, NCES and ETS review the representatives’ comments and concerns and take action when appropriate. When the standing committee has finalized its choice of items, these items are submitted to NAGB who makes the final determination as to which items will appear on the pilot tests.

NAGB’s involvement in item review is through representation by members of the framework committee to the item development committee. The Board (by law) looks at bias and appropriateness of each item. Before this review, training is conducted on item development policy and general process for good items. During this review the Board does have the right to comment on other item characteristics. Any comments on items are sent to NCES. The Assessment Design Committee (ADC) of NAGB does a separate review of items by teachers, principals, and policymakers.

NAGB does review the reading passages that are included in NAEP assessments. The Board is given a booklet of passages and a large number are reviewed at once. The Board is responsible for ensuring that passages are engaging, appropriate, and current. Each passage receives a rating of “definitely use,” “possibly use,” or “definitely not use.” Many of the passages are taken from published texts so edits are not always possible. Approximately 15–20 percent of the passages are rejected during this process. NAGB’s comments on passages are funneled through NCES to ETS. The three step process is as follows: NAGB first reviews the passages, then the passages, items, and scoring guides, and finally the passages, items, scoring guides, and pilot data (passages are reviewed three times). NAGB reviews reading passages first to assist with the efficiency of the development process. If a passage is rejected, there is no need to write, review, or pilot test items that would be related to the passage.

The ADC of NAGB also has the responsibility of reviewing all the subject-specific background questions (e.g., number of science classes taken) and the reporting committee reviews the generic background questions. Based on policy (NAGB, 2002a) NAGB is responsible for developing the framework and specifications for these questions including specification of which topics should be included. According to policy (NAGB, 2002a) NAGB is responsible for reviewing the questions under federal legislation P.L. 107-110 based on the following criteria:

- A. Background information is needed to fulfill the statutory requirement that NAEP report and analyze achievement data, whenever feasible, disaggregated by race or ethnicity, gender, socioeconomic status, disability, and limited English proficiency. Non-cognitive data may enrich the reporting and analysis of academic results, but the collection of such data should be limited and the burden on respondents kept to a minimum.
- A. All background questions must be related to the primary purpose of NAEP: the fair and accurate presentation of academic achievement results.
- B. Any questions on conditions beyond the school must be non-intrusive and focused on academic achievement and related factors.
- C. Questions shall be free from racial, cultural, gender, or regional bias.
- D. All questions must be secular, neutral, and non-ideological. Definitions of these terms, accompanied by clarifying examples, are presented in Appendix A [of NAGB's document], as adopted in the Governing Board Policy on NAEP Item Development and Review.
- E. NAEP must not evaluate or assess personal feelings or family beliefs and attitudes unless such questions are non-intrusive and have a demonstrated relationship to academic achievement.
- F. Issues of cost, benefit, appropriateness, and burden shall be carefully considered in determining which questions to include in background questionnaires. These factors must also be considered in determining the frequency with which various questions shall be administered and whether they shall be included in both national and state samples.
- G. Background questions that do not differentiate between students or have shown little change over time should be deleted or asked less frequently and to limited samples. (p. 5)

Pilot testing

A pilot test is administered to a nationally representative sample of approximately 500–1,000 students, representing the full range of ability. At least two items are pilot tested for each operational item that is needed. Item statistics are analyzed and items and item blocks are examined for difficulty and possible bias with differential item functioning (DIF) analysis (analyses of group performance at the item level when controlling for ability). Items may be dropped or reworked if necessary. The results are reviewed by the standing committee, and in the case of the reading and math assessments, the items may undergo a second pilot test. The items and item blocks that performed well then go on to make up the operational exams. NAGB has one final review of the items before the assessment becomes operational.

Other Test Item (Question) Development Activities

Cognitive item development is a continuous process. Roughly every ten years new assessment frameworks are developed which require updated item sets. Also, about one-fourth to one-third of NAEP items is released after each assessment. Therefore, continual replenishment of the item pool is necessary. NCES and the item development contractors determine which items to release so that the items are representative of the NAEP assessment.

Three sources of quality control were noted for the item development process. First is the extensive review process. Items are reviewed by the standing committee, by the state reviewers, and by NAGB. This multistage process is used to ensure match to the test specifications, appropriate difficulty, and fairness. Second are the statistical analyses that are incorporated within the item development process. Specifically, DIF analyses are used to evaluate potential bias and sensitivity across groups, the relative performance across ability levels, and performance is explored across time (by large samples and as a group comparison). Third, at each review session, NCES collects comments about each item and is forming a coding system to organize these comments.

The trend assessment's process is slightly different from that described above. First, these assessments are not based on frameworks as the Main assessments are. The content was defined by the trend assessments that were constant in the mid-to-late 1980s. Since this time, some items have been replaced with the new items being reflective of the retired items. Bridge studies are currently being conducted to determine if this modified assessment is measuring the same content as the old assessment.

ETS is also responsible for the preparation of translated versions of the assessments (Spanish for Mathematics and Science). In these instances, translations are performed to reduce the potential impact of language on students' opportunity to demonstrate their abilities in Mathematics and Science.

The background questions are developed in much the same way as the cognitive items. Background questions are included in student assessments, in teacher surveys, for students with disabilities (SWD) and English language learners (ELL) student surveys, and in principal surveys (to assess the demographics of the school). The purpose of the background questions is to unobtrusively gather information to aid in the interpretation of cognitive item database. NAGB is responsible for developing the frameworks and item specifications for the background questions and AIR-DC is contracted to develop these items. There are three types of background items developed:

- 1) Reporting—these items are used in NAEP reports and include such variables as region of the country and ethnicity.
- 2) Subject specific—these items measure students' experience with subject matter and related variables
- 3) Other contextual variables—these are designed to measure equitable distribution of resources and opportunity to learn.

AIR-DC has taken a proactive role in the articulation of a model for the background questions, called the Contextual Variable Inference Map (C-VIM). The model allows for a systematic and strategic use of background questions to address important questions related to the influences of certain school, teacher, and student variables on student achievement. In addition, the Item Tracking System (ITS) mentioned previously also has the capacity to include the background questions and this application is currently being finalized. After development by AIR-DC, the background questions are submitted to the standing committee for review and follow a process similar to the one used for the cognitive items. To maintain consistency, many of the same background items are used year after year. In addition, an effort is made to maintain consistency of items across tests (subjects) to allow for comparisons.

Background questions must also be submitted for approval by the OMB. In the past, OMB has requested item revisions. However such changes are considered minimal now by OMB due to the general consistency of items across years.

Conclusions: Developing Test Items (Questions) and Background Questions

In general, the item development and review practices employed by ETS and AIR-DC are consistent with the *Standards* and with sound assessment practices. ETS and AIR-DC work together, and independently, in developing the cognitive questions for the NAEP assessments. A better tracking system to monitor and record changes in items across these two vendors would strengthen the item development program. Because questions that are developed for inclusion in NAEP assessments undergo multiple steps in the development process, such a tracking system would help ensure that all parties in the development and review process are aware of what changes have been implemented and what is the final version of the items. Communications between the test development vendors appears to be strong and mutually supportive.

NAGB's role in this process would benefit from documentation and dissemination of the qualifications of the reviewers, the process it uses to review passages for reading and items for all assessments, and the results of these studies. The importance of independent reviews by qualified experts in this process cannot be overstated. Driven in part by Peer Review Guidance requirements of *NCLB*, current practice in educational assessment involves independent alignment studies that demonstrate that the resultant assessment corresponds to the intended assessment framework in terms of content, cognitive demand, balance of coverage, and sufficient information to support reported achievement levels. Documentation of these review processes is not currently published in a technical manual or supporting literature. Because NAGB has final approval of the items for inclusion in a NAEP assessment, this element of quality control is an important part of the process.

Creating Draft Assessments, Preparing Field Test Designs, and Conducting Field Trials

In the test development process, after the items have been developed, but prior to operational use, the next step is to pilot test the assessments to ensure that they are functioning appropriately. It is important, to the extent possible, that the examinees for the field test are representative of the examinees who will take the test when it is used for reporting NAEP results. It is also important that the administration procedures parallel, as closely as possible, those procedures that will be used in the operational assessment.

Relevant *Standards*:

Standard 3.7: The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If items were classified into different categories or subsets according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented.

Standard 3.8: When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended.

Although small pilot testing of the items occurs prior to administration of the field trial, most of the critical information about the items is derived from the pretesting of items during the operational

administration. For the NAEP administration, blocks of items are inserted in the operational testing for pretesting items for future use. Because examinees do not know which items are operational items and which ones are to be used only for pretest purposes, the conditions for the pretest and operational items are the same, helping to ensure the veracity of the field test data.

Information was sought about how these pretest blocks were assembled and about their alignment to the assessment frameworks. Also, information was gathered about the logistics for the pilot administration. Because the field testing is subsumed within the operational administration, much of this information was gathered when the operational administration information was assembled. The criteria used in evaluating the results of the pretest were also relevant here as some of the information from item performance is used when assembling operational assessments from the piloted items. For this dimension, ETS and Westat were considered to have active roles.

The block design uses common items to link results across years and for reporting of trend results. Booklets are configured using a modification of a balanced incomplete block design to ensure that all blocks are paired and that all blocks appear in all positions in the assessment. This is a critical issue for the reporting of trend as the current block design reduces the sample size causing certain types of errors that can undermine the linking of assessments across years. Also included in the assessment design are special studies or other booklet components that will affect the total number of assessment formats that are administered. ETS uses proprietary software that calculates the needed booklet formats to accommodate these assembly issues.

To improve the quality of pretest data for *NCLB* content areas, ETS has adopted a practice of pilot blocks. These pilot blocks are constructed to be responsive to several test development issues, such as breadth of content coverage, range of item difficulty, and position effects. These pilot blocks are used in operational settings following pilot testing and kept together as a unit in operational administrations. This has allowed for more confidence to be placed in the item statistics that result from the pilot administrations and has allowed for more efficient use of starting values for operational calibrations and scoring.

Because the field test is subsumed within the operational administration, additional information is contained in the *Sampling Schools and Students* dimension.

Conclusions: Creating Draft Assessments, Preparing Field Designs, and Conducting Field Trials

This dimension appears to be primarily met through the administration of the assessments. The field test sampling procedure adheres to the *Standards* as the field test is administered to an operational sample. In the administration, pretesting occurs for items that will be used in future assessments. There is real strength in this pretesting plan as the students are unaware of which items are operational and which ones serve pretest purposes. This helps ensure the accuracy of the calibrations of the field test items for when they are used operationally.

Collecting Data on NAEP Assessments

Constructing Final Assessments

Once the items for the range of assessments (e.g., Main, State) in their respective content areas (e.g., Reading, Mathematics, Science) have been developed, reviewed, and field tested, the next step in the test development process is to assemble the test forms for operational administration. A test form can be viewed as the collection of items and tasks (i.e., test questions) that were selected to measure the assessment content frameworks. In an assessment program like NAEP that involves multiple forms that sample from different sections of the assessment framework, it is important to ensure the forms meet the requirements for test specifications. Following the assembly of the test forms to test specifications by ETS, the tests must be packaged and prepared for distribution by PEM. Westat is also involved in this process as they provide the student and school information to PEM that is then included in the printing process to ensure that materials are sent to the correct locations. Multiple contractors are involved in this step of the process, so there are necessary communications and handoffs that occur to ensure that the process runs smoothly. Because this step involves many individuals across organizations, there are a number of quality control procedures that must be put into place to ensure proper handling, receipt, and tracking of student test booklets.

Relevant *Standards*:

Standard 3.6: The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

Standard 3.11: Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.

Because this stage of the process involves multiple contractors integrating different parts of NAEP assessment production in real time, many features of the process need to work together like a well-oiled engine to meet strict administration deadlines with tight assembly, packaging, and distribution requirements. Once the multiple test forms have been printed and checked for accuracy through quality control procedures, the test materials must be packaged in pre-determined spiraling patterns for shipment to the multiple assessment administration sites. Procedures for ensuring these steps are accomplished accurately must be monitored and documented. Quality control procedures are critically important at this stage of the test development process.

Both ETS and PEM assume responsibility for this audit dimension. ETS provides PEM with the booklet and spiral “scripts” that are used by PEM for booklet printing and bundling. ETS also reviews print documents for accuracy and technical quality. As the coordinator of the NAEP Alliance, ETS provides many of the internal quality checks for different stages of the process.

Based on printing specifications (i.e. booklet and spiral “scripts”) received from ETS, PEM then has the responsibility for printing the multiple test booklets and ensuring their quality. The integrity of this process is supported by several procedures including dedication of time for reviews of mock-ups that involve multiple review teams within PEM, ETS, NESSI, and AIR-DC. The goal is to catch any printing issues early in the printing process when corrections can be achieved in an efficient and less

costly manner. Once the mock-ups have been approved (and relevant green lights have been provided by government agencies), print runs are completed and delivered to PEM's Cedar Rapids facility. At that facility, specifications are used to prepare the booklets for shipping, including the fulfillment of bundling specifications for packaging the materials for delivery to Westat test coordinators in the field. Several systems are in place to ensure that these specifications are fully complied with, including the use of scanning technology to check for a match with the specifications for booklet spiraling. These specifications are complex and the procedures appear to be effective in monitoring compliance with the specifications.

Conclusions: Constructing Final Assessments

This is an area in which strong communication and cooperation is needed across the contractors and it appeared from our observations that the systems in place are working well and smoothly. Test booklets were printed in accordance with the specifications defined by ETS, packaged and distributed to the desired locations. Because substantive problems were not noted or observed in these areas, we can conclude that the procedures and results of this dimension are strengths within the NAEP Alliance.

Sampling Schools and Students

Unlike most educational testing programs, NAEP assessments (e.g., Main, State) do not report scores for individual students, instead they rely on sampling procedures to obtain representative samples of intended populations (e.g., national, state). Scores from samples of students are used to represent the likely performance of all students had they, in fact, taken the full assessment rather than a sample of the items. Therefore, it is critical that the sampling plan and implementation be sufficient for reporting scores both for intended purposes and intended populations of students.

Relevant *Standards*:

Standard 3.8: When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended.

Standard 15.5: Agencies using tests to conduct program evaluations or policy studies, or to monitor outcomes, should clearly describe the population the program or policy is intended to serve and should document the extent to which the sample of test takers is representative of that population.

Standard 15.6: When matrix sampling procedures are used for program evaluation or population descriptions, rules for sampling items and test takers should be provided, and reliability analyses must take the sampling scheme into account.

Although we have focused on *Standards* as promulgated by AERA, APA, and NCME (1999), NCES has developed and adopted more detailed standards (NCES, 2002) for designing surveys, collecting data, and analyzing data. Because NAEP assessments can be characterized as large-scale surveys, the NCES standards are applicable to these studies.

For this audit dimension, we gathered information about the sufficiency of the sampling design for Main and State NAEP⁶ scores assuming the current intended uses of these scores as indicators of national and state performance, and the strategies for weighting individual scores in order to achieve appropriate representations of subpopulations for reporting purposes. We also considered the representation of the final sample in terms of response rates, school and student replacement rates, and quality indicators for population estimates derived from the sample.

This is another dimension for which multiple contractors share some responsibility; however, Westat has the primary role in sampling schools and students. NAGB and NCES have also helped define the technical expectations for this dimension. ETS provides some information to Westat about the number of booklets that will be used in the administration for Westat to use in completing the sampling plan.

Because the sampling design and procedures have changed since the previous NAEP evaluation, we placed additional emphasis on this dimension of the audit. As part of our review of the sampling procedures, an external member of the evaluation team conducted a document review of the sampling procedures available in the Web-based technical manual from the 2003 NAEP assessment that is under development. Some specific results of that review are included in this section of the report. This full review is included as part of the Westat site visit report (Appendix G10).

More recently, NAGB has paid particular attention to response rates and sample sizes as their responsibilities have expanded regarding the initial release of the reports.

NAEP Sampling Procedures based on 2003 Draft NAEP Technical Manual

The recent decision to combine samples for State and national Main NAEP for greater efficiency represents a significant change to the NAEP sampling design. Until the *NCLB* legislation effectively mandated state participation in NAEP at fourth and eighth grade, an augmentation sample was required to measure students in states that declined to participate in State NAEP. Currently, this state-level augmentation is unnecessary at these two grade levels as states are required to have at least 85 percent of their sample participate for results to be published on NCES's Web site. However, there still appear to be separate samples collected to gather information because of challenges with using a combined sample. The sample, though, is supplemented in many ways to account for other subgroups of interest (e.g., ethnic minority, ELL, charter school, Department of Defense schools).

It is also important to note the differences between the required levels of participation at fourth and eighth grades versus the voluntary participation at twelfth grade. Although a district may refuse to participate, this makes them ineligible for Title I funds. Schools, parents, and students may also refuse to participate. For example, high school science did not meet the 85 percent participation requirement for reporting results. Currently, twelfth grade NAEP assessments are conducted at the national level, but not at the state level. NAGB and NCES have been engaged in ongoing discussions about motivation issues and participation rates at the twelfth grade level. Note, too, that NAEP assessments sample both public and private schools; however, *NCLB*'s legislation focuses on public schools, not private schools. Thus, the inclusion of students within the frame of NAEP assessments is broader than that of the legislation.

NAEP sampling and weighting are accomplished through multiple stages that occur throughout each year of assessment administration. The 2003 NAEP administration (the most recent one where draft technical documentation was available) included Main NAEP, State NAEP, and urban (Trial Urban District Assessment or TUDA) assessments in mathematics and reading. Westat is generally responsible for all aspects of sampling, weighting, and field operations (including data collection) employed in the NAEP program; the processes used by Westat for Main and State NAEP in 2003 are detailed below (some technical documentation for 2003 was omitted from the NAEP Web site that is under development;

⁶ Note that Main and State NAEP sampling characteristics were prioritized within this report due to changes beginning in 2002 and also given the ongoing discussions about uses of Main and State NAEP scores in Reading and Mathematics.

when no 2003 information was available, this section of the report draws upon documentation from the 2002 administration instead).

Sample Design

The NAEP sample design is revised annually through a collaborative effort led by Westat and involving all members of the NAEP Alliance. The sampling plan contains specifications for multiple strata (e.g., public schools, private schools, ethnic minority). The number of sampled schools and the implied number of sampled students are compared to the sample size requirements in the annual sample design. Westat statisticians review tabulation reports showing sample counts by selected characteristics spelled out in the annual sample design. Any samples that do not yield at least the minimum number of students specified in the annual sample design are redrawn. Eligible sampled schools were assigned assessment sessions on the basis of enrollment of students eligible for assessment at the appropriate grades. Although larger schools were assigned more than one assessment session, most schools were assigned a single session.

Sample Design: 2003 Main NAEP

Since changes to the sampling design in 2002, State NAEP samples have included fourth and eighth grade students in public schools in participating jurisdictions (i.e. those that accept Title I funds are required to participate under *NCLB*). In choosing to use combined state samples rather than a single national sample NAEP has traded efficiency (combined state samples are roughly ten times the size of a single national sample) for precision (greater samples allow more precise measurement). If a national assessment was the only purpose, this tradeoff may not be considered worthwhile; however, because precision at the individual state level is also required, there is little reason to prefer a separate national sample solely in terms of the efficiency tradeoff. ETS research has detailed the additional precision of combined state samples, only slight discrepancies between combined and national estimates, smaller standard errors associated with combined estimates, and a reduced need for post-stratification adjustments in using combined samples. The use of combined samples appears to be a change for the better for Main NAEP. However, with this strategy, there is greater sensitivity to changes at the national level that may seem to make small, statistically significant changes appear more meaningful than they actually are. This increased sensitivity could unintentionally influence policy decisions.

To obtain a nationally representative sample for Main NAEP, state samples must be supplemented with public school samples for those jurisdictions that ultimately did not participate in State NAEP as well as a nationally representative private school sample. Public school sample augmentation is relatively straightforward. Jurisdiction school samples were established before it was known exactly which jurisdictions would ultimately participate in the state program. School samples were drawn from all jurisdictions as part of State NAEP—including those jurisdictions that did not ultimately participate in State NAEP—to ensure that the Main NAEP sample was representative. In the state sampling process probabilities of selection were calculated for each school based on jurisdiction. For Main NAEP these probabilities were recomputed to represent the likelihood of selection as part of a national sample (rather than within each jurisdiction).

Inclusion and Accommodations

The target population for 2003 Main NAEP included all students in public or private schools who were enrolled in the fourth or eighth grades in the 50 states and the District of Columbia. Because NAEP is intended to provide achievement estimates representative of all students in state and national populations, every effort is made to include every student capable of participating. Inclusion of students for whom regular NAEP assessments may not be appropriate has represented one of the major challenges to NAEP. Starting in 2002, NAEP required states to use the same standard rules for including

SWD and ELL students in NAEP assessments; these rules were designed to lower the rate of students excluded from NAEP participation. Based on these expectations, the majority of students participating in NAEP completed assessments under standard conditions; the only exceptions to this were students with disabilities (i.e., students with an IEP developed under *IDEA* or those with an accommodation plan under *the Rehabilitation Act's* Section 504 or *ADA*) and students identified by school personnel as having limited English proficiency (with fewer than three years of English instruction). Differential participation, whether due to exclusion or other factors such as absenteeism, could substantially impact comparability of state results.

Although the procedures adopted in 2002 were designed to increase participation and improve the consistency of inclusion across states, whether these goals were accomplished remains an open question. The state-level student participation rates vary substantially. Fourth-grade participation is generally greater than eighth-grade participation; however, differences among states—from a high of 97 percent participation of North Dakota fourth-graders (in both math and reading) to a low of 85 percent of New York eighth-graders participating in mathematics—remain substantial. It is well known that participation in assessments such as NAEP is related to student characteristics, the degree of interstate variability in participation could impact the state-by-state comparability of NAEP scores.

Once school and student samples are selected, Westat delivers to PEM files containing school, grade, session, student, and shipping information. PEM uses these files to prepare preprinted Administration Schedules and to assign and track assessment booklets. Prior to delivery, the content of files prepared for PEM is compared to a master file. To determine whether transmission was successful, PEM returns the files and they are compared to the master file. If summary counts and frequencies suggest discrepancies between files sent to PEM and files received from PEM, the system is reviewed for possible programming errors. The process is repeated until returned files match those transmitted.

Weighting

NAEP weighting programs are updated annually to account for changes in state and national populations. Student weights for the National sample contained three components: a base weight, an adjustment for school nonparticipation, and an adjustment for student nonparticipation. Weights may also be scaled (post-stratified) so that sums of weights for appropriate subgroup estimates are consistent with known national totals of assessable students across the nation. Weights for students sampled but excluded from assessment are estimated in a similar manner.

In addition to overall estimation weights, replicate weights—used to estimate sampling variability of NAEP estimates—are also provided for each student, excluded student, and school. Replicate weights are important to the jackknife variance procedure currently used to generate approximately unbiased estimates of sampling variance results.

Quality Control Procedures

Westat has well-established algorithms to check the accuracy of weighting programs. Weighting programs are run using test data that will produce known outcomes if the programs work properly. Test-generated weighting values are compared with known weighting values as a quality check; deviations are flagged for further review. Weighting programs are adjusted as appropriate and the testing process is repeated until differences fall within a specified tolerance range.

Final trimmed weights must be delivered to ETS for use in NAEP score estimation. Prior to delivery the content of files is compared to a master file. To determine whether file transmission was successful, ETS returns the files and they are compared to the master file. Discrepancies in summary counts and frequencies trigger a review of the system for possible programming errors; this process is repeated iteratively until returned files match those transmitted.

Conclusions: Sampling Schools and Students

Because NAEP relies on a sample of students, instead of a full census administration, this is a critical element to ensuring the validity of score interpretations. There have been some changes in the sampling procedures and methods since the last NAEP evaluation, so additional focus was put on this dimension. Some areas were identified where additional studies could help inform whether the current sampling methods and procedures support sound measurement practices. Specifically, attention needs to be addressed to the inclusion/exclusion policies of states, accounting for school and student nonresponse and refusal to participate—particularly at the 12th grade, ensuring adequacy of state samples, impact of repeated sampling of schools and districts across multiple assessment administrations, and the methods for estimating sampling variability of NAEP estimates.

Administering NAEP Assessments

Systematic and consistent procedures must be followed to ensure comparability of the testing experience for students who take the assessment. The comparability of the testing experience is essential for the interpretation of the results. Especially with a large-scale, national assessment program that uses many administrators, procedures need to be in place to ensure proper shipment and receipt of the materials. Because of the magnitude of NAEP assessment administration, it is important that the training program for administrators provide support for standardization across sites. Security is also critically important and procedures need to be in place to protect the integrity of the assessments and the validity of the results.

Relevant Standards:

Standard 5.1: Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or a test taker's disability dictates that an exception should be made.

Standard 5.2: Modifications or disruptions of standardized test administration procedures or scoring should be documented.

Standard 5.3: When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.

Standard 5.4: The testing environment should furnish reasonable comfort with minimal distractions.

Standard 5.6: Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.

Standard 5.7: Test users have the responsibility of protecting the security of test materials at all times.

When examining the administration procedures for NAEP assessments, our audit focused on procedures for selecting and training test administrators, the logistics for administration, and accommodation policies. Special attention was also given to security procedures for administration of the NAEP assessments. Although Westat has the primary responsibility for administering NAEP assessments, PEM and the NAEP state coordinators also play important roles in this part of the process.

Westat's staffing needs for administering NAEP assessments are great and vary by the administration. Some years (e.g., 2005) have greater administrator needs than other years (e.g., 2003) because of the number of assessments in the cycle (e.g., reading, mathematics, science, writing, etc.). In 2005 there were 5,000 field staff needed to administer NAEP assessments compared with 3,500 in 2003. Most administrators and field staff members are retired educators (approximately 90 percent) and there is relatively small turnover in the group (attrition was estimated by Westat to be 15 percent). Before training begins potential administrators undergo a background check and complete a home study course. There are a series of training activities that highlight the key elements of the administration process, particularly the ones that have the greatest chance to impact the validity of scores. These are well documented in the training manuals for the assessment coordinators (ACs) and assessment administrators (AAs). The training manuals also highlight characteristics of the administration process that are potential threats to validity for which field staff should monitor. This is a somewhat novel approach to training that goes beyond just the specific, operational expectations and provides some assessment literacy about how this component fits into the bigger picture of the NAEP assessment system. It also helps with quality control because administrators are more aware of the potential problems.

Assessment coordinators are responsible for assembling packages for the schools and are familiar with the forms, supervisors, and school questions. They also conduct pre-assessment visits in January to prepare the school for the specifics about the administration. There is a Quality Control Booklet that provides a scripted protocol for the pre-assessment visit to ensure standardization. As part of the quality control procedures, there is a Quality Control log and information gathered from debriefing interviews that may impact the process.

Because of the detail-oriented nature of the six week administration period for the operational NAEP assessments, another layer of challenge is added when special studies that may require deviations from the typical administration practices are included. For example, NAGB requested three special studies during the 2005 administration making the logistics to include these more difficult, particularly when the request was during a year where a greater number of administrators were already needed. Because of their experience in administering the NAEP assessments, Westat's operations leaders are often given an opportunity to provide input on the design of some of the special studies (e.g., arts–clay, dance sequence; foreign language–performance assessments; science–manipulatives). However, there is some tension between efforts by NAGB to be “cutting edge” versus what is practically and economically feasible within the scope of the contract.

As part of the ongoing monitoring of the assessment administration, supervisors visit each administration team 1–2 times during the administration. Following administration, Westat conducts callbacks to 25 percent of the schools to interview local representatives to ask about the administration. If something negative arises from the callback, they will contact all of the schools of the individual who was responsible for the administration. PEM also plays a role in the process by monitoring the delivery, receipt, and return of materials through the PEM Alert System. As a limited external quality check on the administration process, HumRRO also conducts site visits to a few sites (approximately 15 schools) and submits observation reports to NCES.

Feedback on the administration process informs the design of the administration system. Debriefing forms and meetings with staff members, state coordinators, and NCES are all part of the process to learn about what worked and what could be improved about the administration process. This information is then integrated into the feedback loop when changes are suggested. Westat provided two examples during our site visit of such changes. First, there is a policy that precludes administrators from opening bundles of booklets until one hour before the assessment. Although this is an important security precaution, for large schools that may be administering multiple subjects, the administration team likely needs more time to prepare. Second, the timing of the pre-assessment visits currently occur 2–3 weeks in advance of the assessment so there is a standardized amount of time before each administration. There

has been a request to move all pre-assessment visits to January to make it easier to manage some of the logistics involved in the operational administration.

Because 2007 will be a big year in the administration schedule due to a greater number of assessments administered than in other years, it will be important to stay on the critical path and carefully consider the number of special studies that could interfere with the primary purpose of the assessment. NAGB is encouraged to consider special studies in the context of the assessment schedule as opposed to the relatively short notice of the more recent studies. This is especially important during administration years that include a third subject (e.g., science; writing–2007). The additional subject areas require large increases in staffing and the addition of special studies then requires augmentation to the training activities for those administrators who will be responsible for administering elements of the special studies.

PEM is also very involved in the administration process as they are responsible for packaging, shipping, and receiving the test booklets from the administration. Use of several communications systems help support assessment administrators once the materials are in the field, including customer hotline support and fax communications. Communication linkages with Westat are also maintained when the assessments are in the field to keep both partners fully informed of issues related to assessment receipt and delivery. PEM has put into place several “customer friendly” procedures to help ensure that the administrator in the field can achieve the intended administration procedures, maintain accurate assessment records, and return the materials in an efficient manner.

Once the assessments have been returned to PEM, additional systems are in place to monitor receipt control and security. PEM attempts to protect the security of the assessment through inventory systems to track receipt of all materials that were shipped. Materials are held in an “alert” area until receipt control issues are resolved. The inventory systems are generally tracked electronically.

As a third partner in the leadership of NAEP assessment administration, state coordinators are responsible for several activities during the NAEP administration. The amount of time required by this activity depends on several factors (e.g., if the state was selected to participate in a pilot study, how many schools in their state were selected to participate in NAEP, the type/number of assessments being conducted that year, and if there is a state mandate for NAEP participation). Some states have legislation requiring participation in NAEP for any school that is selected; however, this is inconsistent across states. Although *NCLB* requires participation in State NAEP in fourth and eighth grade for schools that receive Title I money, the requirements for schools that do not receive these federal dollars are state-specific. Without such legislation to assist the process, the NAEP state coordinator must spend time recruiting schools that have been selected in the sample. This activity may involve several forms of personal communication (e.g., letters, phone calls, visits) which can be quite extensive. After recruitment, state coordinators are responsible for entering information about participating schools into the school control system. Coordinators expressed frustration with this system because the information cannot be uploaded electronically. As the administration date approaches, state coordinators commonly serve as a liaison between schools and the NAEP field staff in making preparations. During the day of administration, state coordinators often observe as many administrations as possible and try to intervene with any administration problems.

The NAEP state coordinators noted several problems with the administration of NAEP. First, some of the coordinators suggested that there were not enough field staff available during the administration. This issue is likely to be state specific due to differences in student populations and accommodations policies. The staffing concern is particularly related to years when there are larger samples needed because of a greater number of administrations. Some of the state coordinators indicated that many of the field staff in some states were unprepared and quit (in some cases a third) during the administration. These situations, though, appeared to be isolated and not nationally representative. They speculated this was due to poor recruitment, low pay, and unrealistic workloads. A related issue may also be the difficulty that the NAEP state coordinators have in balancing their responsibilities between

their state Department of Education's request and the requirements of NAEP. A second problem noted was that the NAEP questionnaires for students with disabilities were too long and required extensive time to complete. In addition, many school assessment coordinators were faced with reviewing the individualized education programs (IEPs) and related forms for SWD and all ELL students for NAEP assessments to evaluate accommodations that were acceptable.

Conclusions: Administering NAEP Assessments

The administration dimension involves coordination and cooperation across multiple members of the Alliance, specifically Westat and PEM. NAEP state coordinators also play an important role at the state level to assist in fulfilling the sampling plan. One of the important operational components that allow the administration to flow smoothly is the electronic monitoring systems in place to ensure tracking of the materials from the time they leave the warehouse until their safe return. Security is a highlighted component for this dimension as the integrity of the NAEP system depends of the security of the assessments. Additional attention to the training of field administrators for their role in administration could improve the integrity of the scores; however, given the number of administrators and sites, some variability is inevitable and likely does not substantively threaten the validity of the scores.

This page intentionally left blank

Scoring and Analyzing NAEP Assessment Data

Scoring NAEP Assessments

Once the completed paper-pencil assessments have been shipped from the schools, the responses need to be scanned, scored and prepared for analyses. This stage in the process is necessary to transfer the hard copy responses into an electronic format that can be more easily used in the data analyses. Multiple-choice items are typically scored by machines (i.e. optical scanning) and the accuracy of the machine scoring should be verified. Open response items (e.g., short answer, extended response) are scored through a separate process, often using trained human raters. Sometimes these responses are also electronically scanned, but may also be scored in the hard copy format. Monitoring is needed to verify the accuracy of these scores, regardless of the mode in which the performances are scored, over time. Once the scoring is completed, these results need to be analyzed to provide interpretable results. The final database of student scores is the input to the next phase in the assessment process: Creating scale scores and links.

Relevant *Standards*:

Standard 5.8: Test scoring services should document the procedures that were followed to assure accuracy of scoring. The frequency of scoring errors should be monitored and reported to users of the service on reasonable request. Any systematic source of scoring errors should be corrected.

Standard 5.9: When test scoring involves human judgment, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented.

Standard 13.10: Those responsible for educational testing programs should ensure that the individuals who administer and score the test(s) are proficient in the appropriate test administration procedures and scoring procedures and that they understand the importance of adhering to the directions provided by the test developer.

For this audit dimension, the focus is on the quality and integrity of the scoring procedures for both multiple-choice and open-response items. For open response items, we directed our attention to the selection and training of the scorers, evidence for the quality of scoring and quality checks. Attention was also given to the procedures for collecting and storing student data. Security procedures for the collection and storage of examinee data were also considered. Two members of the Alliance play key roles in scoring NAEP assessments: PEM and ETS

Once students' test booklets are prepared for scanning, several checks are in place to protect the integrity of the scanned capture of the student responses. Multiple-choice responses are captured electronically and prepared for transmittal to analytical scaling and linking procedures that are completed at ETS. Open responses are also captured by proprietary scanning software and prepared for use in human scoring under the direction of PEM's scoring processes. To score the open responses, scorers work on computer terminals that bring in the scanned image of the student's written responses and then assign their ratings electronically.

Another of PEM's roles in this dimension is in their preparation of the scorers for responses to constructed-response NAEP prompts/items. The responsibility for training of the scorers switches from ETS (the item and rubric development) to PEM as the open response questions move from pilot (when they are still in development) to operational, post calibration status. In the scoring procedures, different

issues are in place depending on whether the open response questions serve a trend or non-trend role. ETS has the responsibility for identifying and developing the training sets, and depending on the status of the questions (pilot or operational pre-calibration or not) ETS may or may not have additional training responsibilities. Regardless of whether PEM or ETS conducts the training, the scorers are recruited by PEM to meet scorer eligibility and scoring is conducted in PEM's scoring facilities.

Current research studies are in place to explore alternative strategies for scoring procedures for trend responses. In the past, trend question scoring occurred as pre-planned (and nontransparent) events in the scoring procedures. A stronger psychometric design for scoring of trend questions would be that they occur without knowledge of their "trend" status, integrated within the other constructed-response questions assigned to the scorers.

Procedures for gathering validity and reliability evidence involve the use of "backreading" by the scoring supervisor and randomly obtaining a second score for a percentage of the papers (either 5 percent or 25 percent depending on the volume of responses). Backreading is implemented as a mechanism for monitoring the calibration of scorers with intervention strategies in place for a scoring supervisor to take different actions depending on the severity of the problem. Supervisors may simply communicate (directly via face-to-face conference or indirectly via e-mail) with the scorer to alert him or her to concerns about score decisions or the supervisor may make a decision to "reset" a question and reseed the responses into the scorers' scoring set, perhaps following a retraining of one or a group of scorers.

Several issues were raised through the discussion about open response scoring. First, there does not appear to be a systematic use of "validity" papers, either for the non-trend or trend questions. For non-trend questions, it would be highly desirable to include validity check papers in the papers seeded to scorers. This is common practice in the scoring of performance assessments. Monitoring of scores on these validity papers would provide additional information to the scoring supervisor regarding the need for retraining or disqualification of a scorer. Instead of systematic use of validity papers, PEM uses "backreading" by senior graders as a means of identifying graders who may need retraining. The issue of maintaining the level of scorer quality is particularly important when most performances are only scored by one scorer. The issues are more complex with trend papers due to the changes that have occurred over time regarding the scoring of these papers and the need to replicate whatever idiosyncrasies might have been in place in the prior scoring procedures.

Second, the decisions regarding how the results from a second scorer and supervisor's backreading results are used should be reconsidered. These results are used only for quantifying inter-rater reliability and for identification of scorer drift. These score values, regardless of whether they bring into question the accuracy of the first scorer's score value, do not alter the first score even when evidence might suggest they are inaccurate (unless the supervisory decides to disqualify, i.e., "reset", this question, retrain, and then have the question reentered into the scorers' set of questions to score). Although, it could be perceived that it is PEM's responsibility only to provide the obtained score records to ETS for use with their scaling algorithms (which would be analogous to how ETS uses the scanned responses from the multiple choice questions), another perspective is that it is PEM's responsibility to ensure the validity of these constructed-response scores that are transmitted to ETS for their processing. This would be similar to the steps that PEM now carries out to ensure the validity of the scanned images for both the multiple-choice responses and the open responses. Additional attention to the validity of the scores provided for the open responses is desirable.

Following the completion of these multiple data capturing procedures, data files are prepared and made available electronically to ETS, Westat, AIR-DC, and NCEs. PEM stores student test booklets and ancillary materials used in NAEP assessments for an indeterminate period. Once these data are available in electronic format, the responsibility then transitions to ETS.

ETS shares responsibility for scoring the constructed response items with PEM; ETS has this responsibility for the NCLB content areas of reading and mathematics, even when these items are not

yet operational. In the Alliance arrangement, PEM is an independent contractor, whereas in the past PEM was a subcontractor to ETS for NAEP scoring. Although ETS does not have direct responsibility for some of the scoring practices, they maintain responsibility for the validity and reliability of the scoring as it impacts the quality of the data that is used for subsequent analyses. Therefore, ETS serves in an oversight capacity in the monitoring of scoring that is done by PEM.

Conclusions: Scoring NAEP Assessments

Although scoring procedures for NAEP assessments were generally consistent with expectations in the *Standards*, concerns were raised about current practices for scoring constructed-response items, particularly the need for better interspersed use of validity papers and the need for an improved system for scoring trend papers. The systematic use of validity papers provides evidence of both consistency and accuracy among scorers. Although, backreading is currently conducted to help ensure quality of the scores, the additional use of validity papers is more consistent with sound measurement practice. In the current system, trend papers are treated differently, potentially influencing the precision and attention raters give to these papers. This could distort the comparability of scores for the trends. Also questions were raised about the role of the second rater's score when that score deviates from the first rater's score. The purpose of the second rater's score is for reporting reliability; thus, ignoring known deviations in scores across raters is contrary to good measurement practices.

Creating Scales and Links and Analyzing Data

In most testing programs, special score scales are developed to aid in the interpretation of test results. The creation of scaled scores can be fairly simplistic (such as, for example, putting the scores on a scale from 0 to 100 with a fixed mean and standard deviation) or very complex involving sophisticated equating methodologies. Due to the use of matrix sampling of items and the administration of different blocks of items to examinees, the creation of scale scores for NAEP assessments is even more complex. In addition to NAEP, some international testing programs (e.g., TIMSS, PISA) use a "plausible values" methodology designed to create full assessment records from incomplete assessment results. This strategy uses additional, conditioning information (e.g., background questions) to predict a student's ability if he or she had taken the full form of the assessment rather than just one of the blocks. Because this methodology is used for so few testing programs, and because it is complex, this adds to the lack of transparency of the scoring and scaling procedures used in the NAEP assessment program. To fulfill an additional stated purpose of NAEP scores of being able to track changes in achievement over time, a multistage linking methodology which involves linking and equating test scales over time is used to support interpretations of results over time across annual assessments. This linking methodology involves the use of common (anchor) items (or questions) across years.

Testing programs provide the technical information that supports the scoring, scaling, linking, and equating procedures. This information should provide evidence of the reliability and validity of intended score interpretations over time.

Relevant *Standards*:

Standard 2.1: For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

Standard 3.22: Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions

for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally.

Standard 4.1: Test documents should provide test users with clear explanations of the meaning and intended interpretations of derived score scales, as well as their limitations.

Standard 4.2: The construction of scales used for reporting scores should be described clearly in test documents.

Standard 4.9: When raw score or derived scores scales are designed for criterion-referenced interpretations, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained.

Standard 4.11: When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions.

Standard 4.13: In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented.

Standard 4.17: Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.

For this audit dimension, two major components were considered. In the first component, we sought information regarding the procedures used for creating the scaled scores and the links to maintain score interpretation. In the second component, we sought evidence of the technical quality of the resulting test scores, including evidence to support reliability and validity of score interpretations. Although AIR-CA and HumRRO have supporting roles for this dimension, the major responsibility for this dimension is with ETS.

Determining NAEP scaled scores involves several critical steps. Because of this complexity, several data quality checks are included throughout the process to ensure that the data are accurate and appropriate. A statistical analysis (called a principal components analysis) that seeks to identify the questions that provide the most predictive information is conducted on the background questions to reduce the number of variables used in subsequent analyses (involving conditioning) to those principal components that summarize at least 90 percent of the variance represented in the full set of background questions. This is done both at the national level and then separately for each state for state-by-state reporting. Because states have differing characteristics, the number of principal components used for the state-by-state analyses can vary substantially, from as small as 100 to as many as over 400. The relative contribution of these variables is also unique to the national or state-by-state analyses. No analyses are done to identify whether there is a common set of background variables across the states. Other strategies could be used to ensure some commonality in the principal components information that are used for state-by-state reporting, such as forced entry of some of the contrasts used in the principal component analyses conducted for the states. Following the creation of these principal components, plausible values methodology is used for the final scaling. This methodology is both complex and controversial. It would be helpful if a more “user friendly” (e.g., simpler) explanation of this process could be prepared and shared with both the psychometric and lay communities. Although the plausible

values methodology is published in the professional literature, the comprehensibility of the approach continues to be elusive as it is applied in so few testing programs. Common items are used in the assessment for linking purposes in order to keep the results on a common scale.

ETS staff also described their procedures for evaluating items for differential item functioning (DIF). Again, DIF is an empirical, statistical procedure to evaluate items (test questions) for potential bias. They also summarized specific instances where items flagged for DIF were removed from NAEP assessments.

As a contractor that provides external quality control for NCES, HumRRO serves only a minor role in this audit dimension. Some of the special studies they have conducted have looked at the replication of the full parameter estimates used in the Item Response Theory (IRT) scaling and replication of Long Term Trend scaling, equating, and conditioning. AIR also serves only a minor role in this dimension. Some of the special studies conducted by AIR have looked at the replication of the full parameter estimates used in the IRT scaling and the potential for other indicators to be used in conditioning variables for scoring.

Conclusions: Creating Scales and Links and Analyzing Data

This dimension, for all practical purposes, is the purview of ETS, which has been creating the scales and links for the NAEP program for many years. Although the methodology used for creating the full data matrix is not without controversy, the plausible values methodology has been reviewed and evaluated in the literature and in previous NAEP evaluations. Therefore, it was not a focus of this audit study. These methods have passed the test of time, and are consistent with those used in similar large scale assessment programs that sample content and student performances (i.e. PISA, TIMSS). The methods used to create scales and links to analyze the data are generally consistent with the *Standards* and sound measurement practices for this type of assessment program.

This page intentionally left blank

Interpreting and Using NAEP Assessment Scores

Writing, Reviewing, and Disseminating Reports and Data

Communicating results in a meaningful and useful manner is obviously important to a successful testing program. A testing program that employs excellent technical and psychometric procedures that produce reliable and valid scores is not a successful program unless the scores can be used and interpreted in a meaningful way by test users. It is the responsibility of the testing program to provide documentation on the technical quality of the results at the time scores are released. Providing this information increases the transparency of the testing program and assists users in understanding the appropriate uses of scores.

Relevant *Standards*:

Standard 5.10: When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Standard 5.12: When group-level information is obtained by aggregating the results of partial tests taken by individuals, validity and reliability should be reported for the level of aggregation at which the results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established.

Standard 6.1: The documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

Standard 6.3: The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standard 15.11: When test results are released to the public or to policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data.

NAGB and NCES play major policy roles in reviewing and disseminating reports of NAEP results, but within the Alliance, ETS has the responsibility for consolidating and preparing many of these materials for dissemination. NAGB, NCES, and ETS work with two contractors to facilitate the release and dissemination of NAEP results: GMRI and Hager Sharp. NAEP state coordinators also help facilitate the interpretation of these results by providing feedback during the review process and assisting in local (state-level) dissemination of results.

In considering this audit dimension, information was sought on the report development process and any evidence of intended interpretations by stakeholders and test users. Information was gathered about the dissemination of results and whether the results were readily available to appropriate audiences in a timely manner. Because of the key role of the Web site in providing NAEP assessment results, information was also sought about the usability and accessibility of the Web-based presentations of NAEP results. The general topic of the utility of NAEP assessment scores and reports was one of the

prioritized areas of the evaluation as a whole. Special studies were conducted to add more information about this dimension. The audit focused on the development and review processes for technical and result-oriented reports. Because the dissemination of NAEP reports falls under the control on NCES, a special interview was conducted with the chief statistician at NCES who oversees the NCES document review policies and procedures.

ETS is working with two additional Alliance contractors on the dissemination of NAEP results: Hager Sharp and GMRI. Although these two contractors have a key role in the dissemination of NAEP reports, ETS has the responsibility for creating the documents for review and subsequent release. Due to the new interactive Web site that allows users to interact with NAEP results in ways that are meaningful to them, ETS has reduced its emphasis on paper printed reports. ETS also works with Westat in providing information about interpretation of NAEP results to the NAEP state coordinators.

In the past, communication about NAEP results was under the auspices of NCES. Although most reports are still released by NCES, NAGB assumed responsibility for initial releases starting in 2004. NAGB is also seeking advice on ways to improve the messaging about NAEP results, hiring their own public relations consulting firm (i.e. Ogilvy). Changing policies about the agency that has primary responsibility for NAEP reporting has created some confusion both within the Alliance and between NCES and NAGB. Further, NCES is the main point of contact with users and ETS may not be consulted when questions are raised about interpretation of NAEP results.

Even though NCES seeks input from Alliance members Hager Sharp and GMRI on format and design, ETS must prepare the text for these firms to use in their preparation of support documents. The audit team urged ETS to conduct usability studies and focus groups to learn information from various user groups about how the information is being interpreted and used. Some research is underway by other Alliance members on report use (e.g., AIR-CA's State Profiles study) and more information about usability is included in the special studies on Utility that are included as part of the full NAEP evaluation.

Because it provides the technology infrastructure for NAEP, GMRI is not responsible for the content in the reports; therefore, it does not play a role in writing or reviewing (for content) the reports. Part of its role in disseminating the information involves verifying the Web site's capability to display and communicate the results of NAEP assessments. GMRI has developed some general criteria that it uses to test the Web site prior to the release of information. These criteria include acceptable functionality, interface usability, browser compatibility, and conformance to NCES's style guidelines.

With respect to the initial release Web site, GMRI provided comments to NAGB and NCES about the potential for user frustration that might be experienced when the information on the initial release site was no longer available after a couple of weeks. This recommendation was considered but not implemented. GMRI also noted that although it may frustrate some users, there is not a consensus within the IT industry about appropriate strategies for managing temporary sites. Therefore, this is an area that will generate further discussion.

Although they are responsible for developing and maintaining the Web sites for NAEP, GMRI has limited control over gathering some of the Web usage information it may need to better inform design or structural decisions. Because the NAEP pages are housed within the NCES site, there may be some confounding of information that GMRI receives from Webtrends (the site usage data collection tool). From the larger dataset that they receive from NCES, GMRI has been able to generate information on monthly traffic flow in terms of page requests. Because GMRI does not have control over how these data are collected, there were some limitations in the interpretation of these data. Some of the NAEP pages did not receive enough hits to make it into the Webtrends.

Within its role of supporting technology infrastructure and disseminating data, GMRI also assists with the development of Web-based tools to be used by visitors to the NAEP Web site. Part of this role is collaborating with NCES and ETS on design and usability of these tools. GMRI carried out a usability study to identify navigation and other issues with a prototype version of the NAEP Data Explorer.

NAGB is responsible for the initial release reports, the Web site for initial release of NAEP results individual state (i.e. State NAEP) and district (i.e. TUDA) reports, special reports, and pilot studies. Other reports that are not special reports are not the responsibility of NAGB. In addition, other reports such as inclusion reports published by NCES are not viewed as initial releases of data and are therefore not the responsibility of NAGB.

In documented NAGB policy (NAGB, 2004a) the Board has listed principles and guidelines for reporting that specify the focus of the reports, the intended audience, rules for reporting subgroup information, and information to be included. This list of policy and guideline statements defines the extent to which NAGB influences the content of the report before the writing begins. In addition, the NAGB policy on 2005 report specifications (NAGB, 2004c) includes reporting requirements that focus on the structure and presentation of different types of results for the reports and the Web sites. NAGB is responsible for reviewing the reports (even at the outline stage) that affords them opportunities to make suggestions for change to the proposed content or framework.

Although NAGB does not appear to be responsible for writing these reports (i.e. the content), they are involved in the extensive, multistage review process.

NAGB is given several opportunities to provide feedback on the reports during the review process:

- 1) Format—NAGB can comment on the proposed format of the report and specifically highlight any ways in which the NAGB policy for reporting is violated.
- 2) Proposed content—NAGB can look at the proposed content, executive summary, and table shells of the report. Comments are gathered from the staff and Board and are sorted into four categories:
 - a. Policy issues (these are non-negotiable changes to be made)
 - b. Strong recommendations
 - c. Questions, needed clarifications
 - d. Editorial comments (grammatical issues)
- 3) Final Proof—The Board has final say on whether or not to release the NAEP reports. To date, they have not held the release of any report and suggested that the only reason why one would be held is in the case of a policy violation. However, the 2005 NAEP 12th Grade Mathematics Assessment has experienced delays in its release.

NAGB is responsible for the dissemination of many NAEP reports and has published a reporting schedule for the 2005 assessments on the Web site (NAGB, 2004b).

NCES is charged with making these many reports understandable. Starting in the 1980's NAEP reports became longer and longer. To deal with this issue, smaller "highlight" reports were created. Given their (NAGB's) interpretation of the change in legislation that involved shifting the responsibility for initial releases of NAEP reports, NAGB has now assumed the role of specifying standards for how the reports shall be prepared. NCES strives to ensure that NAEP reports follow the NCES Statistical Standards (www.nces.ed.gov), but occasionally these standards conflict with NAGB's requests for report specifications. NAGB provides specific content and editorial specifications for these reports (i.e. color, content, framework, and number of pages).

The process outlined below is a revised format for report review by NAGB. At each of the listed phases, NAGB is allowed to review the report materials and provide comments to NCES. Specifically, this process is followed for Web pages, Report Cards, State reports-snapshots, and TUDA (each written for two subjects and three grades).

- 1) Outline stage—ETS provides an outline for each report. This is reviewed by NCES and NAGB.

- 2) Table Shells and Figure Designs—ETS again provides this information that is reviewed by NCES and NAGB.
- 3) Pre-division review—In this phase ETS provides the layout of the report without the data that is then reviewed by NAGB and NCES.
- 4) Center-wide review—This includes two individuals from other divisions and the chief statistician. Once the chief statistician approves the report, the review goes to the commissioner.

The review comments provided to NCES by NAGB form a set of consolidated comments from the board and the staff. This appears to have helped in the review process rather than receiving comments from separate sources. Before 2005, NAGB staff members were allowed to look at reports and staff would make policy comments. Board members were never involved under the previous review process. Now, with the change of policy, NAGB provides much more in-depth comments. Occasionally, NCES will negotiate comments and request changes with NAGB until consensus can be reached. In the six-month review process it is not common to have outside reviewers; however, for other reports produced by the agency, it appears to be more common that independent reviews by external content specialists occur. In addition, because the six-month timeline is so short, these reports do not go through review by the Institute of Education Sciences (IES) because the NCES's chief statistician can sign off on these reports.

NAEP technical reports follow a different review process. Starting with the 2000–01 report, the technical reports will be all Web-based and they are working to build this framework and the core elements. This format is intended to allow for quicker production of the reports. Lack of staffing was mentioned as one reason for the delays in getting these reports out as these are of lower priority in comparison to the other reports and activities that are ongoing. The 2000–01 technical reports were expected to be finalized during the summer of 2005. As of March 2007, these technical reports were not yet released and were still in the development/review process.

As briefly mentioned above, the NAEP chief statistician is also involved in the review process and has substantive responsibilities in the review of NAEP reports. One of the responsibilities is to ensure that reports meet the NCES Statistical Standards. These standards, published in 2002, were created through an extensive process that involved internal staff and external reviewers (NCES, 2002).

The review process for documents produced under NCES is as follows. The first step is a divisional review. For NAEP reports, this means that the reports are reviewed by staff within the assessment division. This divisional review for NAEP is dissimilar to the standard review process used by other divisions in NCES. Second, there is a center review that includes the chief statistician's review along with the assistance of NCES or an external contractor, NESSI, staff who reviews the document based on predetermined criteria. The criteria for this review are contained within a 20-page manual that is used by NCES and NESSI to ensure reports meet NCES standards. For the urgent (six-month) reports, the chief statistician strives to complete the center review process within one to two weeks. The comments from the center review are returned to the division and then shared with the author. The author is then given the opportunity to provide reactions to the comments. The chief statistician receives a summary of all comments sent to the author and the author's reactions to each comment. As noted above, the NAEP (assessment division) review process is different from the review process of other divisions. Whereas other divisions include an initial review by program staff (e.g., program officer), NAEP reports are immediately submitted to the divisionwide review.

The process described above is also followed for the nonurgent reports (e.g., secondary analyses). In addition, after the center review, these reports are sent to IES who conducts both an internal and external review. All comments are consolidated and sent to the reviewer. The issue of the significant lag time in release of NAEP reports (other than the six-month reports) was addressed. The office has recently averaged a 21-day turnaround for the initial review and a 57-day total turnaround

time to completion of the NCES center level review. However, the process by which these reports are passed between agencies often requires reviewers to re-familiarize themselves with reports as this iterative process often involves multiple drafts. In addition, reports that are of lower priority often seem to get “lost” in the process of author’s revisions which can add significant lag time to the process. One specific problem noted was the NAEP technical reports. Because these are perceived as having a lower priority, these reports take the longest to produce. The next technical reports to be released (2000–01) will use a new online format but will also be available in paper format. Although a Web-based presentation of the technical manuals has been discussed, they have been shifted to a lower priority given other concerns in the testing program. Again, the online technical manuals for the 2000–01 and subsequent years were not published as of October 2006.

Based on this evaluation, it became apparent that the reporting task for NAEP is quite substantial. Although there are reports that need to be produced in shorter timeframes, we do understand that there are other reports that are being produced through this program. To illustrate this, we have provided Appendix H that details the volume of reports that have been released in the past year. This Appendix includes publications and products from NAEP since the beginning of this evaluation (October 2004). Each publication is noted with the month and year of release by NCES and is grouped within one of three categories. The Results publications include initial release of results (e.g., Report cards, Snapshot reports), the Technical and Informational reports include any technical reports (e.g., Long Term Trend Technical Report) as well as informational reports (e.g., Education Statistics Quarterly), and the Data Files are all restricted use data files provided by NCES for researchers. In total, since October 2004 there have been 23 reports of results, four data files released, and 30 technical reports published.

As NAEP has gained in national visibility, helping the public understand the results of the assessments has become a larger task. NAEP state coordinators have been important representatives in this process. Specifically, one of the NAEP state coordinators goals is *Data Analysis*; however, their responsibilities here are not related to the operations of NAEP but rather analyses that relate to the dissemination of information. Many of the state coordinators complete the *Data Analysis* goal by reformatting NAEP reports to make them understandable by stakeholders within their state. These reports are designed to highlight findings and data that are important to the state. In addition, several state coordinators reported conducting specific types of analyses such as strand analysis, subgroup exploration, gap analysis, and trend analysis.

NAEP reports are typically provided without interpretation or opinion and the state coordinators are commonly asked by stakeholders within their state to provide meaning of the NAEP results. States want to know the worth of the data to schools and educators. NAEP state coordinators mentioned this being a very interesting aspect of their job; however, some often have inadequate time to address their goal of *Data Analysis*. Several state coordinators reported addressing this goal by developing special reports to be shared at conferences around the state.

Conclusions: Writing, Reviewing, and Disseminating Reports and Data

There are two components of this audit dimension. First, this dimension focuses on the preparation and issuing of NAEP results. ETS has the responsibility for preparing these reports, and the dissemination of the results starts with NAGB, which oversees the release of NAEP results, and with GMRI and Hager Sharp which institute processes to support the utility and ease of use of the results. Because of the critical importance of the interpretation of these reports of NAEP results, a special study is being conducted, within the scope of the evaluation of which this audit is a part, to address the utility of the NAEP reports for various stakeholders. In terms of the preparation and issuing of NAEP results, particularly the initial releases, these have generally met the anticipated timelines. Reports have been made available in electronic form and through various print media sources. The Alliance Contractors

have responded well to these increased pressures to disseminate data. As more of the information is disseminated electronically, the Web site will continue to be an important tool to communicate NAEP results.

The second component of this dimension addresses the review process and the availability of reports that provide the technical information about the NAEP assessment and other supporting documents. This is an area where the current NAEP procedures are out of compliance with the *Standards*. Due to the long delays in getting technical and other reports reviewed and released for public use, limited information is available to support the technical quality of the results. This does not appear to be a fault by the contractors, who are meeting their deadlines for submitting the technical reports for review and release. Instead it appears to be the outgrowth of a multistage review process (i.e. Assessment Division, NCES, and IES) coupled with a limited staff devoted to this part of the assessment program. Although thorough, the review process is limiting the program from providing needed transparency and information to the broader community of users or potential users in a timely manner. The delay in the technical reports, on some level, jeopardizes the integrity of the program by inhibiting the exchange of ideas about the technical adequacy of NAEP.

Setting Achievement Levels

Interpretations of some assessments rely on the use of performance standards, which identify levels of performance on the test that have special interpretative meanings. For example, in educational assessment the score scale is sometimes divided into ranges that support interpretations about the level of performance by students with scores in those score ranges. Student whose scores fall in these score ranges are then classified into a performance category, such as “Basic” or “Advanced”. The process used to identify these score ranges is often called “standard setting” and is used to set the achievement level standards. There are several methods that can be used to set achievement levels, or cut scores, on assessments such as NAEP. In most educational assessment programs, a judgmental process is used for setting achievement level standards that involves expert panelists who have both familiarity with the content and the target population of students being tested. Regardless of the methodology, the resultant cut score recommendation needs to consider the ease or difficulty of the assessment and not be established in a manner that is arbitrary and capricious. It is important that the steps followed in setting achievement levels and the qualifications of the panelist are well documented. However, a cut score decision is ultimately one of the policies that attempt to translate a written description of a performance level into a scale score on the assessment.

Once the achievement level standards for an assessment have been established, statistical equating or linking procedures are then used to adjust the achievement level standards as new assessments are developed using the same Table of Specifications. Although the adjustment may be slight, it is important because the characteristics of two forms of a test will likely be similar, but not identical. Equating, then, allows users to interpret performance on the same scale regardless of the form of an assessment that the student takes.

Relevant *Standards*:

Standard 4.19: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.

Standard 4.20: When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.

Standard 4.21: When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

Our evaluation of the methods used to set achievement levels on NAEP assessments focused heavily on a special study related to a new methodology that was introduced for the 2005 Grade 12 mathematics assessment. Previous NAEP evaluations have been critical of the Item Rating approach that NAGB has used to set achievement levels for other NAEP assessments. These prior evaluations concluded that this procedure was “fundamentally flawed”, in part, because of the inferred cognitive complexity of the task for the panelists (e.g., Pellegrino et al., 1999). However, the documentation provided by NAGB, Loomis and Bourque (2001), and Hambleton et al. (2000), and technical reports by ACT (1995) provide evidence to support their use of the methodology when compared with the relevant standards noted above. These data suggested that qualified panelists engaged in systematic judgments about the likely performance of students at different achievement levels consistent with the Achievement Level Descriptions provided by NAGB. The purpose of including achievement levels as part of NAEP was to aid policymakers in the interpretation of scores beyond a scale score.

Tensions exist between NAGB and NCES about the use of achievement levels on NAEP assessments, in general. NCES’s commissioner has not certified previous achievement levels and these levels have been characterized as “developmental” in reports since their inception. Because the use of achievement levels is ultimately a policy decision, NCES’s resistance to certifying a policy decision rather than reporting an estimated parameter of the population in reports, is understandable. The use of achievement levels is commonplace in educational assessments at the state and local levels and assists in communicating the meaning of scores to policymakers and the public. However, the common language (e.g., Proficient) that is often used across these levels, including NAEP, can also lead to confusion as the definition of performance may differ substantially.

A special study within the full evaluation looked at this dimension of the NAEP assessment program in some depth. Key features that were examined in that study included documentation and analysis of the standard setting procedures, the characteristics of the panelists, the procedural validity, and the external validity of the cut scores. However, because achievement levels have also been established for other NAEP assessments (e.g., Reading), we evaluated the previous achievement levels methodology in the context of the *Standards*. Because there are no “true” cut scores, policy bodies for testing programs play a key role in establishing final achievement levels.

To that end, as the policy body for NAEP, NAGB plays a critical role in crafting the achievement level descriptors that are used in the standard setting process. The actual standard setting process under the current contract is the responsibility of ACT. ACT has subcontracted with Pacific Metrics to undertake the standard setting activities for the 12th grade NAEP mathematics assessment. Validity evidence about the new achievement levels process will be discussed in more depth in the special study on the achievement levels.

The steps in conducting an operational standard setting are generally documented in the technical report for the respective study. After the assessment framework and test specifications committee has decided on content, it is asked to craft achievement level descriptions based on policy. This information is given to those responsible for developing items (to ensure coverage of achievement levels) and those responsible for setting achievement level standards (who will also finalize the achievement level descriptors). During the standard setting process, these achievement level descriptors are revised and are edited to be readable and more easily understood by the public. This means that the achievement level descriptors used in the development of the assessment may change from development to achievement level setting.

ACT/Pacific Metrics was awarded the most recent contract for standard setting work using a new item mapping approach known as “Mapmark” (ACT, 2005a; 2005b; 2005c; Schulz and Mitzel, 2005). One reason NAGB felt it was appropriate to use the Mapmark method was that this method is similar to the Bookmark method that is being used by many state assessment programs. The Mapmark is a test-based standard-setting method (as compared to an Angoff or Item Rating method that was described as item-based). Before implementing this new methodology, NAGB asked for an evaluation of the impact of using the Mapmark method to set the achievement level standards. It was suggested that this new method should be compared to a modified Angoff method referred to as an Item Rating method using the eighth-grade math exam. The results of these studies are documented in the ACT reports shared by NAGB (e.g., ACT, 2005c). One reason that the Board felt it was appropriate to use the Mapmark method was that this method is close to the Bookmark method that is being used by many states

The Mapmark method was developed to take into consideration “domains” or “clusters” of test items. The method was designed to allow panelists to make more informed decisions by providing more comprehensive feedback data. There were two categories of domains that were developed for these studies (1) those developed by NAGB and (2) those developed by ACT (teacher domains, stages in the curriculum, and content domains). These domains were used to help characterize the variety of content that panelists would observe in the assessment.

It is important to note that NAGB’s use of a new standard setting methodology was not a rejection of the method that was used for previous assessments (e.g., the previous achievement levels are still usable for past administrations). Although NAGB did not believe that the public would notice a shift in methods, the research community would be aware of the change. With this shift, the perception may be that the change in the methodology was a response to criticisms of the previous standard setting method. It is important to note that pilot studies conducted prior to the adoption of the new methodology included comparisons with results from the Item Rating method.

Conclusions: Setting Achievement Levels

Two aspects of this dimension were the focus of the audit. First, because new achievement level standards were being set for the 2005 12th-grade mathematics assessment, special attention was devoted to the methodology for setting these achievement standards. Based on our review to date, the new Mapmark methodology appears to meet the *Standards* and show evidence for procedural validity. Additional information about this methodology as well as additional validity evidence is provided in the achievement levels study report as part of the full NAEP evaluation report.

Second, because the achievement levels that were set historically are equated to the newly developed assessment for the grade and content area, consideration was also directed to past standard setting approaches. Although previous NAEP evaluations have been critical of the Item Rating approach that NAGB previously used to set achievement levels for other NAEP assessments, based on the *Standards*, there does not appear to be a rationale for considering the previous method as being unsound psychometrically. Therefore, both the past and current methods for setting achievement levels for NAEP appear to be consistent with sound psychometric practices. Both methods provide information from different sources of validity evidence that informs the resultant policy decision.

Because NAGB and NCES have taken differing viewpoints on this topic, it is difficult for users to know how much confidence to place on the reported achievement levels. This is a topic that highlights the different roles of these agencies. As the agency responsible for the operations of NAEP, NCES’s role is to estimate the scale scores for the variety of assessments that are part of the program. However, as the policy body, NAGB is responsible for defining policy. Because the cut scores that define achievement levels are ultimately a policy decision, they are likely within the scope of NAGB’s responsibilities.

Improving NAEP Assessments

Because NAEP is a long-standing and evolving assessment program, this audit dimension sought to draw attention to the need for continuous monitoring, review, and renewal of NAEP's assessment program. The focus of this dimension was twofold: looking backward, to ensure that the documentation needed to support decisions about the program were in place, and looking forward, to enable the program to remain current with new developments in the assessment community and to respond to evolving needs of stakeholders for NAEP results.

Relevant *Standards*:

Standard 3.25: A test should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may lower the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.

Standard 6.13: When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful additional information or cautions.

External efforts to renew and improve the system were observed in multiple sources. For example, HumRRO's Quality Assurance (QA) contract provides critical and timely information about areas where improvements in practice, policy, and procedures are advisable. Another external source of potential evidence can be seen in the Secondary Analysis Grants (SAG) program. This program of research provides researchers access to NAEP data to conduct a range of studies, many of which provide information useful to renewing and improving the NAEP assessment program. A third external source of evidence comes from the NAEP Consortium's response to recommendations from prior external evaluations.

Within the NAEP Consortium, there are also multiple indicators that research intended to improve the system is being conducted. NAGB and NCES often initiate ideas that are studied internally or through contracts. For example one approach is to maintain advisory panels that provide input on policy and practice issues. One of these advisory groups is the NAEP Validity Studies (NVS) Panel. This group, under contract with AIR-CA, develops and conducts a series of studies related to a research agenda that the group develops. Other examples include the NAEP Alliance contractors that generally engage in ongoing programs of research aimed at identifying improvements to current practice and procedures within their respective roles in the lifecycle.

External evidence

Quality Assurance contract

HumRRO's role in the Quality Assurance (QA) contract can be viewed as one strategy for providing a measure of external quality control and serves as a potential means for renewing and improving the assessment program. HumRRO's work focuses on the quality of the current assessment design, development, delivery, scoring, and reporting. As one of the initial activities within the QA contract, HumRRO identified areas where problems existed previously and their resolution strategies helped to inform procedures for future program design and decisions.

One of the primary activities within the QA contract is to conduct an annual review of contractor Quality Control (QC) plans. These plans serve both an immediate need to ensure quality control through the assessment process and have the potential to provide information that would serve for assessment renewal and improvement. Although the current contract tends to prioritize the extant conditions that support the assessment program, with modest adjustments, these procedures could potentially be adapted by NCES to inform more systematic improvement and renewal of the assessment.

Another HumRRO activity is to conduct site visits that are designed to ensure that contractors comply with their quality control plans. However, these reviews also provide opportunities to gather systematic information about where the system is working and where it needs adjustments. The documentation from the site visits could provide information about areas for assessment improvement, particularly regarding the process for administering the assessment. Currently, the information gathered from these site visits is not systematically being accumulated and evaluated for this purpose, but it could serve as a rich source for systemic program improvement. This effort could be enhanced through a more comprehensive quality control plan for the site visits to ensure that the quality control dimensions across the contractors are considered through the site visit design.

HumRRO also conducts special studies to examine means and mechanisms for assessment renewal and improvement. Some of these studies have evaluated anomalies that have appeared in the data, specific concerns about possible program issues, mechanisms for verifying the accuracy of the reporting of student demographic information, examining motivational issues related to 12th-grade assessments, and improvement of current practices in monitoring the quality of scoring of constructed response questions. Although there does not appear to be a comprehensive plan for the special studies program, NCES could potentially use this part of the program to conduct studies that could more directly inform assessment renewal and improvement.

Secondary Analysis Grant program

The NAEP SAG program was started in 1992 by NCES and serves as another potential source of information to improve the NAEP program. From 1992 until 2003 NCES administered the entire program. In 2003 when IES assumed responsibility for the program, changes were made to the program. Specifically, reviews of grant applications were outsourced and a standardized review process was implemented for all grants awarded under IES.

The purpose of the program is to:

“contribute to improvement of student learning and achievement by (a) identifying programs, policies, and practices that are potentially effective for improving academic outcomes, as well as mediators and moderators of the effects of these programs, policies, and practices, and (b) developing tools or procedures to assist NAEP users in the analysis, interpretation and reporting of state- and district-level NAEP results or to improve precision in the estimation and reporting of NAEP results.” (NAEP SAG call for proposals, 2005, p. 4).

In 1998, NAGB assumed responsibilities for setting priorities for the SAG program in collaboration with the NCES Commissioner. Since 1998, the same priorities have been in place with only some rewording changes made in 2003 and one priority was combined with another. Currently, the four priorities are:

- Projects that use NAEP achievement data alone or in combination with other data sets to assist policymakers and educators in the educational improvement process.
- Projects designed to assist NAEP users in the analysis, interpretation and reporting of state and district level NAEP results.

- Projects that include the development of methodological or analytical procedures which improve precision in the estimation and reporting of NAEP group and subgroup results.
- Projects to analyze and report data using statistical software developed by the project to permit more advanced analytic techniques to be readily applied to NAEP data. (NAGB, 2005).

Currently, the only way that these projects are disseminated is through the NCES Working Paper series on the NAEP Web site. Reports are placed on the site after they have passed divisional review and been approved by NCES.

Prior evaluations

Another element in evaluating renewal and improvement efforts of the NAEP assessment program includes its responsiveness to findings put forth in previous evaluations of NAEP. Linn (2004) describes these previous evaluations and the recommendations that emerged from each. The most recent evaluation of NAEP by the National Research Council (NRC; Pellegrino, Jones, and Mitchell, 1999) offered five recommendations for the program. To illustrate the changing nature of the program, we observed evidence that begins to directly respond to three of the recommendations. These are briefly described here with the evidence of responsiveness.

One recommendation from the 1999 evaluation of NAEP recommended streamlining the sampling and administration plans. Specifically, the evaluation called for NAEP to “reduce the number of independent large-scale data collections while maintaining trend lines, periodically updating frameworks, and providing accurate national and state-level estimates of academic achievement” (Pellegrino et al., 1999, p. 56). Within this recommendation, the 1999 evaluators were asking for a more focused design that eliminated any unnecessary efforts or expenditures in the sampling, administration, and scoring processes. The current evaluation found evidence that this recommendation is starting to be addressed.

Specifically, in 2002 the sampling design was revised to reflect a combined sample of schools for the Main and State NAEP. The Main NAEP sample of students is a subset of the combined State NAEP samples as well as additional Main NAEP samples from states that did not participate in State NAEP (NCES, 2006). It was a logical move to combine these two samples as the Main and State NAEP assessments are based on the same assessment frameworks and items. However, the Trend NAEP assessment is different.

As the intent of the Trend NAEP assessment is to provide a more long-term measure of change in educational progress, the purpose of this assessment appears to be quite different from the Main and State NAEP assessments. Specifically, there is no framework for the Trend NAEP assessment; rather, new items that are written to update the assessment are designed to measure the specific skills previously measured by retired items. Therefore, the assessment frameworks for the Main and State NAEP assessment programs might look very different if an assessment framework were to be developed for the Trend NAEP assessments. Thus, the Trend NAEP assessment remains a separate administration. Although the Trend NAEP assessment is still unique, a degree of streamlining has occurred.

Previous evaluators also recommended that the NAEP consortium increase the level of participation of students with disabilities and ELL to better represent the full student population. In addition, the 1999 evaluation noted the inconsistency in identification and inclusion of these students. The issue of identification and inclusion of students with disabilities and ELL is one that has received some attention in NAEP in recent years and is still an ongoing issue as noted in the current evaluation’s sampling section. Under the current model, states have the authority to (1) identify students as having a disability or as ELL, and (2) determine who will participate in the NAEP assessment. Through this evaluation, it became apparent that there were potential threats to validity as different state policies impact the sampling frame, selection of students, administration procedures, scoring, and interpretation

of scores. State-by-state comparisons are tenuous when different policies for inclusion and accommodations are in place for this targeted group of students.

States that strive to include more students from these populations may be concerned that their overall results suffer as a consequence in comparison to other states. In response to these concerns, both NAGB and NCES reported initiatives to standardize the system for identification and inclusion of students within these populations across states. Specifically, each of the two organizations reported working on a decision tree for this step during the site visit. NAGB later provided the evaluation team with a draft of its decision tree that was shared with the Board and, according to the NAGB meeting summary from August 2004, NCES was incorporating a pilot test of this decision tree with the 2005 assessment to determine if it led to an improvement in the system. A report was prepared for NAGB (Spurlock, 2006) about the implementation of this program.

Feedback was also obtained about any additional components needed to complete the decision tree (as judged by those implementing the tree). However, it was noted that this was used with a Main NAEP U.S. History and Civics exam in 2006—an exam that has not been administered since 2001 and thus an assessment of the improvement to this system was not possible. The decision trees (and accompanying questionnaires) have been revised for the 2007 assessment (Main and State, Reading and Mathematics) at which time a comparison with the 2005 exclusion rates will be possible. Although uniform policies for identification of SWD and ELL and administration accommodations may be desirable, NAGB and NCES cannot mandate these because they may interfere with state-specific laws or regulations. The challenge of uniformly implementing these policies nationally remains an ongoing challenge.

Another recommendation from the 1999 evaluation was that the standard setting process used for setting achievement levels (Angoff-based Item Rating) should be replaced with a method that is less cognitively complex. With respect to the procedures to set achievement levels, NAGB has responded to this recommendation by exploring a new methodology. The standard setting methodology for the 12th grade mathematics assessments was conducted using the Mapmark methodology (Schulz and Mitzel, 2005). Based on the first use of this methodology, there has been discussion about its use with other NAEP assessments in future standard settings. The use of this new methodology is the focus of a special study within the full evaluation.

Internal evidence

NAEP Advisory Panels

The DAC represents one of the standing advisory groups that provide input on NAEP. Funds are included in the NAEP contract for a “dedicated” research program within ETS focused on NAEP. This NAEP program directly relates to improving and renewing the assessment and is accomplished through two different types of research: one directed at solving and resolving immediate operational procedures and processes and a second one that takes a longer view of assessment improvements. Funding differs across these two types of research with the immediate and short-term projects getting approved without full NCES involvement. However, more comprehensive research projects require endorsement by NCES; and therefore, must go through a much more thorough review with the Department of Education’s Contracts Office. Research projects emerge from operational staff members as well as from the DAC. Projects span different operational activities and include such studies as an Item Attribute Study that emerged from test development, ways to improve cross grade scaling, a long term bridge study, and an analysis of the impact of changes implemented in the 2003 NAEP design. Additional research projects have also included an Oral Reading Study and two studies considering the use of online assessments (Math Online and Writing Online).

Although not directly connected with AIR-DC’s Alliance contract for NAEP, AIR has an indirect role in the renewal and improvement of NAEP through AIR-CA’s separate contract with NCES

to coordinate the NVS Panel. This panel has developed an agenda for validity research and members of the panel generally carry out this work. Although this panel is generally able to directly disseminate research, its contract and decision-making authority about which NVS studies to fund rest with NCES. NVS disseminates technical reports outside the typical NAEP review process; thus studies may enter the public domain more quickly. These research efforts are often distributed through professional conferences (e.g., AERA, NCME, Council of Chief State School Officers—CCSSO) or published directly by AIR-CA.

NAGB, NCES, and Alliance Contractors

Because of the testing cycle, the operational system does not currently have a way to directly incorporate research innovations into practice without disrupting the system. Some of this is probably because of shortened reporting requirements for reading and mathematics because of *NCLB*. There does not appear to be a decision-making process for reviewing or evaluating new ideas or a budget built into operational practice for planned change. Innovations are recommended through technical reports or research studies, but may not be acted upon. For example, AIR-CA suggested a method for determining how to interpret state assessment achievement levels on the NAEP scale (McLaughlin et al., 2005). ETS suggested an alternative strategy for doing this. The process for reconciling these differences is slow at best and there does not appear to be a well articulated policy for how these new methodologies are considered and then implemented.

Other examples include efforts to operationally implement full population estimates. Analyses in 1998 suggested that observed NAEP gains were due to the increasing rates of exclusions. This is a topic that was formally proposed by AIR-CA in 2002, but the system has been slow to implement these changes. It was also noted that HumRRO conducted an evaluation of the methodology and was to compare AIR-CA's method with an alternative method proposed by ETS. To date, there does not appear to have been an alternative method submitted by ETS to HumRRO for the comparative evaluation. Some form of independent arbitration of these issues similar to the role the HumRRO was to play in this issue would assist NCES in considering competing innovation proposals from contractors within the Alliance.

Promoting Innovation

Many innovations in the NAEP program have involved changes in technology that allow systems that were not possible earlier. We observed evidence from most Alliance contractors and agencies in the NAEP Consortium about efforts to promote innovation in the program. Efforts to promote innovation by the NAEP Alliance contractors are also encouraged through incentives in their contracts with NCES. NAGB has also explored new strategies related to incorporating technology into the NAEP assessment. These discussions are in response to the growing use of technology in education.

Although it is important that NAEP not be locked into one form of administration, issues may preclude transitioning NAEP into a computer-based test. For example, science assessments that require hands-on demonstrations or procedures may not be easily computerized. However, anticipating the growing capacity of technology-based assessment, ETS is currently conducting field tests of computer interactive items for science with a plan for implementation in 2009 or 2011. On the other hand, many students who are learning to write on computers may have difficulty in the future completing quality work on a paper and pencil format test. NAGB acknowledged that many states are ahead of NAEP in incorporating technology into their educational assessments. One additional area that NAGB is exploring looks at incorporating technology into the frameworks by considering measuring technology literacy based on frameworks from the National Academy of Engineering. It should also be noted that changes to the assessment mode may necessitate revisions in the assessment frameworks to ensure alignment. Some additional examples from the Consortium are described below.

Westat also mentioned the need for NAEP to look more closely at computer-based assessment but also noted some of the potential practical challenges to the program. Although historically the hurdles have been perceived as great, as barriers to access and computer literacy are reduced, this is a direction for the program to strongly consider. Some of the challenges to integrating technology into NAEP would include the logistics of computer administration (number of computers needed for administration), student verification (e.g., biometric screening), standardization of the testing environment, technology literacy of field staff, and systems for technology (e.g., security, firewalls). A related challenge to dramatically changing technology would be to provide training to the large contingent of field staff, many of whom may not be as familiar with current technology.

Technology innovations have also played an increasing role in NAEP with the transition of many processes and products to electronic, particularly Web-based formats. Many of the innovations for NAEP have been achieved through this avenue over the past four to five years. GMRI's role in this innovation has been evident through the IMS system, Web CMS, and the variety of Web sites that they have developed. GMRI is currently in the process of transitioning the public Web site for the NAEP Network and updating IMS to a new version (3.0) that will include additional features and functionality. These activities continue to evolve.

PEM has also implemented several software and technological innovations that provide support for the ongoing integrity and quality of NAEP assessments. These include systemic software and documentation systems, clear articulation of specifications for NAEP activities under the auspices of PEM, and the development and implementation of technological solutions to ensure compliance with packaging specifications, shipment and document receipt, and scanning methodologies. Due to the complexities of the NAEP assessment design, and the increased need for ensuring tracking of document shipping and receiving, these systems become more essential.

NCES also provided an operational example not directly related to technology of using interspersed trend papers in the constructed response writing assessments. Writing samples from prior administrations of the NAEP assessments (for calibration) are typically scored before the live scoring of the current administration. It would be better to do simultaneous scoring but before this can happen, a study is needed to determine the impact of this change. This is typically the process for implementing new methodology; a pilot study is conducted (typically during a year when there are not large data collection needs). Also, possible topics are sent to the DAC (ETS), NVS (AIR-CA), and the Quality Assurance Panel (QAP, HumRRO) for review in advance. An innovation clause was put into the Alliance contract to encourage innovation and competition among contractors.

Conclusions: Improving NAEP assessments

All of the members of the NAEP Alliance have systems in place to inform the assessment improvement process. The results of these efforts have resulted in changes to the system (e.g., combined samples for Main and State NAEP; allowing accommodations that are determined appropriate under the IDEA). The various programs of research appear to be an area of strength for the NAEP assessment program.

Contractors in the NAEP Alliance are generally required to be reactive rather than proactive because they are responding to a scope of work that is predefined with some flexibility expected. Therefore, it is often difficult for them to know when they can provide input on proposed changes in the process. Related to the shift in some of the responsibilities for the program, NAGB's policy changes have also led to their increased involvement in the details of the project rather than just at the policy level. It is often challenging for the operational staff to respond to requests for changes or special studies when a particular NAGB committee (e.g., COSDAM) or board member recommending these changes may not appreciate the operational difficulties of the request or how it relates to the contractual responsibilities for the organization.

The NAEP program has been responsive to previous evaluations. In response to the 1999 evaluation of NAEP, changes have been made to the procedures used for setting achievement level cut scores, more attention has been paid by states in addressing issues of inclusion, and efforts to streamline the sampling procedures are underway.

Further, several instances of inclusion of technology into the NAEP program were found. Although there are no immediate plans to administer the assessments using technology, a research effort is examining the feasibility of such a change in the delivery of the NAEP assessments. Technology innovations have improved communications within the NAEP consortium, allowed for better tracking of NAEP assessment in the field, and provided quicker and more interactive access to NAEP results.

One of the challenges to changes or improvements in NAEP's methodologies is a rationale that the need to maintain the validity of the interpretation of trend scores is a compelling reason to retain the status quo. If there are changes to the assessment, the interpretation of the trend data (short or long term) may be questioned. *NCLB* has helped facilitate some changes, but reading and mathematics are being kept together because of their role in the *NCLB* legislation. Because one of the stated purposes of NAEP is to monitor progress over time, resetting baselines too frequently would interfere with this purpose. However, this rationale cannot be used indefinitely when changes in methodology would improve the program and the validity of the results.

This page intentionally left blank

Summary of Key Findings

Based on the information we were able to gather during our review, it appears that most operational components of the NAEP assessment program were functioning well and were in compliance with sound measurement practices and with the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999). There were a few key exceptions, however. The major exceptions that have the potential to threaten the validity of the program were the absence of a formal validity framework to organize and prioritize evidence to support that validity of score interpretations and uses, and the lack of current technical documentations and reports to support the psychometric properties of the NAEP assessment program.

The key findings are organized into two sections. The first presents key findings that were identified as strengths of the program. The second set identifies areas for improvement. Within each section, the findings are organized by importance.

Key Findings Related to Strengths of the Program

Key Finding 1: Main and State NAEP Assessments in Reading and Mathematics are developed, implemented, and maintained in ways that are generally consistent with widely accepted professional assessment standards.

Through this evaluation we were able to explore many aspects of the NAEP program described in the previous section. Except for a few noteworthy exceptions, the methods and procedures used for the Main and State NAEP Assessments in Reading and Mathematics were found to be in compliance with these widely accepted professional assessment standards. This compliance was noted throughout the development, implementation, and maintenance of the program.

The processes used for creating the assessment frameworks are firmly grounded in policy and the review and revision procedures were consistent with sound measurement practices. Further, we found that the methods used by the Alliance contractors to develop and review the NAEP assessment questions are consistent with the *Standards* and follow sound measurement practices. The methods used for field-testing items appear to be technically and psychometrically sound as they involve using embedded field test blocks within the operational administration. This helps to ensure accuracy of the field test data.

We found that systems are in place to support communications and cooperation among the contractors preparing for and conducting the administration. This is an important feature as the administration of the NAEP assessments relies on the coordinated effort of multiple contractors and NAEP state coordinators. We found that the electronic monitoring systems for tracking the materials is a strength of this process as it helped with the administration process and maintaining security of the test materials. Overall, the scoring procedures are generally compliant with the *Standards*; however, there is one exception that is noted in a later finding. In addition, although there is not agreement in the measurement field about which methodologies are the most statistically sound for estimating student performance on a full assessment when they only take a sample of the items, the procedures used for the NAEP assessments are consistent with those used in several other large-scale, international assessments and are generally consistent with the *Standards*. Overall, the psychometric characteristics of the NAEP assessment scores (e.g., reliability, standard error) all support the technical quality of the results. The procedures and timelines for the initial release of NAEP results are in compliance with the *Standards* and the NAEP Alliance responded well to the increased pressure to disseminate results and data in a timely and user-friendly fashion.

We found that there are ample opportunities in the NAEP program for gathering information to support renewal and innovations through several research programs that are a part of the NAEP system.

The topics of these projects span the NAEP assessment program. However, we are concerned that these opportunities are neither systematic nor integrated—this is detailed in a subsequent finding.

Although the majority of the processes in the NAEP system were found to be compliant with professional accepted standards, this evaluation of the psychometric (i.e. technical) quality is limited for two reasons. First, the *Standards* clearly specify that evidence of psychometric quality does not exist in a vacuum. Psychometric quality is related specifically to the defined, intended uses and purposes of the assessment. The intended scope and uses of NAEP assessment results are only defined broadly, leaving room for confusion and lack of clarity about which uses and interpretations are intended and which ones are not. Second, our review of technical criteria was limited to the available NAEP technical manuals (e.g., 2003 NAEP Technical Manual) and some of these conclusions were made based on assumptions drawn from dated material about the NAEP program.

Key Finding 2: Methodologies to establish achievement levels were generally consistent with the expectations of the Standards.

The process of setting achievement levels on NAEP assessments has been both highly criticized (e.g., Shepard, Glaser, Linn, and Bohrnstedt, 1993; U.S. General Accounting Office, 1993) and defended (e.g., Hambleton et al. 2000; Loomis and Bourque, 2001). Two prior evaluations described the NAEP standard setting as “fundamentally flawed” (Shepard et al., 1993; Pellegrino et al., 1999); however some reactions to those evaluations from standard setting researchers were very critical.

These findings are related to the congressional question about the validity and utility of NAEP achievement levels. The *Standards* (AERA, APA, and NCME, 1999) provide guidance on appropriate practice with respect to setting achievement levels (sometimes called standard setting). For example, *Standard 4.19* suggests, “When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented” (p. 59). Also, *Standard 4.20* indicates, “When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria” (p. 60). With respect to the judgmental process, *Standard 4.21* suggests, “. . . The judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way” (p. 60). Within the audit, we reviewed information from the previous methodology used by NAGB to establish achievement levels.

Our findings revealed that one of NAGB’s purposes for developing achievement levels was to assist policymakers and other stakeholders in their ability to interpret NAEP scale scores. To facilitate these activities NAGB also developed Achievement Level Descriptions (ALDs) that provide broad policy definitions of what students should know and be able to do at a given level. These ALDs are then applied to the respective content in more depth during the processes that establish the achievement levels. For these studies, panelists are selected who have content knowledge, some familiarity with the target population of students who would be eligible to take the assessment, and also represent different education stakeholder communities and the public. The results of these activities ultimately represent a policy decision that is within the scope of NAGB’s responsibilities. As is the case with most policy decisions, there is an element of judgment that goes into the final decision. However, in education these types of value-based decisions are also made at the state level (e.g., levels of student proficiency), in a classroom (e.g., assigning grades of A, B, C, D, F), and with individual students (e.g., what is the best instructional strategy to help this student succeed). Given the controversy surrounding this topic, a special study within the full evaluation also reviewed a newly employed method for the 2005 Grade 12 NAEP Mathematics assessment.

Based on the information we gathered during the site visits and through the technical documentation, it appears that the methodologies used to set NAEP achievement levels generally comply with professional technical standards. In particular, there is clear documentation on the rationale

and procedures used for setting the achievement levels. The new methodology applied with the Grade 12 Mathematics assessment had features that were designed to aid the panelists in making their judgments in a manner that is consistent with their knowledge and experience.

Key Finding 3: Current structure of NAEP Alliance contracts facilitate cooperation and communication among contractors.

One of the notable strengths of the NAEP program is the organizational and contractual structure of the contractors responsible for NAEP assessment operations (i.e., the NAEP Alliance). Under the new procurement model that began in 2002, previous subcontractor relationships were changed to direct relationships between contractors and NCES. One characteristic of the change was the establishment of a contract for Alliance coordination to facilitate activities among NAEP contractors. Another feature of the contract is the use of built-in incentives for the members of the Alliance to meet mutually beneficial goals and timelines. This facilitates an atmosphere of cooperation as all contractors benefit when the system is working and all lose out on financial incentives if the system strays from critical path timelines and deliverables.

An additional example of contracts that helped to ensure effective and efficient operations under the new procurement model was the establishment of the Quality Assurance contract that was designed to provide external staffing and support for NCES to monitoring the quality of the NAEP Alliance and operations.

A related strength is the observed communication among Alliance contractors. Within the NAEP Alliance one of the strategies to support this strength is the IMS which facilitates communication among contractors regarding progress, timelines, and discussion and resolution of problems. The features of this online tool provide a common language and structure to the Alliance when integrating systems from different organizations. The IMS also allows for greater decentralization of key personnel because it was developed as secure, Web-based solution and provides a forum for contractors to discuss issues or problems that arise.

Key Finding 4: Psychometric characteristics of NAEP assessment scores are consistent with professional standards for testing.

Our review of technical criteria was limited to the available NAEP technical manuals (e.g., 2003 NAEP Technical Manual) and some of these conclusions were made based on assumptions drawn from dated material about the NAEP program. The technical quality reported in that report provided strong and supportive evidence of technical quality, especially with regard to estimates of score reliability and standard errors of measurement. These technical characteristics support confidence in the scores. This document also provided information about procedures used to ensure the assessments were fair to protected groups through analysis of differential item functioning and item reviews for biasing features. We anticipate that when the technical information is available for the current assessment they will report equally strong evidence of psychometric quality and provide even more evidence of how these assessments comply with the *Standards*. This conclusion is drawn, in part, from historical reports that have been released documenting the NAEP program.

Key Findings Related to Areas for Improvement

Key Finding 5: Intended uses of NAEP assessment scores were not clearly defined.

This finding relates to a critical need for all assessment programs: providing a clear definition of the intended and unintended uses of scores from their assessments. The *Standards for Educational and*

Psychological Testing (AERA, APA, and NCME, 1999) has, in the first chapter, specific expectations of test publishers regarding defining intended uses of test scores and the validity evidence needed to support them. For example, *Standard 1.1* notes that a rationale needs to be presented for each recommended interpretation or use of test scores. Because no test is a gold standard (i.e. valid) for all purposes and all situations, *Standard 1.2* specifies that test developers clearly articulate the intended interpretations and uses of test scores. Because the potential for misuse of assessment data and the resulting consequences are critically important for NAEP, unintended uses of scores are also important to clarify for potential stakeholders. *Standard 1.4* indicates that if a test is being used in a way for which it has not been validated, users need to justify the new use and collect new evidence if necessary. This finding responds to the first congressional question that asked whether NAEP assessments were meeting professionally adopted standards.

The current uses of NAEP are broadly defined by legislation leaving the actual uses open to a range of interpretations. Congress and the wide range of stakeholders may be using NAEP scores for purposes that are not supported by validity evidence. Understanding and clarifying those intended and unintended uses will assist NAGB, NCES, and key stakeholders develop a validity framework for the program and then prioritize validity research efforts to target those intended uses that are most critical to those users. It is important to note that validity research opportunities occur multiple times across many of the NAEP contractors, not the least of which is the NAEP Validity Studies (NVS) Panel that operates under a contract with AIR-CA office. In addition to these efforts, five additional sources of operational validity evidence were cited by NCES: NAEP's DAC, TOC opportunities, assessment development processes, NESSI, and the NAEP SAG program. Research is also funded through separate programs within HumRRO, AIR-CA office (e.g., state analysis contract), and ETS. Our review of several of these research projects suggests that they have the potential to provide critical information that could support the intended uses of NAEP scores and be used in the continual development and refinement of NAEP. However, because specific, intended uses are not currently defined, there is not a transparent validity framework that organizes and prioritizes studies conducted through these various research efforts. This is a lost opportunity to inform, engage, and provide targeted information pertaining to such a validity structure to communicate the strengths of the program and its uses to policymakers.

Key Finding 6: Lengthy review processes limit the availability and utility of NAEP technical manuals and reports.

The protocol for review and dissemination of NAEP-produced technical manuals and reports that document the program's activities is extensive, and in many ways critical to ensuring that NAEP publications are accurate both technically and factually. The review process includes multiple reviews by individuals with different areas of expertise—the specific process is different depending on the type of document being prepared for release. However, because of such an extensive and thorough review process, the outcome is that many important NAEP related documents are not available, therefore missing the opportunity to share high quality, technically and factually accurate information about the NAEP program. Given their role as the agency responsible for program operations, there are more reports that go through the review process at NCES; however, NAGB's review and dissemination practices are also subsumed within this finding. This finding also relates to the first congressional question regarding the program's adherence to professional testing standards.

To highlight this problem, we note that the most recently released technical manual that could be reviewed for this NAEP evaluation was the 1999 Long Term Trend technical report that was released in April 2005. Although we were provided access to Web-based versions of draft technical reports from 2000–03, it is unreasonable that technical documents for assessments that were administered and results disseminated in the years 2000–05 should still be under review. One contributing factor to this delay was that these reports were given a lower prioritization as the focus was primarily on the six-month

reporting requirements thus causing many of the delays in the release of technical documentation. Although we understand the burden presented by the six-month reporting requirement, the lack of available technical documentation violates professional expectations. Another illustration of this timeline is NAGB's initial release of the 2005 NAEP 12th Grade Reading and Mathematics assessment results. These initial releases did not occur until Feb. 22, 2007.

As with the finding of a lack of clearly defined intended uses of NAEP assessment scores, the *Standards* (AERA, APA, and NCME, 1999) expect testing programs to provide documentation for their program(s). For example, *Standard 6.1* suggests that test documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use. In addition, *Standard 6.3* indicates that this documentation should include the rationale for the test, recommended uses, support for such uses, and information that assists in score interpretations. Furthermore, when reasonably anticipated misuses of a test can be anticipated, cautions against misuse should be specified. Although some lag time may be expected due to a comprehensive review process, the current timeline for the release of technical documentation extends beyond what a large-scale testing program should tolerate, and is in violation of the *Standards*.

Key Finding 7: NCES's Assessment Division is understaffed to respond to current demands of the NAEP assessment program.

The NAEP assessment program relies on a series of interactions among the numerous organizations and agencies involved in the development, administration, and dissemination of NAEP assessments and results (See Figure 1). NCES's Assessment Division staff members play a number of roles in the lifecycle. Most important, they oversee the work and deliverables that the Alliance contractors produce. The CORs at NCES are also responsible for facilitating communication among the NAEP contractors and those external to NAEP (e.g., secretary of education, policymakers, evaluators) and assisting in resolving any issues that arise. The Assessment Division of NCES has 20 full-time employees. This is a small staff when compared with other divisions within NCES that have similar budgets but 80 or more full-time employees. Currently, the Assessment Division staff members oversee the work of approximately 1,300 permanent and temporary employees working for various NAEP contractors. Although more than half of these employees are involved primarily in the administration of NAEP assessments, this number is large considering the number of staff within the Assessment Division and responsibilities they have in terms of overseeing quality control procedures for these contractors. Although not directly related to any one congressional question, the capacity for organizations within the NAEP Consortium to respond to the needs of the program s indirectly related to all of the questions mandated in the evaluation legislation.

In addition, as the operations agency for NAEP, the NCES Assessment staff members are responsible for responding to requests for information from multiple stakeholders and responding to questions or inquiries about NAEP results or the proper interpretation of these results. NCES also needs to maintain a close relationship with NAGB to provide input and respond to policies that impact the program's operational activities. The Assessment Division staff members also assume responsibility for reviewing and disseminating technical reports that document program activities (See also Key Finding #6). After noting the many responsibilities of the Assessment Division's staff, it was apparent to the evaluation team that this part of NAEP is dangerously understaffed to respond to these increasing multiple program needs.

Key Finding 8: Some current uses of the NAEP assessments may not be accounted for in the current sampling plan.

Sampling procedures represent an important component in the NAEP assessment program that has a long tradition of driving advances in survey technology. Many of the survey and weighting procedures now used are adequate and consistent with generally accepted methods in sampling. However, the intended uses of NAEP assessments influence how the sampling design is developed and implemented. For example, collecting representative data for the nation requires a different sampling frame than collecting representative data for a state or an urban school district. The sampling frame also extends to student groups. Although this makes intuitive sense, as the intended uses and the policy contexts for NAEP assessment scores are clarified, further evaluation of current sampling practices are necessary. Some of these policy considerations that are unique to NAEP sampling methods are described here and directly relate to the congressional question regarding whether NAEP assessments were conducted as a random sample.

First, appropriate accommodations are expected to be provided to students who require them. Two subgroups of students are most affected by this, specifically SWD and ELL. On the surface, new regulations would appear to lead to an increase in the overall percentage of students included in the assessment as well as the consistency across states in student inclusion rates. However, different inclusion rates and cross-state consistency remain a problem. States differ in their rates of exclusion and also in the accommodations they provide to special needs students who were not excluded. Thus, even included students may have had incomparable test experiences in different states. Differential exclusion rates threaten any state-by-state comparisons.

Second, factors that reduce the initial sample, specifically school and student nonresponse and refusal to participate, represent a potential significant threat to the validity of NAEP assessment scores. Although not directly addressed in the legislation, the *No Child Left Behind (NCLB)* legislation has raised both NAEP's visibility and discussions about intended uses (e.g., comparisons across states). At the same time, *NCLB* has changed the context in which NAEP operates and may indirectly change the nature of student and school nonresponse in NAEP assessments.

Third, state samples must be adequate in size and representativeness to provide reliable estimation of performance. Estimation at the state level has traditionally required sample sizes of about 2,500 students from approximately 100 schools per subject area assessment. Because the specific intended uses of NAEP assessments are not clearly defined (See Finding #5), policymakers' interest in NAEP scores often does not stop at the national or state level for all students. For example reporting is also required for historically prioritized student subgroups (e.g., ethnicity, lunch program status, language proficiency, and student disability). NAEP has traditionally taken steps to oversample students in some key subgroups (e.g., by sampling schools with larger representation of blacks and Hispanics at double the rate of other schools). Today, many states are seeing significant demographic changes; furthermore, demographic characteristics differ substantially from state to state. At the same time, some of the most significant data problems faced by NAEP involve missing Title I data and the representation of these students, uncertain National Student Lunch Program data, and problems with some schools' identifications of racial/ethnic status. All of these issues can affect sampling via less accurate sampling frames and the incomparability of results over time.

Fourth, several schools and districts are sampled with certainty or near certainty across multiple NAEP assessments. As such, what appears to be a random sample in a given year may be more systematic when considered across multiple NAEP administrations. Even though the student sample in certainty schools is refreshed annually, students in these schools may share characteristics that are not shared with students in non-certainty schools. Although this may not yet lead to measurement concerns, as the level of certainty in the sample increases, the more data may be viewed as similar to census rather than sampled information. As school professionals become familiar with the NAEP assessment, scores

of their students may improve in ways that may not be shared with students in districts for which NAEP is a more novel experience. On the other hand, districts repeatedly selected for NAEP participation may experience some fatigue with and resistance to the assessment, adding another potential threat to the validity of these results.

Key Finding 9: Procedures for scoring constructed-response questions are not fully consistent with best practice.

This finding focuses on procedures employed in scoring constructed-response questions for NAEP assessments and relates to the congressional question about whether NAEP assessments adhere to professional standards. Two issues emerged through our evaluation efforts. The first issue relates to protocols for what happens when a student paper is selected for double scoring to estimate inter-rater agreement reliability. In these instances, the score assigned by the second rater's score is not used, even when it deviates from the score assigned by the first rater. Only the score assigned by the first rater is used in scoring. Given the subjective nature of the scoring guidelines for these item types, we noted two concerns with this practice. First, some raters score at a pace that is more rapid than others when scoring student responses. For these situations, the more rapid raters' scores will be "counted" more often as the operational score. Second, if the scores assigned by the two raters differ, it indicates some potential inaccuracy or at least, uncertainty about our confidence in the resultant score assigned to the performance. Note that if the intended uses of NAEP assessment scores expand in scope beyond the current low-stakes assessment system that does not directly impact individuals, schools, or most districts, these scoring practices would become more critical to our confidence in the resulting scores and decisions.

The second issue within this finding relates to practices for scoring validity papers. Validity papers represent student performances with "known" scores that are included in the scoring process to monitor the consistency and accuracy of raters' performance throughout the scoring process as a quality control strategy. Previously, validity papers were scored as an "event," so that raters knew when a paper would be used as a validity check. This strategy has the potential to influence raters' performance if they know which student performances are being used to monitor the quality of their scoring.

Research and Policy Recommendations

The NAEP assessment lifecycle audit was intended as a broad look at a multifaceted testing program to evaluate important steps in the development, maintenance, and improvement of NAEP processes. Although some select topic areas were evaluated in-depth through special studies within the overall evaluation (See Appendix C), there are aspects of NAEP that could not be investigated in this evaluation because of limited resources but that would benefit from additional study. Some of these (e.g., unclear definition of intended uses of NAEP, limited availability of NAEP technical documentation) have been highlighted in the findings noted above. We have included specific recommendations for the NAEP program that flow from the findings described above and then briefly noted some areas for additional research that were beyond the scope of this evaluation, yet important to the NAEP program.

Recommendation 1: We recommend that the NAEP program develop a transparent, organized validity framework beginning with a clear definition of the intended and unintended uses of NAEP assessment scores (*Standard 1.2*). The specification of intended uses and the development of an organized validity framework should be a joint responsibility of NAGB, NCES, and additional stakeholders (e.g., educators, policymakers). As indicated by *Standard 1.1*, a rationale and supporting research and documentation should be provided to justify the intended use(s). Review of previous or

ongoing NAEP research as described in the body of the report, will likely provide support for the intended uses; however, it is expected that reviewing this body of work will reveal some overlap as well as areas in which sufficient work has yet to be conducted. The validity framework can build on existing research and be organized in a way that supports validity issues in development, program maintenance, and future directions of the program.

Given the importance of a highly visible national assessment program, it is essential that a validity framework be created to coordinate a program of validity research on NAEP, aimed at informing the validity of score interpretation and use. This should be a highlighted component of NAEP; particularly as its perceived role has evolved in the wake of *NCLB*.

Recommendation 2: We recommend that NAGB continue to explore achievement level methodologies as applied to NAEP and consider employing multiple methods with future studies to better inform the policy decision and communicate the policy nature of the decision. The interpretability of NAEP scale scores through the use of achievement levels was an initiative identified by NAGB to aid the public and policymakers. As setting achievement levels is ultimately a policy decision, it is within NAGB's scope to define, establish, and interpret these scores. It is generally accepted among measurement professionals that different methods for setting achievement levels typically produce different results (Jaeger, 1989). Thus, the selection of any one methodology to gather judgments, whether on test characteristics (e.g., Angoff, Bookmark, Mapmark) or examinee characteristics (e.g., borderline group, contrasting groups), only provides one source of evidence for the resultant policy decision. Thus, we further recommend that NAGB consider additional sources of external validity evidence that would be informative to the final policy decision. Some of these sources at the high school level may include results from additional methods, ACT or SAT scores, state university entrance levels, and transcript studies that evaluate course performance. By triangulating these sources of evidence, the cut scores and the resultant impact would strengthen the validity argument.

Recommendation 3: We recommend that NAGB's and NCES's current review and release processes for technical manuals and reports be revised to streamline these efforts while still ensuring high quality and accuracy of NAEP reports. For example, technical information for the aspects of NAEP that have not changed (e.g., test development, scaling procedures) should be publicly available, and information for the most recent tests should be released simultaneously with the test results. This approach would not require reproduction of voluminous technical manuals that repeat much of what is contained in earlier reports, but would rather reference the existing reports and present only information related to the most recent assessments. Although some efforts in this direction have been made as the NAEP technical manuals are transitioning to a Web-based medium, this transition was incomplete during the course of this evaluation.

Recommendation 4: We recommend that the current staffing capacity for NCES's role in NAEP be increased to respond to the increased magnitude of the program. Current NCES staffing levels are inadequate to respond to the operational demands placed on NAEP. To respond to operational needs, some of the activities that may otherwise be conducted within NCES are outsourced to contractors to sustain the program.

Recommendation 5: We identified three areas where additional inquiry is needed in response to the changing policy context of NAEP assessments that have implications for changes in the methods used for sampling.

First, we recommend further study that addresses the impact of differential exclusion and accommodation of special needs students (SWD and ELL) across states. Strategies for estimating the impact of exclusion—including full population estimation (a statistical method for predicting scores in

the full population of students) work done at AIR-CA—appear promising as ways to improve the comparability of State NAEP scores. These and additional strategies should be further explored as well.

Second, we recommend exploration of several questions regarding nonresponse and refusal to participate in NAEP in the current context. Some of these research questions may include: a) What is the impact of nonresponse on NAEP estimates? b) How do the current methods of replacement affect the results? and c) How do these participation rates impact the 12th-grade assessments?

Third, we recommend further exploration of whether NAEP samples as defined by the intended uses are sufficient to support robust estimation of subgroup performance within states or other intended populations because some of these inferences were not necessarily intended at the time these sample sizes were determined. The ability of state samples to provide accurate, valid estimates of subgroup performance in the face of challenges and demographic changes in states and nationally needs to be examined. Related to this recommendation is the need for additional analyses to estimate the impact of repeated administration in units often (or always) selected for NAEP.

Recommendation 6: We recommend that policies and practices related to scoring constructed-response questions, particularly as they relate to the use of the scores assigned by second or subsequent rater, be studied. We also recommend that the NAEP program develop strategies that improve the current practices related to embedded validity papers to monitor the accuracy of raters' performance during the operational scoring procedures. These improvements will help ensure that the validity data derived from these papers more accurately represent the validity of the rating process.

Recommendation 7: We recommend that future contracts for NAEP that involve multiple contractors build on the positive experiences learned in the use of the Alliance, Alliance Coordination and Quality Assurance contracts. The continuation of incentives for cooperative, positive outcomes in an Alliance-like contract is also recommended because it appears to be effective in facilitating collaboration among the members by helping to distribute responsibilities for the success of the program to all contractors within the Alliance.

Additional Research: One additional area of research that has the potential to greatly influence policy considerations is what could be characterized as “alignment”. As used here, alignment refers to the overlap among the NAEP assessment content frameworks and state academic content standards for elementary and secondary education; state assessments and NAEP assessments; and state assessments and NAEP assessment frameworks. Because NAEP is often used by the public as a basis for comparing results from state assessments, whether defined as an intended use or not, further exploration of this area is necessary to ensure valid score interpretations.

This page intentionally left blank

References

- ACT. (1995). *Research studies on the achievement levels set for the 1994 NAEP in geography and U.S. history*. Iowa City, Iowa: Author.
- ACT. (May, 2005b). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Special studies report*. Iowa City, Iowa.
- ACT. (May, 2005c). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Technical report*. Iowa City, Iowa.
- ACT. (April, 2005a). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Process report*. Iowa City, Iowa.
- Allen, N.L., McClellan, C.A., and Stoeckel, J.J. (2005). *NAEP 1999 Long-term trend technical analysis report: Three decades of student performance* (NCES 2005-484). U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. Washington, D.C.
- American Institutes for Research (McLaughlin, D., Scarloss, B., Stancavage, F., and Blankenship, C.). (May, 2005). Using state assessments to impute achievement of students absent from NAEP: An empirical study in four states. NAEP Validity Study.
- American Institutes for Research. (October, 2002). *An Agenda for NAEP Validity Research*. Palo Alto, Calif. By Stancavage, F.B., Beaton, A.E., Behuniak, P., Bock, R.D., Bohrnstedt, G.W., Champagne, A. et al.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Bush, J., and Bloomberg, M. (Aug. 13, 2006). How to help our students: Building on the “No Child” law. *The Washington Post*, p. B07.
- Chaplin, D. (Feb., 2003). Estimating relationships in NAEP: A comparison of IV and traditional methods. Unpublished proposal to U.S. Department of Education (ED 424).
- Chromy, J. (April, 2003). *NAEP Validity Studies: The effects of finite sampling on state assessment sample requirements*. U.S. Department of Education report, NCES 2003-17.
- ETS (Jenkins, F., Qian, J., Braun, H., Davis, S., Kaplan, B., and Pitoniak, M). (April, 2004). The impact of changes implemented in 2003 NAEP—Study 2. Task Order Component Task 2.4.1.1.
- ETS (Pitoniak, M., Bridgeman, B., Braun, H., Donoghue, J., and Kaplan, B.) (April, 2003). Considerations in the use of constructed response items in NAEP. Task Order Component Task 2.4.1.2.
- Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., and Zwick, R. (2000). A response to ‘Setting

reasonable and useful performance standards' in the National Academy of Sciences "Grading the Nation's Report Card." *Educational Measurement: Issues and Practice*, 19(2), 5–14.

- Jacob, B. (Feb., 2003). Test-based accountability and student achievement: An investigation of differential performance trends on NAEP and state assessments. NAEP Secondary Analysis Grant Application.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (ed.) *Educational Measurement 3rd ed.* (pp. 485–514). New York, NY: American Council on Education/Macmillan.
- Jaeger, R.M. (2003). Reporting *the results of the National Assessment of Educational Progress*, NCES 2003-11, NAEP Validity Study, NCES.
- KPMG Peat Marwick LLP. (June, 1996). *Management and Technical Review of the National Assessment of Educational Progress (NAEP)* (study). Washington, D.C.: Huff, L.D.
- Kamata, A. (February, 2003). Differential item functioning analyses for students with test accommodations on NAEP test items. Unpublished proposal R902B030025 submitted to the U.S. Department of Education.
- Linn, R. L. (2004). The influence of external evaluations. In L. Jones and I. Olkin (eds.) *The Nation's Report Card: Evolution and Perspectives*. Bloomington, Ind.: Phi Delta Kappa Educational Foundation.
- Loomis, S. C., and Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, N.J.: Erlbaum.
- National Assessment Governing Board. (2004a). *Policy statement on reporting, release, and dissemination of national assessment results*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2004b). *Reporting Schedule for 2005 NAEP assessments*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2004c). *Specifications for NAEP 2005 reports*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2004d) *Mathematics Framework for the 2005 National Assessment of Educational Progress*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002a). *Collection and reporting of background data by the national assessment of educational progress*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002b). *Framework development*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.

- National Assessment Governing Board. (2002c). *NAEP item development and review*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002d). *Policies and procedures for complaints related to the national assessment of educational progress*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Commission on NAEP 12th Grade Assessment and Reporting. (2004). *12th grade student achievement in America: A new vision for NAEP*. A report to the national assessment governing board. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- Education Sciences Reform Act of 2002, Title III*, National Assessment of Educational Progress, Pub. L. No. 107-279, 20 U.S.C. § 9622.
- No Child Left Behind Act of 2001*, Pub. L. No. 107-110, 115 U.S.C. § 1425 (2002).
- Pellegrino, J. W., Jones, L. R., and Mitchell, K. J. (1999). *Grading the nation's report card*. Washington, D.C.: National Academy Press.
- Schulz, E. M., and Mitzel, H. (April, 2005). The Mapmark standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec.
- Shakrani, S. (June, 2005). In English language learners in state NAEP and state assessments—Shall the twain ever meet? Session presented at the 35th Annual National Conference on Large Scale Assessment, Council of Chief State School Officers, San Antonio.
- Shepard, L., Glaser, R., Linn, R., and Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, Calif.: National Academy of Education
- Spurlock, H. (August, 2006). Field operations report on students with disabilities and English language learner issues. Presented to the NAGB Reporting and Dissemination Committee.
- U.S. Department of Education, Institute of Education Sciences. (2005). *National Assessment of Educational Progress Secondary Analysis Grants: Request for Applications*. Retrieved on June 29, 2005, from www.ed.gov/about/offices/list/ies/programs.html.
- U. S. Department of Education, National Center for Education Statistics. (2003a). A content comparison of the NAEP and PIRLS fourth-grade reading assessments. Working paper series. (NCES-WP-2003-10). Washington, D.C.: Binkley, M., and Kelly, D.L.
- U. S. Department of Education, National Center for Education Statistics. (2003b). NAEP quality assurance checks of the 2002 reading assessment results for Delaware. (WP-2003-10). Washington, D.C.
- U. S. Department of Education, National Center for Education Statistics. (2002). *NCES Statistical Standards*. Downloaded from: <http://nces.ed.gov/statprog/2002/stdtoc.asp>.

U.S. General Accounting Office, (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* (GAO/PEMD Publication No. 93-12). Washington, D.C.: Author.

Warren, J. (August, 2004). High school exit examinations and NAEP long-term trends in reading, mathematics, and science, 1970–2004. Unpublished proposal to the U.S. Department of Education, CFDA # R902.B04.

Appendixes

This page intentionally left blank

Appendix A: Glossary of abbreviations and technical terms used in report

AA—see *Assessment Administrator*

Achievement Level—category used in reporting assessment results of student performance based on scale scores. In NAEP, three achievement levels are used in reporting: Basic, Proficient, and Advanced. A fourth level, Below Basic, is sometimes used but is undefined.

Achievement Level Description/Descriptor (ALD)—the expected knowledge and skills of students categorized within each achievement level.

Achievement Level Standards—test performance expectations for specific achievement levels. The NAEP achievement level standards are typically set by NAGB based on recommendations derived from a *standard setting* process that involves the judgment of expert panelists familiar with the content and target population of students being tested.

ADC—Assessment Design Committee of NAGB

Administration Accommodation—alterations to the administration procedures for students with disabilities or other limitations when such disabilities or limitations unfairly influence test performance. An example of an administration accommodation would be providing large print test materials for visually impaired test-takers.

AERA—American Educational Research Association

AIR—American Institutes for Research

AIR-DC—American Institutes for Research, Washington D.C., office

AIR-CA—American Institutes for Research, Palo Alto, Calif., office

ALD—see *Achievement Level Descriptor*

Alignment—degree of overlap between (a) the knowledge, skills, and expertise measured by a test (as indicated by the test items), and (b) the knowledge and skills included within the test *content specifications*. Alignment can also refer to the degree of consistency between more than one set of content specifications or more than one assessment.

APA—American Psychological Association

Assessment Administrator (AA)—individual who assists with the administration of NAEP in the schools.

Assessment Coordinator (AC)—individual responsible for coordinating the administration of NAEP including preparation of sites and materials for administration sites.

Assessment Framework—see *Content Specifications*

Assessment Mode—the format used to administer an assessment. Assessment modes include, but are not limited to, paper and pencil, computer-based (linear and adaptive), and performance assessments.

Background Variables—information about an examinee’s demographic and educational background. In NAEP, this information is used to estimate an examinee’s scores on the assessment.

Backreading—a quality control procedure in scoring question responses whereby an experienced scorer supervisor checks the accuracy of assigned scores. In NAEP, scoring supervisors backread a small percentage of student responses to monitor scorer accuracy.

Backscoring—see *Backreading*

BIACO—Buros Institute for Assessment Consultation and Outreach (University of Nebraska)

Bias—see *Item Bias*

CCD—see *Common Core of Data*

CCSSO—Council of Chief State School Officers

CEA—Center for Educational Assessment (University of Massachusetts–Amherst)

Common Core of Data (CCD)—This program, that is part of NCES, collects annual data about all public schools (e.g., students and staff demographic data) and state education agencies across the United States.

Conditioning—a process used to incorporate information (see *Background Variables*) into the estimation of an examinee’s score on an assessment in addition to their responses to the test questions. In NAEP, background information provided by examinees is incorporated in the score estimation process.

Constructed Response Item—a test question which requires students to create (write) a response, versus selecting a response from among multiple alternatives.

Content Specifications—an outline or framework of the specific knowledge or ability domains which will be assessed by the test and the number and types of items that will represent each test domain

Contextual Variable Inference Map (C-VIM)—In NAEP, this is a system used by AIR-DC to understand the influence of background characteristics in test performance.

Contracting Officer’s Representative (COR)—these individuals represent the federal contracting officer and advise on technical contract matters as well as serve as liaisons between the contractors and various stakeholders (e.g., NAGB, external evaluators).

COR—see *Contracting Officer’s Representative*

COSDAM—Committee on Study Design and Methodology of NAGB

C-VIM—see *Contextual Variable Interference Map*

DAC—NAGB Design and Analysis Committee

DIF—See *Differential Item Functioning*

Differential Item Functioning (DIF)—a difference in estimated difficulty of an item between two groups after controlling for any differences between the groups in subject-matter knowledge.

ED—U.S. Department of Education

ELL—English Language Learner (see *Limited English Proficiency*)

Equating—the practice of relating test scores from two or more test forms that are built to the same content to make the test scores comparable. A popular equating design utilizes information gathered from a set of common items (also referred to as anchor items or anchor tests) that are administered to all students in order to establish linkage between test scores.

ESSI—see *NESSI*

ETS—Educational Testing Service

Field testing—See *Pilot Testing*

Framework—see *Content Specifications*

GMRI—Government Micro Resources Inc.

HumRRO—Human Resources Research Association

IDEA—*Individuals with Disabilities Education Act*

IEP—Individualized Education Program. These programs are created for students with disabilities and in NAEP, these are reviewed to determine if a student qualifies for an *accommodation*.

IES—Institute of Education Sciences, U.S. Department of Education

IMS—Integrated Management System. This system was created by GMRI as a way for the NAEP Alliance contractors to communicate with one another.

Inter-rater Agreement Reliability—the consistency (agreement) of scores or ratings given by two or more raters for the same set of responses.

IRT—see *Item Response Theory*

Item—a question included on the assessment which may be designed to collect demographic information (see *Background Variables*) or assess the knowledge, skills, or abilities of examinees.

Item Bias—item or test bias occurs when one group is unfairly disadvantaged based on a background or environmental characteristic that is unique to their group.

Item Pool—the group of test questions created for a testing program from which a test publisher or administrator will create a test form.

Item Response Theory (IRT)—a measurement model that mathematically defines the relationships between observed item responses (that examinees provide when taking a test) and one or multiple latent (i.e., not directly observable) traits (e.g., mathematics ability, U.S. history knowledge).

ITS—Item Tracking System

LEP—Limited English Proficiency (students classified as LEP are also known as English Language Learners [ELL]).

Linking—the practice of relating scores from two different tests. *Equating* is a special (stringent) type of Linking.

Mapmark—a standard setting methodology used to set cut scores for the 12th-grade NAEP assessment for mathematics.

Matrix sampling—a process used to select a sample of items to be administered to examinees from an item pool that adequately covers the construct of interest. In a NAEP administration, examinees are only administered a portion of a full exam (e.g., fourth-grade mathematics exams). Examinees' performance on the full exam is estimated based on *background variables* (e.g., math classes taken) and other NAEP data (e.g., how other students did on the other parts of the NAEP mathematics test).

NAEP—National Assessment of Educational Progress

NAEP Alliance—The group of contractors selected by NCES to carry out the development, administration, and scoring of NAEP under the coordination of the Educational Testing Service (ETS).

NAEP Consortium—Agencies, contractors, and organizations involved in the NAEP process that were of consideration for this evaluation.

NAGB—National Assessment Governing Board

NCES—National Center for Education Statistics

NCLB—*No Child Left Behind Act of 2001*

NCME—National Council on Measurement in Education

NESSI—NAEP–Education Statistics Services Institute (formerly ESSI)

NRC—National Research Council

NVS—NAEP Validity Studies Panel.

OMB—Office of Management and Budget of the U.S. government

Open-ended Item—see *Constructed Response Item*

Operational Scoring—scoring of actual examinee item responses using scoring procedures determined during the test development process.

Oversampling—a sampling procedure that disproportionately selects a higher percentage of members from a subgroup than from other groups to be included in a sample. In NAEP, this procedure is used to achieve better precision in the ability estimates for small subgroups.

Parameter Estimate—a statistical quantity which is derived from a sample and is used to make an inference about a population. In NAEP this may refer to an estimate of ability for a particular group or performance on an item.

PEM—Pearson Educational Measurement

Performance Assessment—the measurement of intended knowledge and skills of students, which require students to engage in some type of activity. Performance assessments may include such tasks as writing, conducting a science experiment, or analysis of a portfolio of work.

Performance Standards—also referred to as cut scores, these represent the expected performance (score) of examinees on a measure to be classified within specific achievement levels. In NAEP, performance standards are set for classifying examinees into the Basic, Proficient, and Advanced achievement levels on each assessment.

PIL—Process Improvement Log—This log is maintained by HumRRO and includes the minutes from any meetings of the QCT and QAC to discuss specific issues.

Pilot Testing—part of the test construction process whereby the assessment is administered to a sample of examinees, prior to the operational administration, to assess the psychometric quality of test items. The results of pilot tests are used to develop the final test form.

PIRLS- Progress in International Reading Literacy Study

PISA—Programme for International Student Assessment

Principal Components Analysis—a statistical method that detects relationships within a group of variables in order to reduce a data set to a minimal number of variables. In NAEP, the background information gathered about examinees is reduced to a smaller number of variables using this process.

Psychometrics—the theory and techniques of educational and psychological testing. Psychometrics involves construction of appropriate assessments with the goal of providing valid and fair test score interpretations.

QAC—Quality Assurance Council—The QAC consists of representatives from NCES, the NAEP Alliance, and HumRRO. The purpose of QAC is to facilitate the discussion of quality matters, develop broad quality control policies and standards, and promote a highly functional cross-organizational atmosphere.

QAP—Quality Assurance Panel—This is an external panel whose members serve in an advisory role to HumRRO in their NAEP quality assurance responsibilities.

QC—Quality Control

QCT—see *Quality Control Team*

Quality Control Team (QCT)—The QCT consists of representatives from each Alliance member and HumRRO, who implement standards and policies articulated by the QAC, coordinate quality control activities across the Alliance, develop tools and methods to address quality control issues and inform the QAC of critical quality control issues.

Reliability—the consistency of measurement. In educational assessment, reliability typically refers to internal consistency (consistency of items within an assessment) or test-retest reliability (consistency of test scores across repeated measurements). See also *Inter-Rater Agreement Reliability*.

Response Format—the mode in which examinees respond to an item. Common response formats include (i) selection of the correct response among options, and (ii) constructed response.

RFP—Request for Proposals

SAG—see *Secondary Analysis Grants*

Sample/Sampling—A sample is a subset of the target population (e.g., schools, students or items). Sampling is the process of selecting members of the population to be included in a sample. The NAEP assessment is administered to a sample of students from across the country.

Scale Score—A value representing an estimate of an examinee’s ability on some type of reporting scale. In NAEP, the score scale ranges from 0 to 500 for the fourth- and eighth-grade mathematics, for example. Scores on this scale are estimated based on how examinees respond to questions and NAEP *Background Variables*.

Scale stability—the degree to which values on a score scale possess the same meaning over time or across groups.

Scaling—the process of converting raw scores into equivalent values on an established reporting scale.

Score Equity—the consistency in score meaning across various contexts. In this evaluation, a special study was conducted to evaluate the score equity of NAEP scores across several states.

Scorer Calibration—the process by which human scorers are trained to assign scores in accordance with established scoring rubrics and procedures.

Scorer Drift—when a human scorer deviates over time from the scoring procedures established during *Scorer Calibration*.

Scoring Rubrics—guidelines used to evaluate student responses to a constructed-response item by specifying criteria for scoring that distinguish between possible score points (e.g., a 1-point response versus a 2-point response)

Secondary Analysis Grants (SAG)—this research program is run by NCES (priorities set by NAGB) and provides research funds to conduct studies with NAEP data.

SEM—see *Standard Error of Measurement*

SES—Socioeconomic Status—In NAEP, this is part of the information gathered through the *Background Variables*.

SOW—Statement of Work

Standard Deviation—a statistical value that describes the variance or dispersion of data points around a group average. Higher values indicate more variance in a dataset.

Standard Error of Measurement (SEM)—the degree of error associated with observed test scores. SEM is inversely related to test score reliability.

Standard Setting—the process used to establish cut scores for an assessment. A cutscore is chosen to distinguish between adjacent achievement levels (e.g., Basic and Proficient, Proficient and Advanced). Methods of standard setting include, but are not limited to, the Mapmark method, Bookmark method, and Angoff.

Standards—*Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999)

Statistical Power Analysis—a statistical procedure used to estimate the necessary sample size to achieve measurement precision or to enable the detection of a given effect in a research study (e.g., increase in student knowledge).

SWD—Students with disabilities

Test Specifications—see *Content Specifications*

TIMSS—Trends in International Mathematics and Science Study

TOC—Task Order Component

TOS—Table of Specifications

Trend Item—assessment items that appear in sequential NAEP assessments that are maintained for the purposes of tracking any change in performance over time.

Trend Paper—examinee responses to open-ended questions that have appeared on sequential NAEP assessments. To maintain the trend in NAEP, these responses must be scored in same manner as on previous NAEP assessments.

TUDA—Trial Urban District Assessment

TWG—Technical Work Group

Validity—the degree to which a test is measuring what it is intended to measure. Validity evidence can be gathered through appropriate processes or through research studies, and supports the meaningfulness of the test scores for the intended purpose(s) of the test.

Weights/weighting—Sample *weights* are values assigned to the score of an examinee (based on their subgroup membership) in estimation of the overall performance of a larger group. The value is chosen in such a way to reflect the proportion of the number of group members in the overall population.

Appendix B: Legislation authorizing Evaluation of NAEP

B1. Current Legislative Requirements for the Evaluation of NAEP

In Section 303 of the *National Assessment of Educational Progress Authorization Act*, Title 20, U.S.C.9622, Congress required an independent review of NAEP:

“(f) REVIEW OF NATIONAL AND STATE ASSESSMENTS-

(1) REVIEW –

IN GENERAL- The Secretary shall provide for continuing review of any assessment authorized under this section, and student achievement levels, by one or more professional assessment evaluation organizations.

(B) ISSUES ADDRESSED- Such continuing review shall address-

- I. whether any authorized assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);
- II. whether student achievement levels are reasonable, valid, reliable, and informative to the public;
- III. whether any authorized assessment is being administered as a random sample and is reporting trends in academic achievement in a valid and reliable manner in the subject areas being assessed;
- IV. whether any of the test questions are biased, as described in section 412(e)(4); and whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.

"(2) REPORT.-- The Secretary shall report to the Committee on Education and the Workforce of the House of Representatives and the Committee on Health, Education, Labor, and Pensions of the Senate, the President, and the Nation on the findings and recommendations of such reviews.

"(3) USE OF FINDINGS AND RECOMMENDATIONS.-- The Commissioner and the National Assessment Governing Board shall consider the findings and recommendations of such reviews in designing the competition to select the organization, or organizations, through which the Commissioner carries out the National Assessment.”

B.2. Prior Legislative Requirements for the Evaluation of NAEP

The *No Child Left Behind* legislative language expands upon the 1994 legislative language mandating the prior evaluation:

“(f) REVIEW OF NATIONAL AND STATE ASSESSMENTS

(1) IN GENERAL-

(A) The Secretary shall provide for continuing review of the National Assessments, State assessments, and student performance levels, by one or more nationally recognized evaluation organizations, such as the National Academy of Education and the National Academy of Sciences.

(B) Such continuing review shall address-

- I. whether each developmental State assessment is properly administered, produces high quality data that are valid and reliable, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation); and
- II. whether student achievement levels are reasonable, valid, reliable, and informative to the public.

B.3. Legislative Requirements for the Review of Performance Levels

In addition, recent legislation requires the commissioner of education statistics to rely upon the evaluation for his determination of whether or not the achievement levels are “reasonable, valid, and informative to the public.” Until that determination is made, the law requires the commissioner and the Board to state the trial status of the achievement levels in all NAEP reports.

Appendix C: Special Studies in the Evaluation of NAEP

I. Utility of NAEP Reports

Given the increased national visibility of NAEP, these special studies represent a unique emphasis compared to previous evaluations. Specifically, these studies focus on the utility of NAEP reports as interpreted by a range of stakeholders. These studies were included to respond to congressional questions about valid interpretations of NAEP assessment scores (e.g., do stakeholders correctly interpret data) as presented in reports and various data displays. They also address related questions about how achievement levels may be informative or used by the stakeholders. The reports and data presentation evaluated in these studies include paper and electronic (e.g., Web-based) modes. Data collection for these evaluation activities includes interviews, focus groups, and studies of how well consumers of NAEP results correctly interpret the reported results.

II. Validity of NAEP Achievement Levels

Studies in this topic area were selected to respond to the ongoing discussion about appropriate methods and interpretations of achievement levels in the context of NAEP. These studies were designed to directly respond to the Congressional concerns about the validity of NAEP achievement levels. One of the initial activities was to conduct a comprehensive evaluation of the new Mapmark standard setting (Schulz and Mitzel, 2005) that was recently applied to the 2005 Grade 12 NAEP Mathematics assessment. Additional studies include an analysis of the NAEP achievement levels in math and science in relation to external validity evidence.

III. Score Equity Assessment

This study addresses an important issue in fairness by evaluating whether methods to derive NAEP scores in subgroups (e.g., states) are comparable. This question will be examined across five select states on the Grade 8 NAEP Mathematics assessment and the Grade 8 Reading assessment. This study was identified for inclusion in the evaluation because it addresses the stability of the score scale within and across subgroups. Some educational policy decisions may have requirements about performance over time that could be suspect if the underlying scores cannot be interpreted similarly over multiple years. A key element of this study will be replication of the equating processes for select Main NAEP assessments. This study is also unique in that it allows us to use these data to inform the audit study as we evaluate the equating methodologies and the potential impact on subgroups.

IV. Review of Alignment methods

Alignment is a critical policy consideration for interpreting scores. If there were a low level of alignment between curriculum and instruction in the country when compared with the emphasis in the respective NAEP framework (e.g., reading, mathematics, science), there would be less confidence that the observed performance is a good indicator of achievement as defined by NAEP. The higher the level of alignment, the greater our confidence may be in these score interpretations. This study represents a review of alignment methodologies and available studies that describe alignment with NAEP assessments or frameworks.

This page intentionally left blank

Appendix F-2: NAEP audit site visit timeline

Agency/Contractor	Date of Visit	Audit Team Members	Agency/Contractor Representatives
NCES	June 6, 2005	Chad Buckendahl, Susan Davis	Peggy Carr, Andy Kolstad, Drew Malizio, Janis Brown, Arnold Goldstein, Suzanne Triplett
NAGB	June 7, 2005	Chad Buckendahl, Susan Davis	Charles Smith, Sharif Shakrani, Susan Loomis, Mary Crovo
AIR-CA	June 29, 2005	Chad Buckendahl, Ed Wiley	Victor Bandeira de Mello, Don McLaughlin, Fran Stancavage, George Borhnstedt
HumRRO	June 30, 2005	Barbara Plake, Jim Impara	Laress Wise, Sunny Becker, Felicia Butler, Carolyn Harris, Gene Hoffman, Paul Sticha
Westat	July 11–12, 2005	Chad Buckendahl, Ed Wiley	Nancy Caldwell, Debbie Vivari, David Morganstein, Diane Cadell, Keith Rust, Kavamuimurangi., Catrina Williams
AIR-DC	Aug. 15, 2005	Barbara Plake, Jim Impara	Barry Levine, Sigrid Gustafson, George Borhnstedt, Larry Albright, Helene Mullaney, Kristin Leahy
PEM	Sept.12, 2005	Barbara Plake, Jim Impara	Connie Smith, Steve Kromer, Mary Schulte, Carolyn Loew, Bill Buckles, Erick Hlebowitsh, Russ Vogt, Jim Close, Pat Sterns
State Coordinators	Sept.26–28, 2005	Chad Buckendahl, Susan Davis	Marcie Hickman (North Carolina), Robert Hillier (Hawaii), Wendy Geiger (Virginia), John Kennedy (Maine), Kathryn Sprigg (Washington), Barbara Smey-Richman (New Jersey), Dianne Chadwick (Iowa)
Hager-Sharp	Sept. 26–28, 2005	Chad Buckendahl, Susan Davis	Facilitators of the NAEP Pre-Release Workshops, (not a formal meeting or site visit arranged with Hager-Sharp).
GMRI	Oct. 13, 2005	Chad Buckendahl, April Zenisky Laguilles	Paul Harder, Lori Rokus, Keith Lamond
ETS	Oct. 27–28, 2005	Barbara Plake, Ron Hambleton, Steve Sireci	Jay Campbell, Gloria Dion, Amy Drescher, Robert Finnegan, David Freund, Lydia Gladkova, Eugene Gonzalez, Jeff Haberstroh, Edward Kulick, Michael Lapp, Steven Lazar, John Mazzeo, Nancy Mead, Hilary Persky, Mary Pitoniak

This page intentionally left blank

Appendix G: Site visit reports

Site Visit Reports

- G1: National Assessment Governing Board
- G2: National Center for Education Statistics
- G3: Marilyn Seastrom (Chief Statistician, NCES)
- G4: Educational Testing Service
- G5: American Institutes for Research – D.C.
- G6: American Institutes for Research – Calif.
- G7: Government Micro Resources Inc.
- G8: Human Research Resources Organization
- G9: Pearson Educational Measurement
- G10: Westat
- G11: NAEP State Coordinators

Document Reviews

- G12: Hager Sharp

This page intentionally left blank

Appendix G-1: National Assessment Governing Board (NAGB)

Site visit team: Chad Buckendahl and Susan Davis, Buros Center for Testing

Date of visit: June 7, 2005

Audit Summary

Staff

Charles Smith – Executive Director (works with all Committees)
 Sharif Shakrani – Deputy Associate Director (Nominations Committee) [Now affiliated with Michigan State University]
 Susan Loomis – Assistant Director of Psychometrics (Committee on Standards, Design, and Methodology)
 Mary Crovo – Assistant Director of Test Development (Assessment Development Committee)
 Ray Fields – Assistant Director for Policy and Research (Executive Committee) [Not in attendance on June 7, 2005]
 Lawrence Feinberg - Assistant Director for Reporting and Dissemination (Reporting and Dissemination Committee) [Not in attendance on June 7, 2005]

Prior to the site visit, the audit team contacted NAGB and requested documentation of the processes and procedures used by the organization; however, no documents were provided. Several documents were provided during the site visit and following the site visit, numerous policy documents were accessed from the NAGB Web site.

Organizational characteristics

Brief descriptions of staff member qualifications of each of the above staff members are noted on the NAGB Web site (<http://www.nagb.org/>).

The staffing at NAGB is experienced which aids in responding to new issues. In the past several years, Congress has increased NAGB's responsibilities by putting the Board in charge of the major reporting of NAEP results. In response, NAGB increased its staff to include personnel with relevant skills. In the future, the addition of new responsibilities will necessitate adding new staff; however, current staffing appears to be able to respond to demands. Currently, each staff member is responsible for one NAGB committee (see above). In addition to the core staff listed above, NAGB has support staff to help with day-to-day operations and preparation for Board meetings.

Problem identification tends to occur when there is a conflict with NAGB policy. As the decision-makers regarding NAEP, NAGB's policies take precedence in any conflict resolution strategy that is employed.

The Board represents a range of backgrounds and expertise. Specific issues that come before the Board may be outside the expertise of many Board members. In such cases, NAGB staff arrange for experts to attend Board meetings that can explain relevant issues and educate the committees. For example, the advice of experts has been sought while the Board was exploring the issue of 'preparedness' with 12th grade NAEP

(National Commission, 2004). In addition, contractors often provide expertise to the Board on particular issues (e.g., Achieve, ACT, WestEd, CCSSO).

According to the NAEP legislation (P.L. 107-279) NAGB Board must be composed of the following:

- Two Governors, or former Governors, who shall not be members of the same political party.
- Two state legislators, who shall not be members of the same political party.
- Two chief State school officers.
- One superintendent of a local educational agency.
- One member of a State board of education.
- One member of a local board of education.
- Three classroom teachers representing the grade levels at which the National Assessment is conducted.
- One representative of business or industry.
- Two curriculum specialists
- Three testing and measurement experts, who shall have training and experience in the field of testing and measurement.
- One nonpublic school administrator or policymaker.
- Two school principals, of who one shall be an elementary school principal and one shall be a secondary school principal.
- Two parents who are not employed by a local, State or Federal educational agency.
- Two additional members who are representatives of the general public, and who may be parents, but who are not employed by a local, State, or Federal educational agency.

According to P.L. 107-279 the secretary of education and the Assessment Board are responsible for ensuring that the Board membership represents “regional, racial, gender, and cultural balance and diversity and that the Board exercises its independent judgment, free from inappropriate influences and special interests.” (section 302, 2(3)). The secretary is responsible for appointing new Board members. These appointees are chosen from nominations received from organizations represented above (e.g., Chief State school officers, Governors). Each organization is asked to nominate six persons who have the desired qualifications. A term as a Board member cannot exceed four years and members may not serve more than two terms.

The selection criteria and process for nomination to the Board is not included on NAGB’s Web site. This information is requested from NAGB to add transparency to the nomination process.

The flow of the decision-making process appears to begin with the respective NAGB staff member bringing an issue to a NAGB subcommittee that is then discussed among the members. The NAGB subcommittees are: Committee on Standards, Design and Methodology Reporting and Dissemination Committee, Assessment Development Committee, Nominations Committee, and the Executive Committee. Once the subcommittee has responded to the issue, the topic may be brought to the full Board. Generally, there appears to be a timeline that brings particular topics to the full Board first as information and then as an action item at a subsequent meeting. There did not appear to be instances where topics were introduced and then acted upon before being presented to Board members at two or more meetings.

Communications

Among contractors:

Timelines for NAGB contractors are generally built into the contracts. Most timelines are dictated either by the assessment schedule or by Board meetings (to ensure that the Board has needed information to make decisions).

NAGB does not use incentive-based contracts. However, NAGB can withhold payment if contracts are not fulfilled on time. This is expected because the contracts that NAGB oversees are generally much smaller than the ones for which NCES is responsible.

NAGB contracts are run by NAGB staff. Each NAGB staff member takes responsibility for contracts that are within the purview of his or her respective subcommittee. For example, Crovo is responsible for monitoring the contract by CCSSO and WestEd with respect to the development of science frameworks.

Clarity of roles

NAGB serves as the visible face of NAEP. The staff members are responsible for preparing for and facilitating Board meetings, starting the nomination process for new Board members, monitoring various NAEP-related meetings (e.g., DAC, NVS), and providing updates at these meetings on NAGB activities. The NAGB responsibilities listed on its Web site (<http://www.nagb.org>) include:

- Selecting subject areas to be assessed
- Developing appropriate student achievement levels
- Developing assessment objectives and test specifications that produce an assessment that is valid and reliable, and are based on relevant widely accepted professional standards
- Designing the methodology of the assessment
- Developing guidelines for reporting and disseminating results
- Developing standards and procedures for regional and national comparisons
- Approving all cognitive and noncognitive NAEP items
- Taking appropriate actions needed to improve the form, content, use and reporting of results.

From the “Duties” section (section 302, 5) of P.L. 107-279, NAGB’s responsibilities include the following six components:

(1) IN GENERAL—In carrying out its functions under this section the Assessment Board shall—

1. Select the subject areas to be assessed (consistent with section 303(b));
2. Develop appropriate student achievement levels as provided in section 303(e);
3. Develop assessment objectives consistent with the requirements of this section and test specifications that produce an assessment that is valid and reliable, and are based on relevant widely accepted professional standards;
4. Develop a process for review of the assessment which includes the active participation of teachers, curriculum specialists, local school administrators, parents, and concerned members of the public;
5. Design the methodology of the assessment to ensure that assessment items are valid and reliable, in consultation with appropriate technical experts in

- measurement and assessment, content and subject matter, sampling, and other technical experts who engage in large scale surveys;
6. Be Consistent with section 303, measure student academic achievement in grades 4, 8, and 12 in the authorized academic subjects;
 7. Develop guidelines for reporting and disseminating results;
 8. Develop standards and procedures for regional and national comparisons;
 9. Take appropriate actions needed to improve the form, content, use, and reporting of results of any assessment authorized by section 303 consistent with the provisions of this section and section 303; and
 10. Plan and execute the initial public release of National Assessment of Educational Progress reports.

The National Assessment of Educational Progress data shall not be released prior to the release of the reports described in subparagraph (J).

(2) DELEGATION—The Assessment Board⁷ may delegate any of the Assessment Board's procedural and administrative functions to its staff.

(3) ALL COGNITIVE AND NONCOGNITIVE ASSESSMENT ITEMS—The Assessment Board shall have final authority on the appropriateness of all assessment items.

(4) PROHIBITION AGAINST BIAS—The Assessment Board shall take steps to ensure that all items selected for use in the National Assessment are free from racial, cultural, gender, or regional bias and are secular, neutral, and non-ideological.

(5) TECHNICAL—In carrying out the duties required by paragraph (1), the Assessment Board may seek technical advice, as appropriate, from the Commissioner for Education Statistics and other experts.

(6) REPORT—Not later than 90 days after an evaluation of the student achievement levels under section 303(e), the Assessment Board shall make a report to the Secretary, the Committee on Education and the Workforce of the House of Representatives, and the Committee on Health, Education, Labor, and Pensions of the Senate describing the steps the Assessment Board is taking to respond to each of the recommendations contained in such evaluation. (section 302, 5)

During the site visit, NAGB also described its responsibilities as including approval of the test administration process and special studies (e.g., to ensure that schools are not overburdened). Although broadly stated as having responsibility for the operations of NAEP, NAGB did not directly address the question of NCES's specific role. However, other sections in this report describe the responsibilities of NAGB in relation to the NCES operations. The primary distinction between the organizations is that NAGB's role is more policy-oriented whereas the role of NCES is more on the operations level. The example NAGB used was the recent decision by the Board to include a vocabulary section on the reading assessment. Although the Board decided to

⁷ The National Assessment Governing Board is also referred to as the Assessment Board.

include this section, NCES and ETS were responsible for determining a method for implementing this request.

Given the increased importance of NAEP and the additional responsibilities of each organization, overlap and differences of opinion in interpreting NAGB's and NCES's responsibilities are inevitable. The most recent instance of overlap is the transition of the reporting responsibilities from NCES to NAGB. This change in responsibility was made because Congress wanted the reporting to come from NAGB because of its position as an independent entity. The handoff has taken some time and NAGB has used NCES's resources (e.g., Hager Sharp) as it formulates their own contracts with these organizations. NAGB mentioned that NCES has been very cooperative during this transition process that has taken place over the past two years. The question of where NAGB believes that NCES oversteps its authority was also not directly answered. These disagreements were characterized as differences in interpretation of the responsibilities. It appears, though, that when there are differences in interpretation, NAGB has the ultimate responsibility and therefore decision authority. NAGB views the tension between NAGB and NCES as healthy. NCES and NAGB have joint staff meetings at which they can discuss and resolve issues. The initial release of reports was mentioned as an example of a topic that has been discussed at these meetings.

Intended scope and uses of NAEP assessments

The intended scope of NAEP assessments is set by law, however, there appears to be some ambiguity in how it has been interpreted (e.g., NVS studies comparing state assessment and NAEP performance by school). NAGB's job is to inform policymakers not dictate public policy. For example, NAGB was recently given the responsibility of releasing NAEP results. From P.L. 107-279, NAGB's duties include:

“7. Develop guidelines for reporting and disseminating results; ... 9. Take appropriate actions needed to improve the form, content, use, and reporting of results of any assessment authorized by section 303 consistent with the provisions of this section and section 303; and 10. Plan and execute the initial public release of National Assessment of Educational Progress reports.” (section 302, 5)

NAGB's responsibility in this situation is to articulate how NAEP data should be reported and how it should not be reported. NAGB avoids telling states directly how to interpret NAEP results in relation to state test data; however, states are free to make their own comparisons. NAGB's responsibility is to ensure that NAEP reports include caveats that such comparisons are difficult as NAEP is a survey (not a census) testing program and the NAEP frameworks are built differently than the state frameworks.

With *NCLB* legislation, there has been increased interest in NAEP frameworks across the country. The Board cannot advocate the NAEP frameworks for states (NAGB, 2002e); however, it makes the frameworks available to any states that request them. In the introduction to the current NAEP Mathematics Framework (NAGB, 2004e), it states:

Of critical importance is the fact that this document does not attempt to answer the question: What mathematics should be taught (or how)? This is an assessment framework, not a curriculum framework. It was developed with the understanding that some concepts, skills, and activities in school

mathematics are not suitable to be assessed on NAEP, even though they may be important components of a school curriculum. (¶ 3)

In this sense, NAGB has to be more reactive rather than proactive with states' use of the frameworks to comply with its mission and scope of work.

The validity of inferences on NAEP scores is built on the NAEP frameworks. Although they serve as the foundation for NAEP development and reporting, the creation of these frameworks alone does not ensure appropriate interpretations of NAEP results. During the development process, the frameworks are reviewed by a panel of experts that look at the frameworks in late draft form. There are also additional formal and informal reviews during the framework development process. After the frameworks are developed, items are created to match the frameworks; however there do not appear to be any alignment studies conducted independent of the item development contractor. NAGB board members review the match between the framework and item pools as reported by the contractors and send an observer to item development meetings. Overall, NAGB stated that it is not involved with any work that considers the validity of the use of NAEP reports or results; however, it does attend to the needs of the public through one of its subcommittees (i.e. dissemination and reporting). Its recent work on the 12th-grade NAEP assessments is an example of perceiving the public's need for specific information from NAEP. NAGB's projects of this nature are generally shorter and may not collect data as would be done in research studies. The Hager Sharp study that was reported during the November, 2004 Board meeting did describe information gathered from NAEP coordinator focus groups. Many of these reports are prepared to inform the Board before it makes important policy decisions.

NAGB staff noted that the validity work conducted by some of the groups involved with NAEP (e.g., NESSI) is very relevant to the work that NAGB is responsible for; however, this work or the validity framework is not necessarily shared with NAGB. This has led to some duplication in efforts. For example, both NAGB and NCES have contractors working on motivation-related research simultaneously. Once NAGB found out about the other work it requested information; however, this information was not shared. This communication about work done related to the broader validity framework appears to be a source of some frustration for NAGB.

NAGB is responsible for setting the priorities of the secondary analysis grant program. However, reviews of the proposals are conducted by an external peer review panel organized by the Institute of Education Sciences (2005). NAGB is not responsible for ensuring a match between the noted priorities and completed work of the secondary analysis grants.

Evaluating consequential validity does not appear to be within the scope of NAGB's responsibilities. However, at the NAGB board meeting in May 2005 (and previous meetings) there were some discussions within the reporting and dissemination subcommittees regarding information that board members could have to respond to media requests after the initial release of data. It would appear that these materials serve as an opportunity to encourage appropriate interpretation of NAEP data in addition to discouraging inappropriate uses.

Roy Truby, in a report to NAGB, noted problems with 12th-grade NAEP. A commission was charged to examine these problems. Several meetings and papers resulted from this issue and the commission prepared a report that included five recommendations (National Commission, 2004). The issue of college preparedness is

what is being discussed first by the Board. The prioritization of this issue is tied with concerns across the country that the nation should be producing qualified students. Measuring preparedness will mean changes for NAEP frameworks. In the past, 12th-grade NAEP has assessed the mathematics skills needed by all students. The skills needed to be prepared for college may not be the same as those needed for the workforce or military. A second important issue presented in this report is that of motivation and effort on the 12th-grade assessment.

NAGB recognizes the need to consider incorporating technology into the NAEP assessment in response to the growing use of technology in education. Therefore, it is important that NAEP is not locked into one form of administration; however, issues may preclude transitioning NAEP into a computer-based test. For example, the science assessments that require hands-on demonstrations or procedures cannot easily be computerized. In contrast, many students who are learning to write on computers may have difficulty completing quality work on a paper and pencil format test. NAGB admitted that many states are ahead of NAEP in incorporating technology into its assessment. In the future, NAGB will look at incorporating technology into the frameworks by considering measuring technology literacy based on frameworks from the National Academy of Engineering.

Develop assessment framework and test specifications

According to NAGB policy (NAGB, 2002b) contractors for framework development are selected based on a competitive process facilitated by NAGB. The proposal evaluation team includes NCES, Board members, and outside individuals. The Board helps in developing the SOW for the RFP and a subset of these individuals (who help develop the SOW) help in reviewing proposals. Crovo is involved in both the Board meetings and contractor meetings. During the process of framework development (approximately 18 months) the Board has several opportunities to review the work of the contractors and then the framework goes for Board approval. After approval, approximately 20–25 percent of the framework committee must serve on the NCES standing committee for the item development process.

The Framework Development policy (NAGB, 2002b) describes who is involved in the process and documents the need to have content experts, educators, members of the public, and policymakers on the panel. There is an international perspective to these frameworks as many individuals on NAEP framework committees have served also on international assessment committees (e.g., PIRLS, TIMMS, PISA). NAEP was able to borrow from these frameworks and subsequent research has examined the overlap between these frameworks. Because the typical NAEP framework panel consists of approximately 20 percent teachers it appears that most committee members do not have classroom teaching experience. It was unclear whether the criteria for panel membership included content knowledge or familiarity with the target population of students.

The NAGB Framework Development Policy (NAGB, 2002b) specifies seven guiding principles by which these frameworks should be developed. This policy lists the following seven principles as those guiding the framework development:

Principle 1: The Governing Board is responsible for developing an assessment framework for each NAEP subject area. The framework shall define the scope of the domain to be measured by delineating the knowledge and skills to be tested at

each grade, the format of the NAEP assessment, and preliminary achievement level descriptions.

Principle 2: The Governing Board shall develop an assessment framework through a comprehensive, inclusive, and deliberative process that involves the active participation of teachers, curriculum specialists, local school administrators, parents, and members of the public.

Principle 3: The framework development process shall take into account state and local curricula and assessments, widely accepted professional standards, exemplary research, international standards and assessments, and other pertinent factors and information.

Principle 4: The Governing Board, through its Assessment Development Committee, shall closely monitor all steps in the framework development process. The result of this process shall be recommendations for Board action in the form of three key documents: the assessment framework; assessment and item specifications; and background variables that relate to the subject being assessed.

Principle 5: Through the framework development process, preliminary achievement level descriptions shall be created for each grade being tested. These preliminary descriptions shall be an important consideration in the item development process and will be used to begin the achievement level setting process.

Principle 6: The specifications document shall be developed during the framework process for use by NCES and the test development contractor as the blueprint for constructing the NAEP assessment and items in a given subject area.

Principle 7: NAEP assessment frameworks and test specifications generally shall remain stable for at least ten years. (p. 3–4)

Often, the frameworks are informed by standards for national learned societies; however, the frameworks do not necessarily follow these standards. Where possible, they are included as one piece of information to be considered. Some examples are listed here:

- Economics—follows fairly close
- Science—two sets of national standards exist
- Reading—no national standards exist

Given the lag time between framework development and administration of the operational NAEP assessment, the framework development process requires forward thinking (where do we want to be in X number of years when this assessment becomes operational?) and the need to reflect best practice. The panel works well together and is not dominated by one type of panel member (e.g., policymakers, teachers).

Frameworks are reviewed whenever there is a major change in the direction of state or international assessments. The decision to change a framework is weighed between the desire to maintain a trend in the assessment and wanting to keep the assessment current. For example, in a survey of state policy makers concerning the 2005 NAEP mathematics assessment it was apparent that an update was needed in fourth- and

eighth-grade mathematics but the desire was to maintain trend. The geography framework will be ready for an update in 2010 and the subgroup will revisit the framework but again, there is the desire to maintain trend.

Test specifications are often developed in parallel with the frameworks. The proportional weighting of content is determined by emphasis at grade level. The content requirements are the first priority followed by the appropriate item formats given the objectives within the frameworks.

Develop items and background questions

The NAGB NAEP Item Development and Review policy (NAGB, 2002c) lists the following principles as guiding the item development and review process:

Principle 1: NAEP test questions selected for a given content area shall be representative of the content domain to which inferences will be made and shall match the NAEP assessment framework and specifications for a particular assessment.

Principle 2: The achievement level descriptions for basic, proficient, and advanced performance shall be an important consideration in all phases of NAEP development and review.

Principle 3: The Governing Board shall have final authority over all NAEP test questions. This authority includes, but is not limited to, the development of items, establishing the criteria for reviewing items, and the process for review.

Principle 4: The Governing Board shall review all NAEP test questions that are to be administered in conjunction with a pilot test, field test, operational assessment, or special study administered as part of NAEP.

Principle 5: NAEP test questions will be accurate in their presentation and free from error. Scoring criteria will be accurate, clear, and explicit.

Principle 6: All NAEP test questions will be free from racial, cultural, gender, or regional bias, and must be secular, neutral, and non-ideological. NAEP will not evaluate or assess personal or family beliefs, feelings, and attitudes, or publicly disclose personally identifiable information. (p. 3)

These principles are detailed in specific procedures required to satisfy each policy requirement. After the items are created, a clearance package is created that shows the content match and the intended cognitive level. This information is then shared with the Board.

NAGB's involvement in item review is the representation of the framework committee to the item development committee and the Board (by law) looks at bias and appropriateness of each item. Before this review, training is conducted on item development policy and general process for good items. These training materials were not provided. During this review the Board does have the right to comment on other item characteristics. Any comments on items are sent to NCES; however, comments are quite

rare. The Assessment Design Committee (ADC) does a separate review of items by teachers, principals, and policymakers.

NAGB does review the reading passages that are included in NAEP assessments. The Board is given a booklet of passages and a large number are reviewed at once. The Board is responsible for ensuring that passages are engaging, appropriate, and current. Each passage receives a rating of “definitely use,” “possibly use,” or “definitely not use.” Many of the passages are taken from published texts so edits are not always possible. Approximately 15–20 percent of the passages are rejected during this process. NAGB’s comments on passages are funneled through NCES to ETS. In this three step process, NAGB first reviews the passages, then the passages, items, and scoring guides, and finally the passages, items, scoring guides, and pilot data (passages are reviewed three times). NAGB reviews reading passages first to assist with the efficiency of the development process. If a passage is rejected, there is no need to write, review, or pilot test items that would be related to the passage.

The ADC of NAGB also has the responsibility of reviewing all the subject-specific background questions (e.g., how many science classes have you taken) and the reporting committee reviews the generic background questions. Based on policy (NAGB, 2002a) NAGB is responsible developing the framework and specifications for these questions including specification of which topics should be included. According to policy (NAGB, 2002a) NAGB is responsible for reviewing the questions under federal legislation P.L. 107-110 based on the following criteria:

- A. Background information is needed to fulfill the statutory requirement that NAEP report and analyze achievement data, whenever feasible, disaggregated by race or ethnicity, gender, socio-economic status, disability, and limited English proficiency. Non-cognitive data may enrich the reporting and analysis of academic results, but the collection of such data should be limited and the burden on respondents kept to a minimum.
- B. All background questions must be related to the primary purpose of NAEP: the fair and accurate presentation of academic achievement results.
- C. Any questions on conditions beyond the school must be non-intrusive and focused on academic achievement and related factors.
- D. Questions shall be free from racial, cultural, gender, or regional bias.
- E. All questions must be secular, neutral, and non-ideological. Definitions of these terms, accompanied by clarifying examples, are presented in Appendix A [of NAGB’s document], as adopted in the Governing Board Policy on NAEP Item Development and Review.
- F. NAEP must not evaluate or assess personal feelings or family beliefs and attitudes unless such questions are non-intrusive and have a demonstrated relationship to academic achievement.
- G. Issues of cost, benefit, appropriateness, and burden shall be carefully considered in determining which questions to include in background questionnaires. These

factors must also be considered in determining the frequency with which various questions shall be administered and whether they shall be included in both national and state samples.

- H. Background questions that do not differentiate between students or have shown little change over time should be deleted or asked less frequently and to limited samples. (p. 5)

Set achievement level standards

After the framework and test specifications committee has decided on content, it is asked to craft achievement level descriptions based on policy. This information is given to those responsible for developing items (to ensure coverage of ability levels) and those responsible for setting achievement level standards (these experts will also finalize the achievement level descriptors). During the standard setting process, these achievement level descriptors are revised and are tweaked to be readable and marketable to the public. This means that the performance level descriptors used in the development of the assessment may change from development to achievement level setting.

ACT/Pacific Metrics was awarded the most recent RFP for standard setting work using the new Mapmark methodology (ACT, 2005a; 2005b; 2005c; Schulz and Mitzel, 2005). Before implementing this new methodology, NAGB first asked for work that assessed the impact of using the Mapmark method to set the achievement level standards. It was suggested that this new method should be compared to Angoff using the eighth-grade math exam. The results of these studies are documented in the ACT reports shared by NAGB. The Board felt it was appropriate to use the Mapmark method as this is close to the Bookmark method that is being used by many states. The field test of this method was based on a less than desirable sample and used imputations (ACT, 2005b).

The Mapmark method was developed to look at “domains” or “clusters.” The method was designed to allow panelists to make more informed decisions by providing better feedback data. There are two types of domains (1) those developed by NAGB and (2) those developed by ACT (teacher domains, stages in the curriculum, and content domains). The Mapmark is a test based standard setting method (as compared to Angoff which was described as item-based).

NAGB specified that using a new standard-setting methodology is not a rejection of the method that was used for previous assessments (e.g., the old standards are still usable for past administrations). Although it did not believe that the public would notice, the research community will be aware of the change.

Write, review, issue, disseminate reports and data

NAGB is responsible for initial reports, the Web site for initial release of NAEP results, individual state and district reports, special reports, and pilot studies. Other reports that are not special reports are not the responsibility of NAGB. In addition, inclusion reports by NCEES are not viewed as initial releases of data and are therefore not the responsibility of NAGB.

In documented NAGB policy (NAGB, 2004a) the Board has listed policy principles and guidelines for reporting that specify the focus of the reports, the intended audience, rules for reporting sub-group information, and information to be included. This list of policy and guideline statements defines the extent to which NAGB influences the

content of the report before the writing begins. In addition, the NAGB policy on 2005 report specifications (NAGB, 2004c) includes reporting requirements that focus on the structure and presentation of different types of results for the reports and the Web sites.

NAGB is responsible for reviewing the reports (even at the outline stage) that affords them opportunities to make suggestions for change to the proposed content or framework. Although NAGB does not appear to be responsible for writing these reports, it is involved in the extensive, multistage review process. NAGB staff members indicated during our visit that a documented flowchart describing this process exists; however, to date, this document has not been made available. NAGB is given several opportunities to provide feedback on the reports during the review process:

1. Format—NAGB can comment on the proposed format of the report and specifically highlight any ways in which the NAGB policy for reporting is violated.
2. Proposed content—NAGB can look at the proposed content, executive summary, and table shells of the report. Comments are gathered from the staff and Board and are sorted into four categories:
 - a. Policy issues (these are nonnegotiable changes to be made)
 - b. Strong recommendations
 - c. Questions, needed clarifications
 - d. Editorial comments (grammatical issues)
3. Final Proof—The Board has final say on whether or not to release the NAEP reports. To date, it has not held the release of any report and suggested that the only reason why one would be held is in the case of a policy violation.

As far as inclusion policies, NAGB is aware of the variation across states in how students from special populations are included in NAEP assessments. With the implementation of *NCLB*, this problem has only gotten worse because of how these students could be identified by the states. The goal of NAGB is to have more equitable inclusion criteria. To explore this issue, NAGB asked state representatives for their opinion on this matter. Many states currently offer students an alternative assessment (NAEP does not have an alternative assessment), and several states use various accommodations that, if used on the NAEP assessment, would change the content of the test (e.g., reading a student the passage on a reading test). To deal with this problem NAGB brought a nationally representative panel together for two days to discuss this inclusion issue and had other panelists review its work. The technical report or meeting minutes from this activity were not provided. This widespread participation in this project increased support from the states. This work resulted in a decision tree that focused on inclusion of these students rather than exclusion. This decision tree was provided in subsequent communication with NAGB.

The motivation issue appears to be most prominent at the 12th grade level. One of the five recommendations from the National Commission on 12th Grade NAEP Assessment and Reporting (National Commission, 2004) was that NAEP's leaders find ways to increase motivation of the 12th grade students as results showed low participation rates (as compared to other grades) and low motivation as indicated by unanswered questions. During the November 2004 NAGB Board meeting, Reingold Inc. (2004) presented a study exploring the use and success of public-private partnerships as a way to increase motivation in testing. More recently, NAGB conducted a session at the National Large Scale Assessment Conference where members of the public (mostly

NAEP State Coordinators) were asked how to improve motivation of 12th grade students (NAGB, 2005).

NAGB is responsible for the dissemination of many NAEP reports (see list at the beginning of this section) and has published a reporting schedule for the 2005 assessments on the Web site (NAGB, 2004b).

NAGB has the responsibility for responding to complaints about NAEP. As indicated by NAGB policy (NAGB, 2002d) the process for complaints are as follows. All complaints should be sent in writing to NAGB and the executive director will make a decision (administrative determination) after consultation with the commissioner for education statistics. This administrative determination can be appealed and the chair of the Board will determine if the appeal needs to be discussed by the full Board or by a group of Board members.

Renew and improve the assessment

Work by contractors is received by NAGB staff and reviewed for quality. Quality control procedures are commonly built into the process for completing the work. The responsibility for reviewing contractor work is held by the staff member (varies by contractor) that is most closely associated with the work. For example, Lawrence Feinberg would likely monitor and review work from contractors that conduct studies relevant to the reporting and dissemination committee. However, it is more common that the work of contractors is reviewed or monitored by more than one NAGB staff member.

Final comments

The NAGB staff reiterated that the strength of the Board was the way in which it was created and how it works. One staff member noted that “This is not an inside the beltway board.” This group represents a wide range of expertise and background. The Alexander-James report (Alexander and James, 1987) noted the need for such a policy body to assume responsibility for these activities. It was important that this board be nationally representative and independent from the federal government.

NAGB reiterated the need to improve NAEP reports by making them more understandable for the general public and thus increasing the utility of these reports. NAGB staff members referenced studies by Hambleton (1997; 2002) that suggest most policy makers got their NAEP information from the popular media (e.g., newspapers). In the past, NAEP reports have been filled with statistical jargon that makes many of the reports unreadable to the general public. Congress identified this problem and transferred the responsibility from NCES to NAGB for overseeing report preparation and release of the reports. NCES and NAGB each have their own standards for reporting and these contain some conflicting ideas. NAGB realizes that if NCES were to take all of NAGB’s recommendations for reporting it would violate many of its own policies and therefore, many of the comments on reports are discussed between the two organizations.

NAGB underscores the need to avoid the “black box” terminology that makes NAEP processes sound secretive. NAGB feels that with educational modules many concepts (e.g., scaling, conditioning) should be understandable by the public.

The final key to this reporting issue is that timely reports are produced for the public without sacrificing quality control of the information that is reported. The challenge comes with making sure that the contractors have the staff and organization to

execute this plan for timely reporting—both NAGB and NCES need quality control checks in place.

Findings and Recommendations

Overall we would like to commend the NAGB staff members for their work with the national assessment. The range of expertise represented by the core staff appears to be well matched to the various functions that NAGB is responsible for. It was very apparent from the conversations during the site visit that the staff is highly capable of handling the workload before them and facilitating the work of the Board. In addition, we would like to acknowledge NAGB's efforts to attend to the needs of the public. This focus is apparent in several facets of NAGB work. For example, the efforts by NAGB to improve the reporting show a focus on the needs of the public and increase the usability of the reports by making them readable by a larger audience. These efforts are supported by research and advice from experts and reputable agencies in the field of reporting and dissemination.

The staff and Board should be commended for their ability to adapt to changes in federal education policy (i.e., *NCLB*) which has changed the focus of NAEP and increased the public visibility of the results. Although the Board interprets its role independent, it may have been inevitable that this change in K–12 educational policy at the federal level would alter the focus of NAEP and influence the way in which NAEP was administered and reported.

The policy documentation from NAGB staff and Board is also noteworthy. Many documents described in this report were available on NAGB's Web site and constitute official policy documents that detail principles and guidelines for the execution of activities that are under the responsibility of NAGB. These documents are available to the general public and are very readable.

Based upon information collected in the onsite interviews, observations at NAGB Board meetings, and review of documents, we also have a few recommendations that could benefit NAGB's operations and the NAEP. First, in presentations and documents, NAGB has stated the intended and unintended uses of NAEP data and results. However, it is apparent that NAEP data and results are being used for purposes not included in the intended scope. Although NAGB does not have the power to enforce proper use of NAEP data and results, the policy body is strongly encouraged to be more vocal about its position on the proper and improper use of NAEP data. For example, although NAEP is specified to report on the larger group level (primarily national and state level), school-level data are being computed and used in disseminated research (e.g., McLaughlin and Bandeira de Mello, 2005).

Second, the validity evidence for use of NAEP results appears to come from various sources including the Secondary Analysis Grants, the NAEP Validity Studies Panel (NVS), the NAEP–Educational Statistical Services Institute (NESSI), Assessment Development, and the Task Order Component. However, there does not appear to be a unified validity framework for the program. Moreover, there does not appear to be an individual, panel, or agency that is responsible for synthesizing this information and ensuring that it is used to improve the NAEP system. As the policy body for NAEP responsible for interpreting and determining the scope and use of NAEP assessments, we believe that NAGB should take a leading role in this effort to define the validity

framework. More importantly, it could serve a key role in monitoring the validation and research efforts conducted under this framework.

Finally, there continue to be differences of opinion regarding the roles and responsibilities of NAGB and NCES. This results in part from the lack of clarity of the legislation, but also from the differential interpretation of the NAEP legislation.

Materials reviewed:

ACT. (May, 2005b). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Special studies report*. Iowa City, Iowa.

ACT. (May, 2005c). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Technical report*. Iowa City, Iowa.

ACT. (April, 2005a). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Process report*. Iowa City, Iowa.

Alexander, L., and James, T.H. (1987). *The nation's report card: Improving the assessment of student achievement*. Stanford, Calif.: National Academy of Education.

Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scales and reports more understandable? In R. W. Lissitz and W. D. Shafer (eds.), *Assessment in educational reform* (pp. 192–205). Boston, Mass.: Allyn and Bacon.

Hambleton, R. K. and Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Teaching.

Institute of Education Sciences. (2005). *National Assessment of Educational Progress Secondary Analysis Grants: Request for Applications*. Retrieved on June 29, 2005, from www.ed.gov/about/offices/list/ies/programs.html.

McLaughlin, D. and Bandeira de Mello, V. (June, 2005). Comparing NAEP and state measurement of achievement standards, trends, and gaps: Methods, issues, and results. Presentation at the annual National Conference on Large-Scale Assessment, San Antonio, Texas.

National Assessment Governing Board. (June, 2005). Twelfth grade NAEP-new mission and purpose in an era of high school reform? Presentation at the annual National Conference on Large-Scale Assessment, San Antonio, Texas.

- National Assessment Governing Board. (2004a). Policy statement on reporting, release, and dissemination of national assessment results. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2004b). *Reporting Schedule for 2005 NAEP assessments*. Washington, D.C. U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2004c). *Specifications for NAEP 2005 reports*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2004d) *Mathematics Framework for the 2005 National Assessment of Educational Progress*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002a). *Collection and reporting of background data by the national assessment of educational progress*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002b). *Framework development*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002c). *NAEP item development and review*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002d). *White paper: Prohibition on using NAEP to influence state and local standards, tests, and curricula*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board. (2002e). *Policies and procedures for complaints related to the national assessment of educational progress*. Washington, D.C.: U.S. Department of Education. Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- National Assessment Governing Board—November 2004 Board meeting, Unpublished Briefing Book.
- National Assessment Governing Board—May 2005 Board meeting, Unpublished Briefing Book.
- National Commission on NAEP 12th Grade Assessment and Reporting. (2004). *12th grade student achievement in America: A new vision for NAEP*. A report to the National Assessment Governing Board. Retrieved on June 8, 2005, from <http://www.nagb.org/>.

Reingold, Inc. (2004 November). Summary of research findings for the national assessment governing board. Unpublished report presented at the NAGB Board Meeting, Washington, D.C.

Schulz, M.E. and Mitzel, H. (April, 2005). Research and developments in standard setting. Paper presented at the meeting of the National Council on Measurement and Education, Montreal, Canada.

Title III–*National Assessment of Educational Progress Authorization Act*, Pub. L. No. 107-279. (2002). Retrieved on June 8, 2005, from <http://www.nagb.org/>.

This page intentionally left blank

Appendix G-2: National Center for Education Statistics (NCES)

Site Visit Team: Chad Buckendahl and Susan Davis, Buros Center for Testing
Date of Site Visit: June 6th, 2005

Audit Summary

Staff

Peggy Carr – Associate Commissioner
Andy Kolstad - Senior Technical Advisor for NAEP
Andrew Malizio – Program Director, Assessment Development and Quality Assurance
Janis Brown – Statistician, Assessment Development and Quality Assurance
Arnold Goldstein – Statistician, Reporting and Dissemination
Suzanne Triplett - Program Director, State Support and Constituency Outreach

Prior to the visit, Buros provided NCES with a list of topics and approximate time estimates for each topic they would like to discuss during the meeting. In turn, NCES reacted to an agenda provided by Buros and a copy of the NAEP Alliance Statement of Work (SOW) and vitae for all key NAEP staff. Buros also reviewed documents in preparation for the site visit (list provided at the end of the site visit report).

Peggy Carr provided introductions for everyone and Chad Buckendahl briefly reviewed the purpose for the visit and the agenda for the day. Throughout the day, NCES staff members were available to respond to questions. Marilyn Binkly, one of the NAEP contracting officer's representatives (CORs), was not available for the June 6 meeting. Following the site visit, Buros conducted a teleconference with her on July 6, 2005, about the NAEP item development process. Responding to a recommendation from NCES, a subsequent interview was conducted with Marilyn Seastrom, chief statistician for NCES.

Organizational characteristics

The key NCES staff members for NAEP operations and their respective responsibilities are as follows. Arnold Goldstein is the COR for ETS and is responsible for overseeing the NAEP reporting (both internal and external), dissemination of NAEP reports, and special populations work such as private schools, charter schools, students with disabilities (SD) and English language learners (ELL). Suzanne Triplett is the COR for Hager Sharp, is responsible for overseeing NAEP on the Web, constituency outreach, and serves as an advisor to Peggy Carr. In addition, she is responsible for overseeing the work of states, more specifically, the state NAEP coordinators. Rima Zoyban (not in attendance at the June 6 meeting) works with Triplett to coordinate the efforts of the state coordinators. Andrew Malizio is the COR for PEM and oversees the materials distribution and scoring for NAEP. He is also the project director for the assessment development and quality assurance, and assists with the budget under Peggy Carr. Janis Brown is the COR for HumRRO, coordinator for NAEP Validity Studies Panel (NVS), the Design and Analysis Committee (DAC), improvement activities, and the COR for the High School Transcript Studies. Peggy Carr is the associate commissioner for the division and is responsible for general oversight and management of NAEP. Andy Kolstad is the senior technical advisor for NAEP and his primary responsibilities include

reviewing publications, assisting in making decisions regarding design features for NAEP operations, and being involved in research to develop new statistical or psychometric methods.

A consistent comment during our visit was NCES's need for more full-time staff members on both the technical and managerial level. Due to limited technical staff, NCES relies heavily on the NAEP–Education Statistics Services Institute (NESSI). NESSI is a division of AIR but is designed to operate separately from other AIR operations related to NAEP. Currently, NCES has approximately 39 NESSI staff working on NAEP. However, NESSI staff members are outsourced and not supervised by NCES. This outsourcing strategy is needed due the challenges of creating new positions, even if the functional roles would be better served by in-house staff. NESSI has high turnover of staff; however, NCES has input when new staff are hired to allow for match between applicants' skills and needed qualifications. Contractors (including NESSI) have expressed concerns with finding high quality staff given the increased importance on testing. Additional information was gathered about the scope of NESSI following the NCES site visit.

There was also an observed need for NCES to employ additional managerial staff. Current staff members mentioned a specific need for a person to oversee the NAEP budget. NCES CORs have substantial responsibilities in addition to overseeing the work of NAEP contractors. When compared to other divisions in NCES, the Assessment Division is forced to respond to the unique challenges of NAEP. While other divisions have staff solely responsible for overseeing one contract—this is not possible for the NAEP division given the current staffing limitations. One staff member also indicated the need for a deputy director to help coordinate all NAEP efforts and work of contractors. In a later part of the discussion, staff members mentioned that it would be helpful to have additional staff dedicated to compiling and organizing NAEP validity research and identifying areas in which research from different agencies could be used to improve NAEP operations. Drew Malizio and Janis Brown are jointly working to try to broadly oversee the validity framework, but it was noted that these efforts are in their early stages. More importantly, these efforts could not interfere with the operational activities for which they were already responsible.

The ability of NCES to hire additional staff is in large part related to availability of funds within the division budget. The Assessment Division of NCES has two budgets: one for projects and one for salary and travel. This arrangement of the budget has resulted in limited travel funds, which the NCES staff feels limits their ability to do their job well. This is a department wide problem; however, the Assessment Division has more travel needs because of the operational and logistical demands of NAEP. Proposals for an integrated budget have failed in Congress. Because it is an important component of the quality control processes within NCES, CORs must have the opportunity to closely monitor the activities of contractors for which they are responsible. As many of these activities occur outside the Washington, D.C., area, travel funds are needed for effective contract management and accountability.

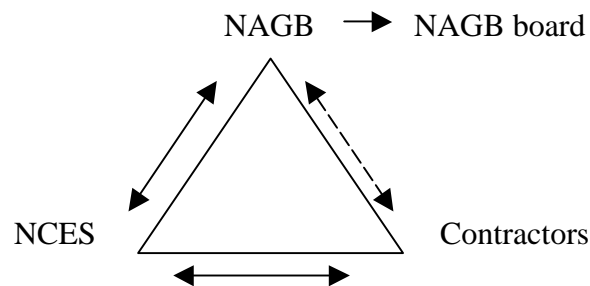
At the time of this data collection, there was an acting commissioner of NCES (Russ Whitehurst, director of IES). NCES staff felt the need for this position to be filled, as this person would help with the workload and serve as an advocate for the Division when differences in interpretation with NAGB regarding the legislation that sponsors NAEP occur. Specifically, the commissioner serves as a primary contact for NAGB and

facilitates NAGB requests for information and activities. Since October 2005, Mark Schneider has served as commissioner of NCES.

Communications

There are several mechanisms for communication within the NAEP system. For internal communication among the staff, Carr occasionally holds staff retreats to discuss issues and solve problems. One staff member used Figure 1 below to explain the communication structure within the NAEP organization. NCES is responsible for communication with the contractors and communication between the contractors and NAGB staff. The NCES CORs are in contact almost daily with their respective contractors for a variety of purposes (contractors must consult NCES before making any major design decisions). More formal teleconferences between the contractors and NCES are held approximately every two weeks. Other communication is Web-based. The Alliance contractors created the Integrated Management System (IMS) for virtual discussions, sharing of materials, and review of materials. The IMS system appears to offer NCES the ability to monitor discussion and work among the contractors within the Alliance. Suzanne Triplett uses WebEx for Web conferencing with NAEP State Coordinators; each week, there are three training sessions that state coordinators can attend. These are recorded and can apparently be played back at a later date. In addition, the state NAEP coordinators are brought together twice a year for group meetings.

Figure 1—NAEP communication structure



NCES and NAGB hold two joint meetings prior to each board meeting to discuss the meeting agenda and materials needed prior to the meeting. The first is six months prior and the second is approximately three weeks prior. Occasionally, NAGB has initiated direct contact with the NAEP contractors which can lead to confusion when information is shared in this manner. Depending on the nature of the communications and the requests, it also has the potential to challenge contractors to remain within the agreed upon scope of work.

Responsibilities

Throughout the site visit, NCES described different facets of their role but primarily focused on their managerial role in NAEP. NCES is responsible for executing board policy decisions and overseeing the work of NAEP contractors. The focus of NAGB's role is policy-oriented. The NAGB staff members are responsible for facilitating

board meetings and preparing issues and information to bring before the board. The responsibility of the Board is to exist as a policy body for NAEP that is independent of the government.

Recent legislation (P.L. 107-279) has changed the policy for preparation and dissemination of NAEP reports. The new legislation appears to expand NAGB's role into areas that had historically been within the purview of NCES leading to confusion about responsibilities. This confusion has likely facilitated some differences of opinion between NCES and NAGB regarding the interpretation of this legislation. Clarification of the roles is a critical step for the program.

Reference was made to the KMPG (1996) study which indicated that NAGB occasionally infringed on the operational side of NAEP. By legislation, NAGB is responsible for oversight of policy. However, due to the technical expertise of some NAGB staff, there are occasions when NAGB becomes involved in more of the operations side of NAEP. In turn, decisions or policies made by NAGB in these instances often overlap with existing NCES policies. Examples included NAGB establishing a policy for participation rates when policies already exist for these data in NCES Statistical Standards, the specifics mandated by NAGB for the execution of the fall pilot study, and requests for projects or changes to frameworks that are outside the bounds of NAEP's limited budget (e.g., addition of a vocabulary scale to reading, foreign language assessment).

When a difference of opinion arises between NCES staff and NAGB staff, the issue is first taken to NAGB staff. If this is not a viable solution or does not resolve the issue, Carr may need to take the issue to the commissioner for resolution or assistance.

Intended scope and uses of NAEP assessments

NCES indicated that it has a minimal role in defining the intended scope of NAEP assessments and that NAGB is responsible for creating the frameworks and test specifications. NCES is invited to attend these planning meetings. NCES is responsible for translating the frameworks and test specifications into the operational NAEP assessments.

NCES listed six sources of validity evidence within the NAEP system that can be used to support the inferences made from NAEP data.

1. NAEP Validity Studies (NVS) Panel: The NVS is facilitated through AIR's Palo Alto office. The NVS prioritized several validity issues in *An Agenda for NAEP Validity Research* (AIR, 2002). Research by this group has resulted in work that has been presented at conferences and published by the NVS on comparing state assessment and NAEP data, on inclusion and exclusion policies, and accommodations.
2. NAEP Design and Analysis Committee (DAC): The DAC does not necessarily conduct or set an agenda for validity research in NAEP; however, in its advisory capacity to NAEP, their work relates to validity issues. The DAC deals with real time problems and monitors ETS's assessment development and maintenance activities. The DAC focuses primarily on methodologies, statistical quality, and provides technical advice.
3. Task Order Component (TOC): This is a subset of the NAEP Alliance contract and involves specific research studies requested by NCES or NAGB and may include quick

turnaround projects (e.g., Inclusion decision tree) that are requested by NCES throughout the duration of the contract.

4. **Assessment Development:** Much of the work conducted and documented by ETS during the development of the assessments can be viewed as contributing to validity evidence (e.g., attribute study—how much of an item is related to an irrelevant construct). These procedures, methodologies, and results are included in technical reports; however, the most recent publicly released technical report is from the 1999 Long Term Trend and may not reflect current procedures. A Web site is currently under development that would present the technical report online.

5. **NAEP–Education Statistics Services Institute (NESSI):** As a subsidiary of AIR, this group may conduct special studies related to validity as part of their broader responsibilities for NCES. For example, one study focused on analyses conducted with a reliance on the assumption of a normal distribution of scores.

6. **Secondary Analysis Grants:** Work from these grant projects sometimes contributes to the validity framework of NAEP. For example, some work on accommodations has come from this program that has helped inform NAEP policy. However, because these are run as a grant program, there is little input or control over the final products of this work. A revision in the proposal review process has appeared to focus the priorities of the program and incorporated an external, independent process for proposal review and selection.

In addition to the six sources noted above, NCES also reviews work by contractors to consider any validity implications (e.g., work on Full Population Estimates that arose from the state analysis project). The issue of perceived competition between contractors was addressed during the discussion. NCES feels that even though there is some overlap in work conducted by contractors, the resulting competition can be beneficial for NAEP (e.g., ETS released software used to conduct their analyses because AIR distributed a similar version). Some competition is fostered by NCES to get the best work possible and these contractors are encouraged to take this work to the professional community through conference presentations and professional journals.

12th Grade NAEP

In response to the 12th grade “preparedness” issue (e.g., Are students prepared for the workplace, college, and the military?), NCES wrote a 30-page response that included several foreseeable challenges related to this proposed change for NAEP. Buros requested that NCES share this document for our review but did not receive a copy. Included in this list was the issue of motivation which NAGB is currently considering with the help of Achieve. Given the preliminary stages of this effort, the operation of NAEP in the near future is not likely to be affected. Given their relevance to this effort, NAGB now wants to look closely at the High School Transcript Studies. It appears that through the HumRRO Quality Assurance panel, NCES is also investigating the issue of motivation.

Develop items and background questions

Cognitive Item Development (Information gathered during telephone conference with Marilyn Binkly on July 6, 2005)

NCES is responsible for overseeing the item development process and making sure it follows the specific frameworks created by NAGB. Specifically, the process is overseen by standing committees made up of roughly 12–20 content specialists from the national, university, state, and local levels. Typically, one-fourth to one-third of the members of the standing committees will also be members of corresponding framework committees. The standing committees meet between two and four times per year.

In the first phase of cognitive item development, pilot items are written by different contractors based on content area. AIR is responsible for developing items for the writing and social science assessments and background questions. AIR hires content specialists and trains them on item writing procedures and their work is supervised by AIR staff. ETS is responsible for writing items for the reading, math, and science assessments. The content specialists at ETS are employed on a permanent basis and may work on other projects in addition to NAEP. NCES oversees both contractors and helps with the training of the item writers to ensure the items conform to specifications and fit the frameworks specified by NAGB. Roughly twice as many pilot items are written as will be included on the final NAEP assessment to account for attrition that may occur during the piloting process.

The items are then passed to NCES and the standing committee for review. Each item (with related scoring guides when appropriate) is individually examined for match to the NAGB framework, appropriateness of the difficulty level, clarity of the question and response options, and appropriateness of scoring. Items may be rewritten by the group during the review process to achieve greater agreement among the reviewers. The items are returned to the contractors for revision, and then sent back to the steering committee for further review. A larger goal of this process is to ensure that the frameworks are being properly interpreted by the contractors (i.e., did the contractors do their job in writing items to match the NAGB framework). Also at this point, the standing committee may determine that the frameworks need additional clarification.

After the standing committee has completed their review of the items, NCES conducts a state item review. NCES pays for two representatives from each state to participate in the review (states may send more representatives at their own expense). The state representatives may be curriculum specialists, state testing coordinators, or teachers. While the feedback from these representatives may not directly affect which questions will ultimately appear on NAEP, NCES and ETS review the representatives' comments and concerns and take action when appropriate.

When the standing committee has finalized its choice of items, these items are submitted to NAGB who makes the final determination as to which items will appear on the pilot tests.

The pilot test is administered to a nationally representative sample of about 500–1,000 students, representing the full range of ability. At least two items are pilot tested for each operational item that is needed. Item statistics are analyzed and items and item blocks are examined for difficulty and possible bias (DIF analysis). Items may be dropped or reworked if necessary. The results are reviewed by the standing committee, and in the case of the reading and math assessments, the items may undergo a second

pilot test. The items and item blocks that performed well then go on to make up the operational exams. NAGB has one final review of the items before the assessment becomes operational.

Cognitive item development is a continuous process. Roughly every ten years new frameworks are developed which require updated item sets. Also, about one-fourth to one-third of NAEP items are released after each assessment. Therefore, continual replenishment of the item pool is necessary. NCES and the item development contractors determine which items to release so that the items are representative of the NAEP assessment.

Three sources of quality control were noted for the item development process. First is the extensive review process. Items are reviewed by the standing committee, by the state reviewers, and by NAGB. This multistage process is used to ensure match to content specifications, test specifications, appropriate difficulty, and fairness. Second are the statistical analyses that are incorporated within the item development process. Specifically, DIF analyses are used to examine bias and sensitivity across groups, the relative performance across ability levels, and performance is explored across time (by large samples and as a group comparison). Third, at each review session, NCES collects comments about each item and is forming a coding system to organize these comments.

The trend assessments process is slightly different than that described above. First, these assessments are not based on frameworks like the Main assessments are. The content was defined by the trend assessments that were constant in the mid to late 1980s. Since this time, some items have been replaced with the new items being reflective of the retired items. Bridge studies are currently being conducted to determine if this modified assessment is measuring the same content as the old assessment.

The background questions are developed in much of the same way as the cognitive items. Background questions are included in student assessments, in teacher surveys, for SWD and ELL surveys, and in principal surveys (to assess the demographics of the school). The purpose of the background questions is to unobtrusively gather information to aid in the interpretation of cognitive item database. NAGB is responsible for developing the frameworks and item specifications for the background questions and AIR is contracted to develop these items. There are three types of items developed:

- 1) Reporting—these items are used in NAEP reports and include such variables as region of the country and ethnicity.
- 2) Subject specific—these items measure students' experience with subject matter and related variables.
- 3) Other contextual variables—these are designed to measure equitable distribution of resources and opportunity to learn.

After development by AIR, the background questions are submitted to the standing committee for review and follow a similar process used for the cognitive items. To maintain consistency, many of the same background items are used year after year. In addition, effort is made to maintain consistency of items across tests (subjects) to allow for comparisons.

Background questions must also be submitted for approval by the Office of Management and Budget (OMB). In the past, OMB has requested item revisions. However such changes are viewed as minimal now due to the consistency of items across years.

All materials in development are sent password protected. Due to the documented, intended nature of NAEP (low stakes for students) there are no cheating

analyses. Any teachers who are granted permission to stay in the room during testing must sign a confidentiality agreement.

During the site visit, NCES staff shared with Buros two procedural manuals for providing secure access to NAEP materials (NCES, n.d.; 2004). These manuals detail the procedures and guidelines by which certain individuals may obtain access to secure NAEP items. To date there has been one instance of stealing; a block of materials was posted to the Web.

Write, review, issue, and disseminate reports and data

NCES is responsible for creating understandable reports. Starting in the 1980 NAEP reports became longer and longer. To deal with this issue, smaller ‘highlight’ reports were created. Given their interpretation of recent legislation regarding NAEP reports, NAGB has now assumed the role of specifying standards for how the reports shall be prepared. NCES strives to ensure that NAEP reports follow the NCES Statistical Standards (www.nces.ed.gov), but occasionally these standards conflict with NAGB’s requests for report specifications. NAGB provides specific content and editorial specifications for these reports (color, content, framework, and number of pages).

The process outlined below is the new format for report review by NAGB. At each of the listed phases, NAGB is allowed to review the report materials and provide comments to NCES. Specifically, this process is followed for Web pages, Report Cards, State reports-snapshots, and TUDA (each written for two subjects and three grades).

- 1) Outline stage—ETS provides an outline for each report. This is reviewed by NCES and NAGB.
- 2) Table Shells and Figure Designs—ETS again provides this information which is reviewed by NCES and NAGB.
- 3) Pre-division review—In this phase ETS provides the layout of the report without the data which is then reviewed by NAGB and NCES.
- 4) Center-wide review—This includes two individuals from other divisions and the chief statistician. Once the chief statistician approves the report, the review goes to the commissioner.

The review comments provided to NCES by NAGB are now a set of joint comments from the board and the staff. The comments provided by NAGB are grouped in three categories:

- Possible violations of NAGB policy
- Editorial and Design
- Other

Before 2005, NAGB staff members were allowed to look at reports and staff would make policy comments (board members were never involved). Now, with the change of policy, NAGB provides much more in-depth comments. Occasionally, NCES will negotiate comments and requested changes with NAGB until consensus can be reached. In the six-month review process it is NOT common to have outside reviewers. In addition, because the six-month timeline is so short, these reports do not go through IES review (Marilyn Seastrom can sign off on these reports).

NAEP technical reports follow a different review process. Starting with the 2000–2001 report, the technical reports will be all Web-based and they are working to build this framework and the core elements. This format is intended to allow quicker production of the reports. Lack of staffing was mentioned as one reason for the delays in getting these reports out as these are of lower priority in comparison to the other reports and activities that are ongoing. The 2000–2001 technical reports were expected to be finalized during the summer of 2005. As of Feb. 1, 2007, this report had not yet been released.

Following dissemination of NAEP reports, Hager Sharp is responsible for obtaining comments and feedback from the public. Part of this process is conducted at professional conferences. Most of the feedback about NAEP reports is positive and is focused on the highlights reports (report cards).

Renew and improve the assessment

HumRRO is responsible for reviewing and providing feedback on quality control plans submitted by all NAEP contractors. HumRRO provides content guidelines to the contractors for these reports but not formats. The CORs are provided the quality control plans by the contractors and have a chance to review them and make comments before they are sent to HumRRO. Sometimes the review by the COR must be done in parallel to the HumRRO review. Each year the quality control plans are updated. When a problem in the NAEP assessment system is noted, the processes leading up to it are revisited. Some of the requested changes are implemented before the quality control plans can be changed. Sometimes HumRRO requests processes that are actually in place but not documented properly in the QC plans.

Additional Quality Control Checks include:

- 1) Contractors are expected to have checks and balances built into their QC plans to ensure quality control throughout the process
- 2) HumRRO conducts site visits to NAEP contractors and to ensure that the checks and balances (as well as other QC components) are in place and working as planned. HumRRO provides feedback from these visits to the COR to be shared with the contractor. When HumRRO conducts these site visits they review contractor materials prior to the visit and determine what is supposed to happen and then what is actually happening. HumRRO staff from across the country assist with these site visits. If there is any problem that needs immediate attention, the COR is informed that day.
- 3) HumRRO constructed a process model of the NAEP QC plans. The first of these was static and the second includes timelines and simulation (requires software to run). This model can be used to identify feedback loops.

Due to the shortened reporting time, many NAEP contractors have moved their QC checks to earlier in the process, and there is now greater automation involved in checking the data for errors. There is also the Quality Assurance panel that talks about emerging issues and improvement models. This process helps serve as an additional independent check and prepare for unanticipated consequences.

When asked about incorporating innovative procedures into the NAEP system, NCES provided the example of using interspersed trend papers. Writing samples from prior administrations of the NAEP assessments (for calibration) are typically scored before the live scoring of the current administration. It would be better to do

simultaneous scoring but before this can happen, a study is needed to determine the impact of this change. This is typically the process for implementing new methodology; a pilot study is conducted (typically during a year when there are not large data collection needs). Also, possible topics are sent to the DAC (ETS), NVS (AIR), and/or QAP (HumRRO) for review in advance. The innovation clause was put in to encourage innovation and competition among contractors.

Final comments

NCES stressed the need to continue with activities that would move the assessment program forward including research and development studies on program methodology, psychometrics, and any efforts that help them maintain the status as the “gold standard.”

With the increased importance of NAEP and pressure to produce usable results it is difficult to keep everything current. One of these examples included the use of technology. As the scope of NAEP increases, NCES senses the need to find ways to integrate technology into the NAEP assessment.

Due to the broader visibility of NAEP, NCES finds it increasingly difficult to conduct research when its work is very much in the public eye. In turn, this will make the program hesitant to try new methodologies. In addition, NCES staff often does not have opportunities or funding to monitor current research through the professional literature or attend professional conferences at which relevant work is being discussed. Greater opportunities for professional development would make NCES staff more able to become specialized in their roles and better suited to evaluate work by NAEP contractors.

The issue of governance remains. NCES staff referred again to the KMPG (1996) study and the clarity of the responsibilities between NAGB and NCES.

Findings and Recommendations

Based on the information gathered through the site visit and review of NAEP documents, Buros would like to commend the staff at NCES for the job that they do overseeing the NAEP assessment program. Given the limited number of staff and the vast array of responsibilities placed upon this organization, the staff appears to operate efficiently in managing such a large operation. The staff noted several areas in which additional staff would be helpful and we would like to underscore the need to have a person who is responsible for managing the validity-related NAEP work. NCES noted six different sources from which this information could be drawn; however, there did not appear to be a program in place by which this information was organized in a way that could be used to improve NAEP operations. Developing, overseeing, and periodically updating a unified validity framework would strengthen the program.

In addition, Buros would like to commend the staff of NCES for continuing to strive to maintain a high quality testing program. They noted several ways in which their work and experience could be enhanced that would help them advance the testing program and improve their managerial perspective. We would like to specifically encourage the NCES staff to find ways in which NAEP work could go through the process of peer review including publishing in academic journals or presenting at conferences in this field. This would provide more opportunities for those in the field of educational testing to learn about NAEP and provide insight as ways to continually

improve the program. Although we encourage NCES to pursue these avenues they will need additional support from funding and organizational sources to have the opportunity to realize this potential.

There continues to be differences of opinion regarding the roles and responsibilities of NAGB and NCES. This results in part from the clarity of the legislation, but also from the differential interpretation of the NAEP legislation (modified in 2002).

Materials Reviewed:

- American Institutes for Research. (October, 2002). *An Agenda for NAEP Validity Research*. Palo Alto, Calif.: Stancavage, F.B., Beaton, A.E., Behuniak, P., Bock, R.D., Bohrnstedt, G.W., Champagne, A. et al.
- KPMG Peat Marwick LLP. (June, 1996). *Management and Technical Review of the National Assessment of Educational Progress (NAEP)* (study). Washington, D.C.: Huff, L.D.
- Jones, L.V., and Olkin, I. (eds.). (2004). *The Nation's Report Card Evolution and Perspectives*. Bloomington, Ind.: Phi Delta Kappa Educational Foundation.
- National Center for Education Statistics. (n.d.). NCES Statistical Standards. Downloaded from: <http://nces.ed.gov/statprog/2002/stdtoc.asp>
- National Center for Education Statistics (n.d.). *Providing Access to Secure NAEP Items: Procedural Manual*. Washington D.C., NCES.
- National Center for Education Statistics (2001). *The NAEP 1998 Technical Report*. Washington D.C.: Allen, Donoghue, Schoeps.
- National Center for Education Statistics (2004). *NCES Procedures for Providing Researchers with Access to Secure NAEP Cognitive Items: Internal Procedural Manual*. Washington D.C., NCES.
- Title III–*National Assessment of Educational Progress Authorization Act*, Pub. L. No.107-279. (2002). Retrieved on June 8, 2005, from <http://www.nagb.org/>.
- Vinovskis, M.A. (1998). *Overseeing the Nation's Report Card. The Creation and Evolution of the National Assessment Governing Board (NAGB)*. University of Michigan, Department of History, Institute for Social Research, School of Public Policy.

This page intentionally left blank

Appendix G-3: Chief Statistician - NCES

Site Visit Team: Chad Buckendahl and Susan Davis, Buros Center for Testing
Date of Site Visit: Aug. 17, 2005

Audit Summary

Staff

Marilyn Seastrom: Chief Statistician for NCES

As part of the evaluation of the NAEP audit study, Buckendahl and Davis met with Marilyn Seastrom on Aug. 17 to discuss her role in the NAEP process as chief statistician for NCES. Three of the 14 audit dimensions were discussed with Seastrom: organizational characteristics; write, review, and disseminate, data and reports; and renew and improve the assessment.

Organizational characteristics

Seastrom is chief statistician for the National Center for Education Statistics (NCES). The assessment division (which includes NAEP) is one of four divisions within NCES. Working under her, Seastrom has four NCES mathematical statisticians and two data confidentiality technicians. In addition, several NESSI staff members are available through an outsourcing contract and involved in reviewing documents and performing various quality control projects (conducted to ensure proper interpretation of new standards). Specifically, there are four full-time-equivalent (FTE) research assistants, and four FTE mid-level analysts at NESSI that work directly with her on projects and tasks. There is some turnover with NESSI staff; however, the specific needs of NCES are considered when hiring and Seastrom has been involved in the hiring process.

When asked about the effect of NCES being without a full-time commissioner at the time of this data collection, it was noted that this situation has resulted in the senior leadership developing good working relationships. However, NCES has been challenged by not having a strong advocate in this position to protect their interests. However, since Oct. 2005, Mark Schneider has served as commissioner of NCES.

An NCES staff member indicated that she has the opportunity to attend both national and international conferences and meetings each year for professional development. In addition, there are several local organizations that offer professional development opportunities for NCES staff members (e.g., Washington Statistical Society, Federal Committee on Statistical Methodology).

Write, review, issue, and disseminate reports and data

Seastrom has substantial responsibilities relating to the review of NAEP reports. One of her responsibilities is to ensure that reports meet the NCES statistical standards. These standards, published in 2002, were created through an extensive process that involved internal staff and external reviewers (NCES, 2002).

The review process for documents produced under NCES is as follows. The first step is a divisional review. For NAEP reports, reports are reviewed by staff within the assessment division. The divisional review for NAEP is dissimilar to the standard review process used by other divisions in NCES. Second is the center review, which includes Seastrom's review and she, along with the assistance of NCES or externally contracted NESSI staff, reviews the document based on predetermined criteria. The criteria for this review are contained within a 20-page document used by NCES and NESSI to ensure that a report meets NCES standards. Versions of this document exist for both the technical reviewers and research assistants (Seastrom has provided Buros with a copy of each of these documents). The NESSI staff members who are responsible for reviewing reports include research assistants and a mid-level analyst. For the urgent (six-month) reports, Seastrom strives to complete the center review process within one to two weeks. The comments from the center review are returned to the division and then shared with the author. The author is then given the opportunity to provide reactions to the comments. Seastrom is provided a summary of all comments sent to the author and the author's reactions to each comment. This iterative process continues as Seastrom or staff members provide comments to the authors' reactions. As noted above, the NAEP (assessment division) review process is different from the review process of other divisions. Whereas other divisions include an initial review by program staff (e.g., program officer), NAEP reports are immediately submitted to the division-wide review.

The process described above is also followed for the nonurgent reports (e.g., secondary analyses). In addition, after the center review, these reports are sent to IES who conducts both an internal and external review. All comments are consolidated and sent to the reviewer and Seastrom usually only reacts to the IES comments when there is a question concerning the interpretation of NCES statistical standards.

Review of reports by NAGB was also discussed. Seastrom noted that when there are conflicting policies between agencies (NAGB and NCES) and the difference could result in a violation of NCES standards, the commissioner is consulted. His decision is the final authority on report preparation (see Addendum A for clarification of review process among Seastrom, NCES, and NAGB).

Seastrom addressed the issue of the significant lag time in release of NAEP reports (other than the six-month reports). Her office has recently averaged a 21-day turnaround for the initial review and a 57-day total turnaround time to completion of the NCES center level review. However, the process by which these reports are passed between agencies often requires reviewers to refamiliarize themselves with reports as this iterative process often involves multiple drafts. In addition, reports that are of lower priority often seem to get "lost" in the process of author's revisions which can add significant lag time to the process. One specific problem noted was the NAEP technical reports. Because these are perceived as having a lower priority, these reports take the longest to produce. The next technical reports to be released (2000–01) will use a new online format but will also be available in paper format. Although a Web-based presentation of the technical manuals has been discussed, they have been shifted to a lower priority given other concerns in the testing program.

Materials Reviewed:

National Center for Education Statistics. (2002). NCES Statistical Standards.

Downloaded from: <http://nces.ed.gov/statprog/2002/stdtoc.asp>

NCES Internal Documents (provided by Seastrom):

- Reviewing NCES reports—Technical Reviewers
- Reviewing NCES reports—Research Assistants

Addendum A

After completing all site review reports, it was apparent that discrepancies existed between the documentation on the report review process. Although all agencies had reviewed each summary for factual accuracy, follow-up questions were submitted to clarify this discrepancy. The following summary represents the current understanding of the review process.

Review Process for NAEP reports

The eight steps below outline the process from initial creation to release.

- 1) Shell or Outline Review—This draft is presented to provide an overview of the report framework and intended comments. This step involves the report coordinator, assessment division staff and NAGB (staff and board).
- 2) Pre-division review—This initial draft of the report is presented with data and only that text which is not data dependent (e.g., description of a survey process). This step involves the report coordinator, assessment division staff and NAGB (staff and board).
- 3) Division-review—This is the draft report with full text. This step involves the report coordinator, assessment division staff and NAGB (staff and board).
- 4) Center-wide review—This is the complete report. This step involves the report coordinator, assessment division staff, NAGB (staff and board), and the NCES chief statistician who is assisted by NCES staff (outside the assessment division) and NESSI.
- 5) Commissioner—The report is sent to him/her for review after the chief statistician signs off.
- 6) IES Review—For six-month reports, the IES director will review the report; however, the non-6 month report involves a more extensive review. IES ultimately determines what type of review but the commissioner typically recommends if the review should be either internal, external, or both.
- 7) Short editorial review
- 8) Release

NAGB—At each of the first four steps, both NAGB staff and Board review the report. The staff and Board present NCES with a combined list of comments grouped into four categories:

- 1) Policy and guideline issues
- 2) Strong recommendations
- 3) Questions
- 4) Other editorial comments

According to NCES, for the initial report cards, NAGB comments are integrated as they fit with the NCES statistical standards. For the other reports they are considered to be more advisory comments.

NCES indicated that each phase involves only one round of review including the author response and approval of changes. Seastrom indicated that the center-wide review can sometimes involve multiple iterations - the author is allowed to provide feedback about comments and then whoever provided the comments can react.

Sources:

E-mail correspondence with Andrew Malizio (NCES) and Charles Smith (NAGB).

Site visit summary reports from NCES, NAGB, and Marilyn Seastrom (chief statistician with NCES).

This page intentionally left blank

Appendix G-4: Educational Testing Service (ETS)

Site Visit Team: Barbara Plake, Buros Center for Testing; Ronald Hambleton, Stephen Sireci, University of Massachusetts Amherst
Date of Site Visit: Oct. 27–28, 2005

Audit Summary

Staff

Jay Campbell—Project Director and Alliance Coordinator
Gloria Dion—Senior Program Administrator
Amy Drescher—Research team member
Robert Finnegan—Manager, NAEP Web reporting activities
David Freund—Director of NAEP data analysis
Lydia Gladkova—Member, Research Team
Eugene Gonzalez—Project Director, Field Services and Quality Control
Jeff Haberstroh—Project Director, Test development [Jeff Haberstroh did not actually attend the meeting, though he participated in preparations.]
Edward Kulick—Data Analyst
Michael Lapp—Project Director, Alliance coordination
Stephen Lazer—Vice President, Assessment Development
John Mazzeo—Associate Vice President of Research, head of statistical analysis and psychometrics research
Nancy Mead—Project Director, Reporting
Hilary Persky—Associate Director of Center for Technology
Mary Pitoniak—Lead Program Administrator, Research and Development

In advance of the visit, Buros shared with Arnold Goldstein, COR for ETS, information regarding the purpose of the audit, the comprehensive plan for the audit, and primary audit dimensions relevant to ETS. Using this information, Jay Campbell and Eugene Gonzalez coordinated the preparations by ETS for the site visit. They communicated directly with Barbara Plake. Gonzalez prepared an agenda prior to the visit that was shared with Plake, Hambleton, and Sireci in advance of the meeting. In addition, Buros was provided a number of documents prior to the site visit for review and additional documents were provided during the site visit and after the site visit.

Following introductions and a brief overview of ETS's contract with NCES and a brief summary of the NAEP audit goals, presentations were made by ETS staff members. These presentations were organized around the dimensions of the matrix that were identified as relevant to ETS. These interactions served as the primary information gathering process during the site visit.

Organizational Characteristics

With regard to qualifications of key staff members for their functions on the ETS NAEP contract, brief biographical statements were provided to Plake following the site visit. Senior NAEP staff members include Jay Campbell, project director and Alliance

coordinator, Catherine McClellan, director of NAEP psychometrics, David Freund, director of NAEP data analysis, and NAEP Program Directors: Nancy Mead (Reporting), Eugene Gonzalez (Field Service and Quality Control), Michael Lapp (Alliance Coordination) and Jeff Haberstroh (Test Development). Campbell has been involved with NAEP since 1990, working with NAEP Test Development and serving as the language arts coordinator. He has been the NAEP project director since 2004. McClellan is responsible for overseeing all operational assessment procedures for NAEP. She has been affiliated with data analyses for NAEP assessments since 1999. David Freund has been at ETS since 1980, joining the data analysis staff for NAEP in 1984. He has experience with data analysis and management of complex databases, including the National Longitudinal Study. Mead has a doctoral degree in speech communication (1977) and has been involved with the NAEP project in a variety of capacities starting in 1984. Lapp earned his Ph.D. degree in U. S. history in 1990. In 2000 he became the U.S. history coordinator for the NAEP project. Currently he serves as the director for Alliance coordination. Jeff Haberstroh has worked on several large-scale assessments projects at ETS, including serving as the mathematics development coordinator for the 1992, 1996, and 2000 NAEP assessments. Eugene Gonzalez is the director of NAEP Field Services and Quality Assurance. This team is supported by a number of other NAEP staff members and affiliated professionals within ETS, including Stephen Lazer, John Mazzeo, Henry Braun, Mary Pitoniak, Amy Drescher, Edward Kulick, and Andreas Oranje. The staff resources devoted to the NAEP project are impressive both in their technical quality and expertise, but also in the long time commitment to the NAEP assessment program.

ETS has the responsibility for coordinating the NAEP Alliance which consists of the following contractors for the NAEP project: ETS, Westat, AIR, PEM, and GMRI. A separate presentation was made at the site visit directed at the Alliance coordination responsibility. Prior to the most recent contract procurement model, ETS was the prime contractor for the NAEP assessment but worked with other principal contractors (except Westat) through subcontracts. Under the current model, members of the Alliance have separate contracts with NCES; their work is coordinated through NCES (which oversees all the contractors directly), HumRRO who has a separate contract with NCES to ensure quality across the contractors, and ETS who has a separate contract with NCES for Alliance coordination. ETS sees its Alliance coordinator role as one of “air traffic controller,” ensuring that the project stays on the “critical path” toward fulfilling overall NAEP outcomes and expectations (especially the six-month reporting timeline for reading and mathematics assessment results). In addition its role is as a conduit to ensure that potential problems are brought to the attention of NCES and to focus the alliance on quality control improvements (which overlaps with the roles and responsibilities of HumRRO). ETS accomplishes its Alliance coordination responsibilities through a variety of communication strategies, including regular meetings with contractors, holding an annual NAEP Design Summit, conducting regular conference calls with Alliance partners and NCES, and the use of the Integrated Management System (IMS) that allows for easy sharing of documents between contractors. The IMS also has varying levels of accessibility depending on the sensitivity of the material that is posted; it permits posting of logs of problems with documentation of resolutions. ETS has found that serving as the Alliance coordinator has its challenges because it has no real authority over the contractors but are held accountable for ensuring compliance across the Alliance partners for ensuring NAEP goals are achieved. Strategies used to coordinate functioning of the Alliance have been dynamic over the years of the contract, with changes made in

response to experience with communication procedures and recommendations by Alliance members.

One area of possible tension with communications appears to come from policy decisions, sometimes creating problems with timelines and procedures. For example, recent decisions by the Disclosure Review Board within IES regarding protection of student records required the data analysis division to respond to new policies regarding data perturbation. David Freund along with other ETS staff members worked directly with statisticians at IES and the Design and Analysis Committee (DAC) to come up with an acceptable strategy for working with the 2005 student data records. ETS was able to complete this process and implement the new procedures and still make the six-month reporting window dictated by the contract.

One feature on the contracts with the Alliance contractors is the bonus that is connected to meeting critical time points in their contracts. Because all Alliance members want to qualify for this bonus, and in order to meet these critical deadlines all members of the Alliance need to work cooperatively, this creates a spirit of “all for one and one for all” across contractors.

Intended scope and uses of NAEP assessments

Many of the research projects conducted by the NAEP research division are directed at improving connections between validation efforts and intended uses of results. Validity studies are included in the NAEP program of research. Mary Pitoniak described a long list of research studies aimed at enhancing the validity of item development, test administration, test scoring, data analysis, and score reporting.

The design of the reports and the messaging from Hager Sharp were noted as ways that ETS works to improve the appropriateness of interpretations of score reports. Although ETS prepares technical reports following each NAEP administration, the most recent of these reports (e.g., 2000 and beyond) are not available for review since they are still undergoing review at NCES. The lack of availability of recent technical manuals interfered with our ability to learn about many of the key features of the NAEP assessment program, particularly related to technical quality that would support intended uses of NAEP results. This delay does not appear to be due to ETS as it is required to submit its technical documentation per contract timelines; instead the delay is caused by NCES’s lengthy review process.

Develop assessment framework and test and background specifications

Although originally listed as a responsibility for ETS in the responsibilities matrix, upon discussion with ETS, it was decided that ETS’ involvement in this dimension was indirect and should not be listed as a responsibility. ETS does have an opportunity to serve in an ex officio role in the framework design committees and can (and does) provide feedback on preliminary framework design through NCES. However, this is a very minor level of input and it was decided by the Audit Team that this should not continue to be listed as a responsibility for ETS. ETS is responsible for ensuring that the items it develops for the assessment align with these frameworks.

Develop items and background questions

ETS has two major roles in the development of items and background questions. First, it has major responsibility for the development of items for the *NCLB* content areas (Reading, Mathematics, and Science currently). Second, it is responsible for final sign off on all items that eventually appear in an operational form of the assessments, regardless of content area. Further, it is responsible for the preparation of translated versions of the assessments (Spanish for Mathematics and Science). ETS uses mostly in house item writers for Reading but has a fairly substantial pool of external items writers for Mathematics. It uses external item writers for some other content areas. Most of the item development work is still in paper form, although it does receive item development files electronically from AIR. There are a number of possibilities for item review, at various stages in the item development process. ETS compiles all the comments from item reviewers.

Although AIR has the contract for training item writers for its item development efforts, ETS provides orientation to the history of NAEP and training about NAEP item formats for AIR training activities. ETS assumes responsibility for all items that appear in operational assessments and therefore uses its own item review processes for the items that are developed by AIR. All items that are selected for use operationally must also be reviewed by NAGB. NCES posts these comments and ETS provides comments and reactions to these recommended changes from the NAGB review. However, final decision about the items is the responsibility of NCES and NAGB, not ETS. Because ETS has responsibility for all NAEP items (whether it had the primary role in their development or not), both for content and background questions, a “*” has been added to this activity in the responsibility matrix (see page 1-121).

Create draft assessment, prepare field design and conduct field trials

The block design uses common items to link results across years and for reporting of trend results. Booklets are configured using a modification of a balanced incomplete block design to ensure that all blocks are paired and that all blocks appear in all positions in the assessment. This is a critical issue for the reporting of trend as the current block design reduces the size and certain types of errors that can undermine the linking of assessments across years. Also included in the assessment design are special studies or other booklet components that will affect the total number of assessment formats that are administered. ETS uses proprietary software that calculates the needed booklet formats to accommodate these assembly issues.

To improve the quality of pretest data for *NCLB* content areas, ETS has adopted a practice of pilot blocks. These pilot blocks are constructed to be responsive to several test development issues, such as breadth of content coverage, range of item difficulty, and position effects. These pilot blocks are used in operational settings following pilot testing and keep together as a unit in operational administrations. This has allowed for more confidence to be placed in the item statistics that results from the pilot administrations and has allowed for more efficient use of starting values for operational calibrations and scoring.

Construct final assessment and field design

Although ETS does not *conduct* the field trials (this is Westat’s responsibility), nor does it prepare the physical test booklets (this is PEM’s responsibility), it does provide to PEM the booklet and spiral “scripts” that are used by PEM for booklet printing and bundling and ETS provides to Westat information on sample size needs to provide for good estimates for use in scoring. ETS also reviews print documents for accuracy and technical quality.

Sample schools and students

Again, ETS does not actually do the sampling of schools and students (that is Westat’s responsibility), but it does provide counts to Westat for fulfillment in its sampling for administration.

Score the assessment and prepare final analysis database

ETS shares responsibility for scoring the constructed response items with PEM; ETS has this responsibility for the *NCLB* content areas, even when these items are not yet operational. In the alliance arrangement, PEM is an independent contractor, whereas in the past PEM was a subcontractor with ETS for NAEP scoring. Although ETS doesn’t have direct responsibility for some of the scoring practices, it maintains responsibility for the validity and reliability of the scoring as they impact the quality of the data that is used for scoring. Therefore, ETS serves in an oversight capacity in the monitoring of scoring that is done by PEM. This is a mutually beneficial relationship and is viewed as cooperative and not adversarial. Again, the “one for all” perspective, enhanced by financial rewards tied to meeting critical deadlines, was highlighted as a mechanism for the cooperative spirit that is enjoyed across the alliance partners.

Two areas were identified for possible revisions in the scoring procedures used for constructed response items. A limited number of papers are used for rescoring, and the results of the second reader are used only for computation of scorer reliability (percent of exact and adjacent agreements). To date, there was no operational use of the second scorer’s values, even when they were different (although a rater who is consistently found to be “off scale” may be singled out for retraining). It was recommended that when the two raters results are adjacent, some random process be used to assign the score for the performance; if the raters are more different than adjacent, the results could be averaged. Another area for consideration was the use of trend or validity papers in the operational scoring to mask better the appearance of these trend papers (see PEM site visit report).

Create scales and links and analyze data

ETS has primary responsibility for this dimension and the responsibility matrix has been updated to indicate this by adding a “*” to this cell in the responsibility matrix (see page 1-121). The scoring of the assessment is quite complex, involving several critical steps. Because of this complexity, several data quality checks are included throughout the process to ensure that the data that are analyzed are accurate and appropriate. A principal components analysis is conducted on the background questions to reduce the number of variables used in subsequent analyses (involving conditioning) to

those principal components that summarize at least 90 percent of the variance represented in the full set of background questions. This is done both at the national level and then separately for each state for state-by-state reporting. Because states differ in size and policy, the number of principal components used for the state-by-state analyses vary substantially, from as small as 100 to as many as over 400. No analyses are done to identify whether there is a common set of background variables across the states. Other strategies could be used to ensure some commonality in the principal components information that are used for state-by-state reporting, such as forced entry of some of the contrasts used in the principal component analyses conducted for the states. Following the creation of these principal components, plausible values methodology is used for the final scoring. This methodology is both very complex and controversial. It would be helpful if a more “user friendly” (e.g., simpler) explanation of this process could be prepared and shared with both the psychometric and lay communities. Presentations or training workshops at professional meetings about the plausible values methodology would likely be welcomed and well attended. Common items are used in the assessment for linking purposes in order to keep the results on a common scale.

As indicated earlier, data perturbation was added to the data preparation step for the 2005 reporting. This step was added in order to be in compliance with the federal regulations for protection of the privacy of student records. ETS developed a “data swapping” strategy that was accepted by the DRB for use with NAEP data.

ETS staff also reviewed procedures for evaluating items for differential item functioning (DIF). The evaluators asked for details regarding how items flagged for DIF were handled. ETS staff described the post-DIF item review process and distributed examples of the comprehensive information provided to the DIF item review committees. They also described specific instances where items flagged for DIF were removed from NAEP assessments.

Write, review, issue, and disseminate reports and data

ETS is working with two additional contractors on the issuing of NAEP results: Hager Sharp and GMRI. Although these two contractors have a key role in the dissemination of NAEP reports, ETS has the responsibility for creating the documents for release. It is constrained by new government reporting rules in the creation of these documents (that detail such things as a prohibition of footnotes and the total number of pages allowed). Due to the new interactive Web site that allows users to interact with NAEP results in ways that are meaningful to them, ETS has reduced its emphasis on paper/print reports. ETS also works with Westat in providing information about interpretation of NAEP results to the NAEP State Coordinators.

In the past, communication about NAEP results was under the auspices of NCES; NAGB has taken on this responsibility starting in 2004. NAGB is also seeking advice on ways to improve the messaging about NAEP results, hiring its own public relations consulting firm. These changing policies about who has primary responsibility for NAEP reporting have created some confusion both within the Alliance and between Alliance members, NCES and NAGB. Further, NCES is the main point of contact with users and ETS may not get access to questions raised about interpretation of results.

Even though NCES is seeking input from consulting firms, including Hager Sharp and GMRI, ETS must prepare the text for these firms to use in their preparation of support documents. Questions were raised by the audit team about the research

underlying NAEP reporting policies and procedures. The team urged conducting usability studies and focus groups to learn information from various user groups about how the information is being interpreted and used. Some research is underway by other Alliance members on report use (e.g., AIR's State Profiles study). More information about the usability of NAEP results is the focus of a separate study being conducted under the NAEP Evaluation Project.

Renew and improve the assessment

Funds were included in the NAEP contract for a “dedicated” research program within ETS focused on NAEP. The NAEP research program directly relates to ways of improving and renewing the assessment. This is accomplished through two different types of research, one directed at solving or resolving operational procedures and processes and another one that takes a longer view of assessment improvements. Funding differs across these two types of research. Immediate and short-term projects, with smaller price tags, can be approved without full NCES involvement. Larger ticket research projects require endorsement by NCES and therefore must go through a much more thorough review with the Department of Education's contracts office. These reviews take more time than do the reviews for smaller projects. Research projects emerge from operational staff members as well as from the DAC. Projects span different operational activities, such as the Item Attribute Study from test development, studies on ways to improve cross grade scaling, a long term bridge study, and an analysis of the impact of changes implemented in the 2003 NAEP design. Special studies have included an Oral Reading Study and two studies considering the use of online assessments (Math Online and Writing Online). There is a very strong program of research underway at ETS focused on NAEP.

Examination security

Security appears to be a serious consideration by ETS in the completion of its responsibilities for item development, data analysis, and reporting. All external item writers and reviewers are required to sign nondisclosure statements (as were the members of the site visit team). Backups of the databases, which reside only on the mainframe, are done daily. Business resumption plans are in place.

Findings and Recommendations

Consistent with its long history on the NAEP program, ETS continues its strong contribution in creating the nation's national assessment. ETS serves a fundamentally important role as the Alliance Coordinator. Without its long history with NAEP, it would be very challenging to coordinate all of the parts that make up the NAEP whole. Though we are tremendously impressed with the staff and the technical and logistical procedures in place for implementing NAEP, we do have a small number of recommendations that might improve procedures and practices:

- 1) Changes should be considered in ways that the second scorer's results are used for scoring. It seems that more effective uses can be found.

- 2) The use of trend papers in scoring should be modified to make their presence less obvious to the scorers. This is a potentially important recommendation because it is related to the validity of scoring trend papers and linking of the assessments from year to year.
- 3) Consider using a minimum number of fixed contrasts for the state-by-state principal component analyses of the background questions. Driving this recommendation is our concern that by not standardizing the components across states, a systematic bias in state results may be introduced. We encourage research be carried out to investigate our concern.
- 4) Consider preparing a user-friendly, more simplified presentation of the plausible values methodology for NAEP scoring and making presentations to the psychometric community on this methodology. ETS demonstrated to us that the validity of subgroup comparisons can be substantially aided by plausible values methodology. A better description of the methodology, and more demonstrations of the advantages and proof that the disadvantages are minor would be an invaluable contribution to the measurement field.
- 5) Make decisions about reporting based on a program of research involving usability studies. The trend today is to build score reports based on results from focus groups, experiments, cognitive labs, etc. We encourage ETS to continue that trend with more substantial research on what is one of the least studied aspects of NAEP and one of the most important for the success of NAEP.

Materials Reviewed:

Documents Available for Review Prior to Site Visit

Clement, J.—Inclusion Research Group

Greenberg, E.—Cognitive Labs to Evaluate NAEP Instructions

Mead, N.—Strategies for Reducing Exclusion Rates in NAEP

Braun, H.—SWD and LEP Inclusion/Exclusion in NAEP: Research Design and Instrument Development Study

NCES—NAEP 2002–03 Technical Documentation – Assessment Procedures (Draft). Retrieved 9/9/05 from: <http://cmspreview.naepims.org/nationsreportcard/tdw>

Documents Provided During/Following Site Visit

Biographical descriptions for key NAEP staff

Overview: ETS Fairness Review

Case Study: Development of 2009 NAEP Mathematics Operational Blocks

TCS Workfolder Control Sheet

2005 NAEP Audit—Key Data Analysis Steps

2005 NAEP Grade 8 Math: Breakdown of Assessment Sample Sizes

Description/Membership of the Design and Analysis Committee (DAC) for the National Assessment of Educational Progress (NAEP)

The Nation’s Report Card: Mathematics 2005

The Nation’s Report Card: Reading 2005

NAEP Reading Frameworks

2009 NAEP Reading Framework (Prepublication edition)

Assessment and Exercise Specifications for the 2005 NAEP Reading Framework (Prepublication edition)

Passage search guidelines for 2009 NAEP Reading Framework

2005 NAEP Reading Framework

NAEP Mathematics Frameworks

2005 NAEP Mathematics Framework

2005 NAEP Mathematics Assessment and Item Specifications

NAEP Standing Committees

Description of criteria for NAEP Standing Committee members

Guidelines for test development distributed to all new NAEP Standing Committee members

Information for Outside Item Writers

Sample instructions for outside-item writers with guidelines for writing multiple-choice items (Geography)

Sample materials used for outside item-writer workshop (Mathematics)

Spanish Translation and Adaptation

Materials pertaining to the development process for Spanish translation and adaptation

Committee responsibilities for Spanish translations and adaptation

Report on the NAEP 2003 assessment in Puerto Rico

Sample Documents and Checklists for Development Activities

Description of criteria for item review focusing on linguistic features

Sample of checklist used to review “mock-up” test booklets prior to printing (Mathematics)

Sample form for recording student responses during the question tryout activity (Reading)

NAEP Web site Materials
Web Trends Report
Report of 2005 Mathematics and Reading press release
AIR report on state profiles tool
Prototype of advanced item map for 2006

Report on the Math Online and Writing Online technology-based assessment projects

Appendix G-5: American Institutes for Research (Washington, D.C.)

Site Visit Team: Barbara Plake and James Impara, Buros Center for Testing

Date of Site Visit: Aug. 14, 2005

Audit Summary

Staff

Barry Levine – Managing Research Scientist

Sigrid Gustafson – Principal Research Scientist

George Bohnstedt – Chair of NAEP Validity Studies Panel

Larry Albright – Principal Computer Scientist

In advance of the visit, Buros shared with Marilyn Binkley, COR for AIR, information regarding the purpose of the audit, the comprehensive plan for the audit, and primary audit dimensions relevant to AIR. Using this information, Barry Levine coordinated the preparations by AIR for the site visit. He communicated directly with Barbara Plake. Levine was given a preliminary agenda prior to the visit.

In addition, Buros was provided a number of documents prior to the site visit for review. A list of these documents is attached. Several documents were provided during and following the site visit and are noted on the attached list of documents.

Following introductions and a brief overview of AIR's contract with NCES and a brief summary of the NAEP audit goals, interviews with AIR staff were conducted. These interviews were organized around the six dimensions of the matrix that were identified as relevant to AIR. These interactions served as the primary information gathering process during the site visit. The six audit dimensions identified for AIR are organizational characteristics of the NAEP assessment program, intended uses of NAEP assessments, development of test items and background questions, creation of scales and links and analysis of data, improvement of NAEP assessments, and examination security. Evidence, findings, and recommendations pertinent to each of these dimensions are summarized below.

Organization Characteristics

With regard to qualifications of key staff members for their functions on the AIR NAEP contract, we were able to interview some of these staff members and learn directly about their credentials. However, we were only able to interact with a select group of AIR staff members. An organizational chart showing the personnel structure for the AIR's NAEP project was provided to the site visit team following the visit. Further, staff qualifications were provided for key staff members. The project director, Barry Levine, has several years of experience in project management but no formal education in educational measurement or testing. Project leaders for Quality Assurance and Project Administration (QAT) are Barry Levine and Kristin Leahy. Leahy's responsibilities entail attendance at the QAT meetings as the AIR representative. The cognitive item development team for writing is led by Miriam Fuhrman, who holds a Ph.D. in earth and space sciences and June Zack, who coleads item writing teams throughout AIR, not just for the NAEP project, providing continuity in item development across projects within AIR. Members of this team include Nontas Konstantakos, who has experience as an item writer at ETS. The team developing the background questions is led by Sigrid Gustafson,

who holds a Ph.D. in organizational and industrial psychology and has worked on several projects at AIR prior to joining the NAEP effort. Steven Ferrara is the leader of the Special Projects division. Ferrara has a Ph.D. in educational measurement and many years of experience in large scale testing. He is well-qualified to lead this effort. Jon Cohen has lead responsibilities for software for this project. Cohen is AIR's vice president and director of assessment services. He holds a Ph.D. in methodology and American politics. He has served AIR in a number of roles, including director of the Computing and Statistical Sciences Center. He also served as study director on a NAEP evaluation conducted by NESSI of the statistical methods used in NAEP. Another team has responsibilities for scoring support. These AIR staff members work directly with ETS during scoring. Although the team members appear to be light on psychometric expertise, their collective credential and experience indicate they are qualified for the tasks related to their job positions.

Based on preliminary materials and information gleaned from other site visits, several targeted areas with regard to communications were identified for more detailed inquiry. In particular, we asked questions about the communications with NCES and with other Alliance partners, especially ETS since item development efforts are undertaken by both AIR and ETS. There was some initial concern by the audit team that the two contractors, who already have established systems for item development, would find the transfer of items, data, and files problematic. In addition, there was a concern that the two companies might view their work as somewhat competitive and at worst incompatible. One instance was reported in which a file format from ETS created some compatibility issues with AIR systems, but this was handled in a professional manner. It was reported that AIR's position was that AIR would adapt its systems to be responsive to the file formats provided by ETS instead of requesting ETS adapt its systems to fit AIR. There seems to be an environment of cooperation and task dedication that permeates the relationship between these contractors on the NAEP project. Therefore, the concern about competitiveness and communications problems were alleviated by our discussions with AIR but will be posed again when we conduct the ETS site visit in October.

Communications systems that have been put into place through HumRRO's QAC and QAT seem to be meeting the needs of AIR. The IMS system appears to function adequately to manage the sharing of information across the Alliance members. As an example of the shared communication and responsibilities across the Alliance contractors, Levine described an operations meeting that was held with the responsible contractors (principally AIR, ETS, Westat, PEM, and HumRRO) to discuss and negotiate fixed timelines for materials sharing and handoff for the critical path to getting NAEP assessments ready for the 2006 administration. An atmosphere of respect and dedication to the important and complex set of tasks that underlie NAEP was indicated for these Alliance contractors. Timelines appear to be clear and respected across the Alliance partners.

However, as is expected in a long-standing project, some resistance to change seems to be present, especially with long-standing contractors. This was seen as both an advantage and disadvantage. On the one hand, new "players" in the project (in this case AIR) have the advantage of looking at the extant procedures with an eye toward new systems and innovations. On the other hand, contractors such as AIR who have less long-term experience with such a complex system may not fully recognize the interconnections that might be vulnerable with the implementation of new or changed procedures. Leadership at AIR seems to appreciate this delicate balance and respect the

legacy of the system while time keeping an eye toward change and innovation. This seems like a healthy position for AIR and the other Alliance partners.

As other examples of the cooperation between Alliance partners, one staff member summarized the impact of an OMB decision regarding reporting categories for race or ethnicity and the problem with allowing the surrogate SES variable (free or reduced-price lunch status) be considered a school instead of a student-specific variable. Both of these changes could have serious ramifications across Alliance partner roles. For example, information about SES is currently used by ETS in the conditioning process for plausible values scoring. Sampling plans designed by Westat could be affected as could the ability to maintain reporting systems. NAGB asked for advice that resulted in a meeting of major contractors from the NAEP Alliance. Special studies were designed and through a recent NAGB decision, study designs are being further developed. These examples again illustrate the cooperative environment in which all the Alliance members, including AIR, contributes expertise. In the case of AIR, special questions were designed to address the SES variable through the background questionnaire. Pilot studies may allow for informed decisions about the utility of these questions to provide the needed information for conditioning, scoring, sampling, and reporting.

One area of possible tension with communications appears to come from policy decisions made by NAGB sometimes creating problems with timelines and procedures. For example, with the Arts assessments, delays in making decisions about new item development and the possible inclusion of performance tasks created some pressures within AIR's item development efforts. Further, NAGB's decisions had implications for the configuration of blocks for assessment design and administration, which impacted other Alliance contractors. Although it is clear that there are communications between NCES and NABG staff members regarding implications of NABG policy decisions, and instances when NAGB has sought advice from NCES about pending NAGB policy decisions, these policy decision nonetheless seem to put stress on Alliance partners in their ability to comply with their expected roles and functions.

Develop items and background questions

The major contribution by AIR to the NAEP project is in the development of cognitive items for all scheduled assessments except for Reading and Mathematics (which reside at ETS). AIR brought some new expertise and procedures to the long-standing item development procedures that were used historically by ETS (which had the only item development contract prior to the new Alliance procurement model). AIR directed efforts to improve the evidence of alignment of extant and newly developed cognitive test questions to the respective frameworks. Their efforts to examine item characteristics that provide better differentiated scales has been translated into item development training procedures. AIR is in the process of bringing items from ETS's database into AIR's Item Tracking System (ITS). The ITS has features that enable password- and privilege-dependent access to item writing, item review with comment tracking, item status checking, item statistics database generation, and eventual simulated test creation procedures to monitor compliance with test specifications. The ITS is seen as a strong strategic advantage for AIR in that it can be configured dynamically to fit various contract specifications.

AIR has taken a strong proactive role in the articulation of a model for the background questions, called the Contextual Variable Inference Map (C-VIM). The model allows for a systematic and strategic use of background questions to address

important questions related to the influences of certain school, teacher, and student variables on student achievement. In addition, the ITS mentioned previously also has the capacity to include the background questions and this application is currently being finalized.

There appears to be an issue regarding the transfer to ETS of NAEP items that have completed the full developmental and review cycle at AIR. ETS, whose editors and item developers may decide to make additional changes to the items after NAGB review, does not always articulate these changes to AIR so the ITS can be brought up to date with changes subsequent to the handoff. However, it isn't clear at this point who is the "responsible party" for the final survival or quality status of the items. The test questions must be positively reviewed by NAGB before they are deemed acceptable for use in a NAEP assessment. Further, once the items are used in the field, either in a pilot, field, or operational administration, item statistics are computed to document the technical quality of the items. Some quality indicators of AIR's item development efforts may be distorted if these AIR finalized items receive additional edits and revisions from ETS staff (which may or may not have been deemed acceptable by AIR test developers as they are not consulted following ETS's editorial decisions). Another relevant issue regarding the status of cognitive questions after they have been handed off to ETS by AIR is the final stage of review by NAGB. There appears to be some level of frustration, not just with AIR, that NAGB's standing committees, which have ultimate survival decisions about these test questions, may not be as capable with regard to good item writing practices as might be desired. The possibility of providing some orientation or item writing and test quality information to these reviewers might enhance to information base used by these reviewers in making final item selection decisions.

Security appears to be a serious consideration by AIR in the completion of its responsibilities for item development. The ITS has significant security features that protect the security and integrity of the test questions. All item writers and reviewers are required to sign a nondisclosure statement that identifies serious legal (felony) penalties for revealing NAEP items or information. System backups are in place; multiple locations, and thoroughly documented procedures also help to ensure examination security.

Create scales and links and analyze data

AIR serves only a minor role at this time in this dimension. Some of the special studies conducted by AIR have looked at the replication of the full parameter estimates used in the IRT scaling and the potential for other indicators for SES to be used in conditioning for scoring. It is not clear whether it would be advantageous, or even appropriate, for AIR to assume any greater involvement in this dimension.

Renew and improve the assessment

Although not directly connected with AIR's Alliance contract for NAEP, AIR has an indirect role in the renewal and improvement of NAEP through AIR's separate contract with NCES to coordinate the Validity Studies Panel. This panel has the responsibility to attend to the future directions of NAEP through the articulation of a validity research framework. Funds for the actual implementation of these studies are limited and result in a "favored" position of university-based researchers either who serve

on the panel or who have connections with validity panel members. Although this panel is presented as being “independent”, its contract is directly through NCES, and as indicated in the NCES audit report, leadership of NCES makes the final decisions about which VSP studies to fund or support.

Other contractors, including AIR, have the opportunity to conduct special studies. These special studies also provide opportunities to examine means and mechanisms for assessment renewal and improvement. Examples of these special studies include an Accommodations Validity Study, a project to include teacher pedagogy questions in the set of teacher-specific questions on the Background Questionnaire, and the previously mentioned study on the utility of other background questions to measure the SES construct. The degree to which this special studies program could be more forward looking is a decision to be made in cooperation with NCES. However, the potential to conduct studies that could more directly inform assessment renewal and improvement is present through the opportunity to conduct these special studies. To date, there does not appear to be an articulated purpose or comprehensive plan for the special studies program.

Findings and Recommendations

Overall, AIR appears to be conducting excellent work in item development for the NAEP project. It appears that AIR is working cooperatively and effectively with the other Alliance partners in producing and supporting the NAEP activities. There are only a few areas in which recommendations might improve procedures and practices:

- 1) Conduct an analysis of the goals for special studies program could help inform the direction and decision about additional studies.
- 2) Continue communication with NABG through NCES about how policy decisions and delayed operational decisions affect the capability of Alliance partners to effectively and efficiently complete their responsibilities.
- 3) Establish an item tracking system that carries item modification and performance evaluations back into the ITS in order to keep fully documented records of item development history after it has been handed off to ETS for continued review, modification, and final evaluation by NABG.
- 4) Seek opportunities to provide guidance to NABG regarding the need for NABG standing committee members to have training in good item writing practice to assist them in making good decisions when making final review decisions about NAEP items.

Materials Reviewed:

Documents Available for Review Prior to Site Visit

NCES—Statement of Work for the NAEP Alliance

Finnigan, R.—Design and Implementation of Automated NAEP Cognitive Item Tracking System

Kelly, D., and Ferrara, S.—Developing a New Measure of SES

Neidorf, T.—NAEP, TIMMS and PISA Comparison Studies in Mathematics and Science

Neidorf, T.—NAEP Scoring Guide Studies

Gattis, K.—Pilot Eighth-Grade Mathematics Project Comparing National Assessment of Educational Progress (NAEP) and State Frameworks Assessment

Lapp, M. —Implications of Item Pool Expansion for NAEP Assessments.

Documents Provided During Site Visit

Ferrara, S., and Olmeda, R. (February, 2004). PowerPoint presentation on Proposed NAEP Accommodations Validity Studies (AVS). Summary presented for the NAEP DAC.

Ferrara, S., and Olmeda, R. (February, 2004), Studies of effects on test score validity of test administration accommodations. Excerpted version for the NAEP DAC meeting.

Gustafson, S., Fast, M., Fuhrman, M., and Merola, St. (March, 2004). A proposal for structuring existing background data to address strategic topics: The contextual variable inference map (C-VIM). Submitted to NCES.

NCES (May, 2005). 2006 Economics Assessment Background Questions.

NCES (May, 2005). 2006 U.S. History and Civics Assessment Operational Background Questions.

Appendix G-6: American Institutes for Research (Palo Alto, Calif.)

Site Visit Team: Chad Buckendahl, Buros Center for Testing, and Ed Wiley, University of Colorado

Date of Site Visit: June 29, 2005

Audit Summary

Staff

Victor Bandeira de Mello – Senior Statistician

Don McLaughlin –Member of NAEP Validity Studies Panel

Fran Stancavage – Project Director of NAEP Validity Studies Panel

Prior to the site visit, Buros shared with Janis Brown, one of NCES's CORs for projects led by AIR–Palo Alto, information regarding the purpose of the audit, comprehensive plan for the audit, and the primary audit dimensions relevant to AIR–Palo Alto. Using this information, Fran Stancavage coordinated the local logistics for the site visit. She communicated directly with Chad Buckendahl. Stancavage, McLaughlin, and Bandeira de Mello were given a preliminary agenda prior to the visit. Buros was provided documents in advance for review, some by NCES and some by AIR–Palo Alto. In addition, other documents were provided following the site visit.

Following introductions, a brief overview of AIR–Palo Alto's contracts with NCES, and a brief summary of the audit goals, interviews with AIR staff were conducted. These interviews were organized around the three dimensions of the matrix that were identified as relevant to AIR–Palo Alto. These interviews served as the primary information gathering strategy during the site visit. The three audit dimensions identified for AIR–Palo Alto are Organizational Characteristics, Intended Scope and Use of NAEP Assessments, and Renew and Improve the Assessment. Note that AIR–Palo Alto's contracts are not part of the NAEP Alliance; therefore, AIR does not have any day-to-day responsibilities for the operations of NAEP. Evidence, findings, and recommendations pertinent to each of these dimensions are summarized below.

Organizational Characteristics

Victor Bandeira de Mello, Don McLaughlin, and Fran Stancavage met with evaluation team representatives on June 29, 2005, to gather information about the Palo Alto office of AIR's role in NAEP. Each of these key staff members have been with AIR for a number of years and are well-qualified to lead AIR's efforts. The team will soon experience some change as McLaughlin indicated that he would be retiring from AIR at the end of July.

The Palo Alto office of AIR is not responsible for the operational assessments of NAEP. Its primary role in NAEP is to lead the state analysis projects and also to facilitate the NAEP Validity Studies Panel whose responsibility is to confirm and improve validity of NAEP.

We discussed the review process where technical reports or white papers were submitted to NCES. There appears to be a number of checkpoints in the process that may be perceived as delays that will require the evaluation team to examine other key steps in

the review process to learn more about what factors may be contributing to the turnaround. For example, AIR described the process for a particular report where a preliminary draft was submitted in July 2004. A draft report with analyses and data was then submitted in November 2004. By early February, the first set of comments from NCES was received (approximately 25 pages of comments with a majority focusing on stylistic edits rather than substantive edits); AIR responded to the comments and resubmitted the report in early March. The next round of comments was received in late June. Some of these comments contradicted previous comments, in part because different reviewers were involved in the subsequent round of review. Furthermore, reviewers appeared to have reviewed a report on the Reading assessment though the review was intended for a report on the Mathematics assessment. This round of reviews is within NCES at the assessment division level. At the point of the visit, the report still needed to go through the next level of NCES reviews before it is made public.

Other reports have experienced similar timelines in the review process. Some reports have apparently not made it through the divisional review process and have never been released. Some of this is due to prioritization of particular topics. Within the review process, the comments are often not consolidated so there may be contradictory comments. Moreover, the changes recommended by the reviewer may not be endorsed by the COR. Because of turnaround and feedback in the review process, project timelines must be adjusted to reflect when reviews are received. Although the longer review processes were historically within the assessment division and NCES, the additional reviews at the IES level appear to have lengthened the process. AIR staff sees the power of NAEP in its ability to provide information; as such, these potential delays in the review process threaten NAEP's power and are especially frustrating.

There is also a perception that the current political climate of *NCLB* has contributed to the length of the review process and potential technical disagreements in the comments. The lag time from delivery of a report to publication may also be mitigated by the content of the reports. For example, the data reported on charter schools may not support administration policies; therefore, it may not be approved or released as quickly.

Although direct contact is not prevalent between the Palo Alto office and other contractors because of their role in the operational aspects of NAEP, there appears to be some interaction with HumRRO and ETS regarding efforts to renew and improve NAEP. Efforts to interact with NAEP state coordinators as an extension of support to state data analyses have been curtailed by NCES. Some of this may be a result of Westat's contract to provide support to states. It appears that some of the research that AIR has conducted with respect to the state analyses may not be encouraged or made readily available to the state coordinators.

AIR is involved in a database development project with respect to state assessment data. It appears that some of these efforts were being led by the Policy and Programs Studies Service at ED. AIR's goal was to create this database to provide free access to state data. By 2000 AIR had already collected state information from 48 states and had a contract for the state data analyses. Reports on linking state databases to federal and other sources were mentioned during the interviews and Buros was able to retrieve and review these after the site visit.

Because NESSI is a division of AIR, there is a potential for conflicts of interest given some of NESSI's responsibilities. Specifically, if NESSI is performing contracted work for NCES with RFPs or reviewing reports, there should be a clear separation

between the NESSI staff members and AIR. The only clear, direct collaboration between NESSI and AIR's Palo Alto staff is Stancavage's involvement with NESSI's Bureau of Indian Affairs Indian Education study. Additional information was collected about NESSI's role within AIR and NCES in subsequent interviews.

Intended scope and uses of NAEP assessments

AIR's primary contribution to the validity framework of NAEP is through the NAEP Validity Studies (NVS) Panel. The NVS is an independent advisory group to NCES and is seen as an extension of the trial state assessment evaluation. It can also be viewed as similar to what the DAC does for ETS for the technical characteristics. The NVS is broadly representative of the NAEP research community and has a strong overlap with researchers who were part of NAEP's Trial State Assessment evaluation.

Because NVS is independent of NCES, the review of the study designs and final reports are conducted by panel members before AIR publishes them. However, the determination of which studies are ultimately conducted appears to be greatly influenced by the director of the assessment division of NCES, Peggy Carr. AIR staff members indicated encouragement by NCES to present research at professional conferences and publish in the professional literature.

The state analysis project is seen as a response to calls to use NAEP as a confirmatory tool for state assessment. There was also a suggestion discussed during the interview to discontinue using the phrase "gold standard" when referring to the NAEP assessment because it may be interpreted as suggesting that other assessment practices are invalid. This may communicate an incorrect message to the public and also be inconsistent with measurement theory.

Renew and improve the assessment

NVS disseminates technical reports outside the NCES review process. These are distributed through professional conferences (e.g., AERA, NCME, CCSSO) or published directly by AIR.

Because of the testing cycle, the operational system does not have a way to infuse research innovations into practice without disrupting the system. Some of this is probably because of shortened reporting requirements for particular subject areas because of *NCLB*, but another factor may be ETS being reluctant to implement procedures that it did not develop. NCES has a demonstrated range of technical interests; however, NAGB may not be as interested in modifying methodologies.

There does not appear to be a decision-making structure for reviewing or evaluating new ideas or a budget built into operational practice for planned change. Innovations are recommended through technical reports or research studies but may not be acted on. For example, AIR suggested a method for determining how to interpret state assessment achievement levels on the NAEP scale (McLaughlin et al., 2005). ETS suggested an alternative strategy for doing this, but there does not appear to be an evaluation process to determine whether these methods are appropriate. One staff member also noted that he has received criticism in the past for computing school level NAEP scores for analyses. The distinction may be between computing (as AIR has done in some of its analyses) and reporting (as is not allowed because of technical inadequacy) NAEP scores.

Related to this work, a staff member mentioned a report that has an NCES publication number (from 2001) but has not been placed on NCES's Web site. The paper provides guidelines for linking NAEP and state assessments. There appears to be greater interest in conducting these analyses and showing states that their achievement level decisions are varied. This information may help promote the utility of NAEP data at the state level.

Another example is the efforts to operationally implement full population estimates. Analyses in 1998 suggested that observed NAEP gains were due to the increasing rates of exclusions. This is a topic that was formally proposed in 2002 (McLaughlin, 2002), but not implemented [although these values may be included as an appendix in the 2003 report]. It was also noted that HumRRO conducted an evaluation of the methodology and was to compare AIR's method with an alternative method proposed by ETS. To date, there does not appear to have been an alternative method submitted by ETS to HumRRO for the comparative evaluation. A key element of this research is related to the SWD and ELL questionnaire, an idea that evolved from Trial State Assessment evaluation. One of the challenges of this research is this questionnaire that may be substantially shortened or eliminated.

It appears that budget limitations related to changes in the 12th-grade assessment have precluded NCES from sponsoring new work or research. This is likely an effect of NAGB's policy decisions and has the potential to damage the continued validity research necessary to support the program.

Additional areas of research on which AIR would focus include research on scoring and interpreting accommodated performance. This is an area that has not been researched in the broader educational community. Also, reporting gaps in performance as a percent may be distorting results.

Findings and Recommendations

In general, AIR–Palo Alto appears to be conducting meaningful work related particularly to State NAEP. Because of its role outside the Alliance, some of its work appears to be prioritized lower by NCES and NAGB which may contribute to some of the perceptions and frustrations about how its work is received, reviewed, and implemented. There are only a few areas where recommendations might improve procedures and practices:

- 1) Don McLaughlin's retirement may challenge AIR's staffing in this area because of the long-standing involvement he has had with NAEP. Victor Bandeira de Mello is very capable of taking over as the director of many of the ongoing state analysis projects; however, we encourage AIR to recognize the need for additional staffing assistance to address what appears to be an already full workload.
- 2) We were surprised to discover that the NAEP Validity Studies Panel was not the group that defined the validity framework for NAEP. We recommend that AIR–Palo Alto be involved in efforts by NCES to unify its validity framework for NAEP.
- 3) Related to the scope of the NVS, we recommend that there be a more formal process for identifying studies for NVS. It appears that the current strategy is variable based on priorities and available funding, but appears to be limited to the research interests of the panel's members.

Materials Reviewed:

Documents Available for Review Prior to the Site Visit

Stancavage, F., et al. *An Agenda for NAEP Validity Research.*

Linn. R., et al. *Assigning adaptive NAEP booklets based on state assessment scores: A simulation study of the impact of Standard Errors.*

McLaughlin, D., et al. *Comparison between NAEP and State Reading Assessment Results: 2003.*

McLaughlin, D., et al. *Comparison between NAEP and State Mathematics Assessment Results: 2003.*

McLaughlin, D. *Properties of NAEP Full Population Estimates.*

McLaughlin, D., et al. *Using state assessments to assign booklets to NAEP students to minimize measurement error: An empirical study in four states.*

McLaughlin, D., et al. *Using state assessments to impute achievement of students absent from NAEP: An empirical study in four states.*

Additional Materials Reviewed Following the Site Visit

Harr et al. (2004). *Enhanced Database on Inclusion and Accommodations Variables and Measures.*

Bandeira de Mello, V. and McLaughlin, D. (2004). *Linking the NAEP Database with other State or Federal Databases: School level correlations of achievement 2000, Revised Analysis Plan.*

Bandeira de Mello, V. (2004). *Linking the NAEP Database with other State and Federal Databases: List of databases and variables.*

Harr, et al. (2005). *Participation of and Accommodations for Students with Disabilities.*

Perez et al. (2005). *Participation of and Accommodations for English Language Learners.*

Bandeira de Mello, V. (2004). *State Profile and Report Enhancement: Recommendations on state web profiles.*

Bandeira de Mello, V. (2004). *State Profile and Report Enhancement: Recommendations on the state report generator.*

This page intentionally left blank

Appendix G-7: Government Micro Resources Inc. (GMRI)

Site Visit Team: Chad Buckendahl, Buros Center for Testing; April Zenisky Laguilles,
University of Massachusetts, Amherst
Date of Site Visit: Oct. 13, 2005

Audit Summary

Staff

Paul Harder – Director of federal and civilian programs
Lori Rokus – Project manager
Keith Lamond – TAIC, Quality Assurance Specialist

In advance of the visit, Buros shared with Rich Struense, COR for GMRI, information regarding the purpose of the audit, the comprehensive plan for the audit, and primary audit dimensions relevant to GMRI. Using this information, Paul Harder coordinated the preparations by GMRI for the site visit. He communicated directly with Buckendahl. Harder was given a preliminary agenda prior to the visit.

Buros had difficulty accessing documents for review prior to the site visit. Some of this difficulty was based on our staff's lack of knowledge of where to find information related to GMRI's role in NAEP on the IMS site. During the meeting, Harder provided some of the requested documents and demonstrated how to access the Internal Management System (IMS) system to review additional documentation.

Following introductions and a brief overview of GMRI's contract responsibilities to NCES and a brief summary of the audit goals, interviews with GMRI staff were conducted. These interviews were organized around the two dimensions of the matrix that were identified as relevant to GMRI. These interviews were the primary source of information for this preliminary summary. The two primary audit dimensions identified for GMRI were Organizational Characteristics and Write, review, issue, disseminate reports and data. Because of the technology infrastructure involved and the maintenance of Web sites that are connected to data, it also seemed appropriate to gather some information on recovery and security procedures. Also, because technology usage in NAEP is evolving, it was also appropriate to gather information on how the renewal and improvement processes applied to GMRI.

Organizational characteristics

GMRI's history with NAEP is more recent than other contractors in the Alliance; however, it plays a critical role by providing much of the technology infrastructure that facilitates contractor interaction and dissemination of information about NAEP. The organization has over 20 years of experience and is recognized as a Microsoft Gold Certified Partner for its work in Information Worker Solutions and Integrated e-Business Solutions. Staff members responsible for NAEP activities have extensive experience in software development and project management. Specifically, Paul Harder has almost 20 years of experience managing technical and functional solutions for clients and serves as the project director. As the senior project manager, Lori Rokus has approximately 20 years experience working with commercial and government clients on projects that are similar to those required in the NAEP Alliance contract. Harlan Messinger, development manager, and Alan Wu, senior developer and system architect, are also very well-qualified to respond to the needs of the contract. Because of additional

responsibilities related to training and usability that have evolved in the contract, Timothy Kilby's experience in developing and maintain effective e-learning systems is also important to GMRI's role in NAEP. Allyson Armistead and Elyse Csillag are located on-site in D.C. at NCES to serve in a Web-editing role for GMRI.

GMRI's relationship with NAEP began in 2002 when it was brought in as part of the NAEP Alliance. Its primary roles are to develop and manage the content of the Web site and to develop more effective ways for communication to occur among contractors in the Alliance through the Integrated Management System (IMS) and Web CMS. More recently, additional activities have been included within GMRI's scope. For example, the NAEP Network for the NAEP State Service Center is being transitioned to GMRI given its experience with similar systems. GMRI is also developing the Public Communication Tracking System (PCTS) as a means to respond to inquiries, offer Web reporting, and direct questions dynamically to the contractors that would be in the best position to respond.

GMRI currently hosts review sites for the NAEP public site (Technical Documentation Web Site), NAEP Data Explorer, State Report Generator, State Coordinators, NAEP Network, and NAEP Orientation in addition to the sites it hosts for production sites. The current production sites include the IMS, Web content management system, NAEP state service center Service Desk, public communication tracking system, NAEP Network, NAEP Item Bank, NAEP Incident Tracker, Outstanding Task List, and Training Evaluation.

As part of its quality assurance model, GMRI contracts with an external company, TAIC, that provides a check on the internal processes that in place. This is a commendable strategy. This added measure of quality control is built into GMRI's organizational structure and appears to be part of its general business model that has been used with other clients.

The internal communications among GMRI staff members include biweekly internal manager meetings and team meetings, weekly division senior management meetings, weekly status reports to corporate senior management, and quarterly meetings with corporate senior management.

With one of its roles defined as facilitating communication among contractors in the Alliance, this represents a critical component for GMRI. Staff members at GMRI indicated that they had a very good relationship with their COR, Richard Struense, and that communication with him and other NCES staff members was direct. They also indicated that Struense assisted them by serving as a buffer between them and external requests that may ask them to go beyond the scope of their responsibilities. For the Alliance, GMRI participates in weekly status meetings on Web development and the NAEP Network, monthly progress reports, biweekly conference calls with Alliance directors, monthly in-person meetings with Struense, as-needed meetings with senior NCES or Alliance members, quarterly Web coordination conferences, and quarterly NAEP program reviews. Sample agendas from some of these meetings were provided as evidence of these activities.

The review process and production of technical reports were also discussed. NCES is responsible for reviewing and approving pages on the Web site. This appears to require at least two steps. The first step is the NCES review (Assessment Division and chief statistician) followed by an ETS review as the second step. Because GMRI is at the end of the cycle in the review process, timelines often get pushed to the end limiting opportunities for quality control. Although GMRI is not responsible for constructing technical reports for the assessments, the transition from written reports to Web-based versions has been slowed by the review process. Its Web CMS system has been used more recently as a review mechanism in the past six months. This was viewed as a positive step in addressing some of the delays in the review process.

Write, review, issue, and disseminate reports and data

Because it serves as the technology infrastructure for NAEP, GMRI is not responsible for the content in the reports; therefore, it does not play a role in writing or reviewing (for content) the reports. Part of its role in disseminating the information involves verifying the Web site's capability to display and communicate the results NAEP assessments. GMRI has developed some general criteria that it uses to test the Web site prior to the release of information. The criteria are that functionality meets requirements, interface usability, browser compatibility (i.e., Internet Explorer 5.0+, Firefox), conformance to NCES style guidelines, adherence to Section 508a and W3C accessibility guidelines, adherence to W3C HTML 4.01 coding standards, existence of meta-tags, titles, and keywords, and content testing.

One of the recent challenges for which GMRI created a solution was with respect to the initial release Web site. Because NAGB has assumed leadership of this activity, during the development, testing, and refinement stages, GMRI has received input from NAGB directly and through feedback in a report from Ogilvy that was presented at the NAGB meeting in May 2005. Some NAGB members had been given access to preliminary drafts and have provided comments on the drafts. Because of the contractual relationship to NCES, GMRI has gone to its COR to verify authorization if NAGB has communicated comments or suggestions directly. GMRI also indicated that there were discussions about the potentially different style requirements early in the process. NAGB's comments were included in the revisions and then retroactively fit to the NCES style guidelines to ensure that both were considered. These revisions occurred without major issues.

With respect to the initial release site, GMRI provided comments to NAGB and NCES about the potential user frustration that might be experienced when the information on the initial release site was no longer available after a couple weeks. This recommendation was considered, but not implemented. GMRI also noted that although it may frustrate some users, there is not consensus within the IT industry about appropriate strategies.

The organization also has well-designed and documented quality control (QC) processes and products. Some of these processes are performed internally, whereas some are conducted externally by TAIC or by various stakeholders in the broader NAEP universe. There are a number of QC tools that allow GMRI staff to monitor quality in the range of its activities. For example, within Web sites that they develop and host, they have created a NAEP Incident Tracker (NIT), an integrated spell checker, and tools that check hyperlink linking. The NIT appears to be particularly useful because it serves as a mechanism for collecting customer satisfaction information. ETS also asked to use the NIT as an internal feedback tool. Other QC measures include 508 Compliance validation tools (e.g., accommodations for users), Visual SourceSafe, file comparison difference tools, and standard document templates for greater consistency. Through the IMS system, there are document sharing features built in that allow contractors to access and share information. Finally, because of the range of features that the Web sites offer, GMRI conducts evaluations of users who have participated in training activities to gather procedural validity evidence that it considers in revisions.

Although responsible for developing and maintaining the Web sites for NAEP, GMRI has limited control over gathering some of the Web usage information it may need to better inform design or structural decisions. Because the NAEP pages are housed within the NCES site, there may be some confounding of information that GMRI receives from Webtrends. From the larger dataset that it receives from NCES, GMRI has been able to generate information on monthly traffic flow in terms of page requests. Between August 2004 and September 2005, there were approximately 350,000 page requests per month. Notably there were approximately 500,000 page requests in October 2004 and January 2005. The start pages for some specific

NAEP tools including the NAEP Data Tool, NAEP Questions Tool, and State Profiles were also tracked between September 2004 and September 2005. Although typically the State Profiles tool has been used more, in two instances (i.e., January 2005 and September 2005) the NAEP Questions Tool surpassed it. Because GMRI does not have control over how these data are collected, there were some limitations in the interpretation of these data. Some of the NAEP pages did not receive enough hits to make it into the Webtrends report making it difficult to evaluate the utility of the pages.

GMRI is also able to monitor the IMS site usage and produces monthly reports that document the total number of contacts, new discussion entries, active subscriptions, modified documents, and public folder entries. The system allows users to be assigned to different roles and permission levels (e.g., readers, authors, coordinators). Because of the way visitors to the site are defined, the number of visitors appears to fluctuate with the NAEP assessment cycle; however, the number of unique IPs remains relatively consistent throughout the year with approximately 50 per month.

Within its role supporting technology infrastructure and disseminating data, GMRI also assists with the development of Web-based tools to be used by visitors to the NAEP Web site. Part of this role is collaborating with NCES and ETS on design and usability of these tools. GMRI carried out a usability study to identify navigation and other issues with a prototype version of the NAEP Data Explorer (referred to in the cited report as the NAEP Data Analyst). This study involved analysis of videotapes of a small sample of likely users of the tool to help identify sources of confusion in carrying out different tasks. The users involved varied with respect to their levels of familiarity with both NAEP and computer applications. Many ease-of-use dimensions of the NAEP Data Analyst received considerable attention in the course of the study, and in its report GMRI made substantive recommendations about ways to ease navigation and to facilitate use of the Analyst tool by implementing layout changes, standardizing the use of buttons or tabs for changing views and modes, and improvements in page loading speeds, among other areas. Use of the recently released NAEP Data Explorer (NDE) suggests that many of its recommendations were incorporated into the current version of the tool and serves as evidence of the contribution GMRI had made to support dissemination of NAEP results on the Web.

Because the Web sites it hosts contain information that is widely viewed and also contain information used in the development process, GMRI has an extensive recovery and security plan. There are daily back-up procedures with information housed off-site. There are also daily and weekly incremental recovery efforts in addition to full monthly back-ups. Although there have been no system failures, a parallel system has been built to automatically switch over in the unlikely event of a failure. Because the infrastructure is all housed internally, this parallel system is important. The recovery system, though, is critically important should something happen to GMRI's facilities.

From a security perspective, access to the various Web sites or directories requires authentication as well as password protection with assigned roles and permissions levels. There is continual virus scanning and logs documenting the results of these scans. Also noteworthy is that NAEP information is housed within GMRI's firewalls providing an additional layer of protection for the information contained on these sites.

Renew and improve the assessment

Technology innovations have played an increasing role in NAEP with the transition of many processes and products to electronic, particularly Web-based form. Many of the innovations for NAEP have been achieved through this avenue over the past four to five years.

GMRI's roles in this innovation have been quite evident through the IMS system, Web CMS, and the variety of Web sites that it has developed and currently hosts. GMRI is currently in the process of transitioning the public Web site for the NAEP Network and updating IMS to a new version (3.0) that will include additional features and functionality. Additional innovations that were beyond the scope of the contract, but were benefits to the systems included the PTCS and NIT. GMRI indicated it was encouraged to explore new technologies and have been asked by NCES to "harness the power of the Web" as it relates to NAEP. These activities continue to evolve.

It is also important to note that GMRI has been recognized within and outside of the Alliance for its work on these systems. GMRI noted that NCES and other contractors have provided positive feedback which is rare. Microsoft recognized the IMS system that GMRI developed for NAEP with an innovation award. This type of award adds to the credibility of organization within the IT community and beyond as an external indicator of quality.

Findings and Recommendations

GMRI has used its experience as an IT solutions provider to develop and maintain excellent internal systems for housing information for NAEP and facilitating contractors' communication within the Alliance. Its inclusion into the Alliance in 2002 provides NAEP with an opportunity to continue to explore some of the technologies that are available for maintaining an assessment program of this complexity. GMRI has expanded its role from the original intent following requests to develop more Web-based solutions to operational needs. We were very encouraged by the documentation of processes and products that GMRI provided to us onsite so that we could review them. Because of its unique role in the Alliance, there are only a few areas where recommendations for improved procedures and practices may be warranted:

- 1) We recommend that GMRI pursue strategies that will allow more opportunities to control the usability data. The expertise that the organization has in this area suggests that it be allowed to collect and analyze data that are relevant to the products and services it is providing.
- 2) We also recommend that GMRI consider exploring innovations related to the development of computer-based or Web-based NAEP assessments. Because NAGB and NCES have had preliminary discussions about this topic, it may be beneficial for GMRI to anticipate requests by NCES or contractors to assist or advise them in technology solutions.
- 3) Although current staffing appears to be sufficient, if an increasing number of Web-based solutions and activities are being requested it would likely warrant additional technical staff to help support the increased load.

Materials reviewed (provided after the site visit):

GMRI (October, 2005). Agenda, Web CC 10.05.05. Manassas, Va.: Author.

GMRI (September, 2005). IMS Site Usage Analysis. Manassas, Va.: Author.

GMRI (September, 2005). Monthly Status Report for September 2005. Manassas, Va.: Author.

GMRI (September, 2005). NAEP Site Usage for September, 2005: Monthly Traffic Summary. Manassas, Va.: Author.

GMRI (April, 2005). National Assessment of Educational Progress Semi-Annual Needs Analysis – Review Draft 1.0. Manassas, Va.: Author.

GMRI (October, 2004). National Assessment of Educational Progress (NAEP) Web Development Support: Project Management Plan (PMP). Manassas, Va.: Author.

GMRI (April, 2005). NAEP Data Analyst Usability Study. Manassas, Va.: Author.

Halstead, R. (July, 2004). Software Requirements Specification for NAEP Integrated Management System (IMS) Release 3.0. Manassas, Va.: GMRI.

Harder, P. (October, 2005). National Assessment of Educational Progress: Contractor Assessment (presentation slides). Manassas, Va.: GMRI.

Lazar, P. (November, 2002). Migration Plan. Manassas, Va.: GMRI.

Orban, M. and Halstead, R. (March, 2005). Software Requirements Specification for Public Communication Tracking System (PCTS) Version 2.0. Manassas, Va.: GMRI.

TAIC (date unknown). GMRI 2006 NAEP Quality Plan. Washington, D.C.: Author

Appendix G-8: Human Research Resources Organization (HumRRO)

Site Visit Team: Barbara Plake and James Impara, Buros Center for Testing

Date of Site Visit: June 30, 2005

Audit Summary

Staff

Laress Wise—President

Sunny Becker—Senior Staff, Center for Personal Policy Analysis

Felicia Butler—Research Associate in Instructional Development and Educational Assessment Program

Carolyn Harris—Program Manager, Site visit coordinator

Gene Hoffman - Program Manager of Center for Learning, Evaluation, and Assessment Research (CLEAR)

Paul Sticha—Program Manager of Modeling and Simulation Program

Jay Noell—Department of Education participant

Janis Brown—NCES participant

In advance of the visit, Buros shared with Janis Brown, COR for the Human Research Resources Organization (HumRRO), information regarding the purpose of the audit, the comprehensive plan for the audit, and primary audit dimensions relevant to HumRRO. Using this information, Sunny Becker coordinated the preparations by HumRRO for the site visit. She communicated directly with Barbara Plake, providing a preliminary agenda. Plake made suggestions for revisions and a final agenda was agreed upon. Further, Plake sent to Becker an elaboration on the process and on the specific audit dimensions relevant to HumRRO.

In addition, Buros was provided a number of documents prior to the site visit for review. A list of these documents is also attached. Several documents were provided following the site visit and are noted on the attached list of documents.

The agenda for the site visit is attached. Following introductions and a brief overview of HumRRO's contract with NCES and a brief summary of the NAEP audit goals, HumRRO staff made several presentations. These presentations were organized around the five activities specified in HumRRO's statement of work for the NCES Quality Assurance contract. However, each of the presentations was tailored to be responsive to the audit dimensions that had been communicated to them in advance of the meeting. There was opportunity for interactions between Buros participants and HumRRO staff members during these presentations. These interactions served as the primary information gathering process during the site visit. The four audit dimensions identified for HumRRO are *organizational characteristics*, *create the NAEP scales and links and analyze the data*, and *renewing and improving the assessment*. Evidence, findings, and recommendations pertinent to each of these dimensions are summarized below.

Organizational Characteristics

With regard to qualifications of the staff, it was found that the personnel assigned to the key activities for the Quality Assurance contract are well-qualified for their respective assignments. Laurie Wise has a long history with NAEP and is well-positioned to lead this effort. Wise was the lead staff person on Activity 1 (Past Problems) and participated in key components of Activity 2 regarding an analysis of the Procurement Model. Paul Sticha leads the effort on the Process Model (another component of Activity 2). His background in mathematical psychology makes him well-qualified to undertake this effort. Sunny Becker, deputy director for the contract, also has excellent credential for her roles and responsibilities on the contract through both her Ph.D. degree in quantitative methods and her experience with Prince George's County Research, Evaluation, and Accountability office. The primary person for the Site Visits (Activity 3), Carolyn Harris, also is well-qualified for this position. She has a Ph.D. in educational research and evaluation and several years of experience as an evaluator. Gene Hoffman is the lead person on the Special Studies activity and has a Ph.D. degree in industrial and organizational psychology. His background and years of experience with large scale Assessment programs provide sound credentials for the tasks for which he has primary responsibility. Steve Sellman leads Activity 5 (General). As HumRRO's vice president for strategic planning, he is well-positioned to lead this activity.

Several mechanisms are in place to support communications, both within HumRRO staff who work on the project and among contractors. It was reported that early in the contract, the HumRRO staff members met on a regular basis to share information, raise problems and concerns, and to resolve any issues. However, as experience with the contract matured, these regular meetings with the key staff members have been replaced with mostly e-mail communications on an as-needed basis. More formal and regular communications are maintained and documented with the contractors, principally with the Quality Assurance Council (QAC) and the Quality Control Team (QCT). These groups were formed in December 2003 in response to identified needs to enhance cross alliance communications regarding quality control issues. The QAC consists of representative from NCES, the NAEP Alliance, and HumRRO. The purpose of QAC is to facilitate the discussion of quality matters, develop broad quality control policies and standards, and to promote a cross-organizational atmosphere. The QCT also consists of representatives from each of the Alliance members and HumRRO. This team implements standards and policies articulated by QAC, coordinates quality control activities across the Alliance, develops tools and methods to address quality control issues, and informs QAC of critical quality control issues. The QAC meets quarterly and the QCT holds biweekly conference calls. There is a mechanism for documenting issues identified through these communications on a secure/private Web site that is only accessible to QAC and QCT members. NCES does not have access to this Web site; it was decided that this arrangement would support free and open discussion of problems and issues. HumRRO maintains minutes of these meetings and all issues are logged in the Process Improvement Log (PIL). Unresolved issues remain "open" on the PIL until resolution is obtained.

Communications are also fostered through the involvement of HumRRO through its roles in various meetings, including attendance and preparation of NCES-specific notes. These meetings include NABG, NAEP Validity Studies Panel; the Design and Analysis Committee; and other NCES contractor meetings (including Annual Design

Summit). HumRRO also organizes and coordinates meetings with the NAEP-Quality Assurance Panel and (as identified above) regular meetings of the QAC and QCT. All of these activities help promote an atmosphere of open and informed communications across the contractors and agencies responsible for the various aspects of NAEP.

Mechanisms are in place for problem identification and resolution. This is supported through the regular communications of the QCT and use of the secure Web site for posting of problems and resolutions through the process improvement log. One of the activities supported by the Quality Assurance contract was the Past Problems effort. Through interviews with Alliance members and others, HumRRO was able to document problems that occurred in the past and identify how these problems either were resolved or what steps should be taken to ensure they would not reoccur. Each contractor in the Alliance prepares a Quality Control plan on an annual basis. These QC plans are reviewed by HumRRO to ensure that appropriate QC plans and documentation are in place. This process helps to establish an environment that supports good quality control procedures and has the potential to be proactive in identification of potential problems and facilitate early resolution.

Within HumRRO, staff members have clear delineation of roles and responsibilities. Key personnel and their roles and functions are shown in the organizational chart that is attached.

There does not appear to be any concerns about potential conflicts of interest with other programs or products within the organization.

HumRRO, in terms of its role as evaluator, is responsive to requests and needs from NCES in all phases of its contract. Some aspects of its responsibilities are relatively fixed (e.g., site visits), but there is no QC plan in place to direct HumRRO's efforts.

Create scales and links and analyze data

HumRRO serves only a minor role at this time in this dimension. Some of the special studies it has conducted have looked at the replication of the full parameter estimates used in the IRT scaling and replication of Long Term Trend scaling, equating, and conditioning. There could be additional studies undertaken by HumRRO in this area, under the auspices of special studies. However, there are no plans at this time to undertake such studies. In HumRRO's role with the Validity Studies Panel, some work is done to examine the validity of score interpretations with the validity framework.

Renew and improve the assessment

Much of the work done through the QA contract could be viewed as a means of renewing and improving the assessment, although the focus of the effort is more on ensuring the quality of the current assessment design, development, delivery, scoring, and reporting. However, through these efforts, the potential exists for identification of means and mechanisms for program improvement.

Through the Past Problems activity, areas in which problems existed previously and their resolution strategies help to inform procedures for future program design and decisions. Through the analysis of the Procurement Model, issues related to quality assurance through the coordination of the contractors were identified and procedures put in place to enhance Alliance communication, problem identification, and problem resolution. The development of the Process Model has several potential benefits for assessment improvement and renewal. Although still a work in progress because the

dynamic dimensions of the process are yet to be fully modeled, the static models help to articulate the multitude of components and steps involved in the comprehensive assessment process. Once fully modeled, strategic planning could be aided through applications of the process model. Without this complex and comprehensive modeling, informed strategic planning would be more difficult, if not impossible. In addition, the process model could serve an important role in the preparation of requests for proposals for components of the assessment in the future.

The site visits also provide important information for improvement of the assessment. These site visits are designed to ensure that contractors comply with their quality control plans, but they also provide an opportunity to gather systematic information about how the system is working and where it needs adjustments. The documentation from the site visits could provide information about areas for assessment improvement, particularly regarding the process for administering the assessment. Currently, the information gathered from these site visits is not systematically being accumulated and evaluated for this purpose, but it could serve as a rich source for systemic program improvement. This effort could be enhanced through a more transparent comprehensive quality control plan for the site visits to ensure that the quality control dimensions across the contractors are considered through the site visit design.

The annual review of contractor QC plans again serve both an immediate need to ensure quality control through the assessment process and have the potential to provide information that would serve for assessment renewal and improvement. In its current implementation, through the QA contract, HumRRO tends to look at the static conditions that support the assessment program. With modest adjustments, these procedures could help inform, to a more systematic degree, the improvement and renewal of the assessment. Future negotiations might consider adding this dimension because HumRRO, with its broad and comprehensive knowledge base of the current assessment program, seems well positioned to serve as a conduit for information relevant to assessment renewal and improvement.

The Special Studies activity also provides opportunities to examine means and mechanisms for assessment renewal and improvement. Already mentioned are the special studies that support the strategies and procedures used to analyze the data. Other special studies have focused on anomalies that have appeared in the data, specific concerns about possible program issues, mechanisms for verifying the accuracy of the reporting of student demographic information, examining motivational issues related to 12th-grade assessments, and improvement of current practices in monitoring the quality of scoring of constructed response questions. The degree to which this special studies program could be more forward looking is a decision to be made in cooperation with NCES. However, the potential to conduct studies that could more directly inform assessment renewal and improvement is present through the opportunity to conduct these special studies. To date, there does not appear to be an articulated purpose or comprehensive plan for the special studies program.

Through the enhanced role of NAEP in the *NCLB* legislation, HumRRO anticipated the need to give more focused attention to issues related to examination security. This elevated attention is seen in its review of QC plans, its consideration of security as a component to the site visits, and through the development of the Process models.

Findings and Recommendations

Overall, HumRRO is providing excellent service as the Quality Assurance contractor for NAEP. It is hard to imagine how the NAEP contract, under the current procurement model, could be successful without an overarching agency whose primary role is to coordinate the quality of the component parts. HumRRO has served this role admirably and dynamically, adjusting procedures to be responsive to ongoing demands for communication and information. There are only a few areas in which improvements could be potentially beneficial:

- 1) A comprehensive Quality Control Plan from HumRRO could help support its quality control efforts and ensure that the goals of the contract are being achieved.
- 2) A comprehensive plan for the site visits is needed to ensure that all of the relevant quality control dimensions are being considered through the site visits.
- 3) A system for completing the feedback loop of information gained through the examination of the QC plans, recommendations from the site visits, and the QCT problem identification logs are used for system improvement.
- 4) An analysis of the goals for special studies program could help inform the direction and decision about additional studies.
- 5) An analysis of how information about the quality of the current assessment design, development, delivery, scoring, and reporting could be structured to more systematically inform assessment design and renewal should be conducted.

Materials Reviewed

Documents Reviewed Prior to HumRRO Site Visit

- Becker, D.E. (Sunny), Hoffman, R.G., Schantz, L., Stawarski, C., Schultz, S., Itchkawich, S. (April, 2004). Review of NAEP Quality-Control Plans for 2004. Alexandria, Va.: HumRRO
- Ford, L.A., Hoffman, R.G., Becker, D.E. (Sunny) (June, 2004). Potential Automated Data Checks of NAEP Student Demographic Data—Final Report. Alexandria, Va.: HumRRO.
- Hoffman, R. G., Wise, L.K., Sticha, P.J. (July, 2003). Review of NAEP Quality-Control Plans. Alexandria, Va.: HumRRO
- HumRRO (August, 2004). Development of NAEP Process Simulation Timelines FY04 Special Study Design Plan – FINAL. Alexandria, Va.: HumRRO.
- HumRRO (January, 2005) NAEP Validity Studies Expert Panel Meeting #28. Agenda, Minutes, and Briefing Book. Alexandria, Va.: HumRRO
- U.S. Department of Education, July 03, 2002. Statement of Work: National Assessment of Educational Progress: Quality Assurance of Process and Data Procurement
- a. Amendment to Statement of Work; Responses to Clarifying Questions
 - b. Modification of Contract, Sept. 24, 2002: Modification 9

- c. Modification of Contract, Sept. 24, 2002: Activity 3, Conduct Site Visits of NAEP Operations and Processes

Wise, L.L., Becker, D.E. (Sunny), Ramsberger, P.F. (July, 2003). Report on Past Problems. Alexandria, Va.: HumRRO.

Wise, L., Hoffman, R.G. (November, 2004). Technical Panel Meeting to Discuss the Implementation of Within- and Cross-grade Scaling for the NAEP 2009 Reading Assessment: Meeting Notes. Alexandria, Va.: HumRRO

Wise, L.L., Le, H., Hoffman, R.G., Becker, D.E. (Sunny) (September, 2004). Testing NAEP Full Population Estimates for Sensitivity to Violations of Assumptions—Final Report. Alexandria, Va.: HumRRO.

Documents Reviewed Following HumRRO Site Visit

Quality Assurance Checks for the 2002 Reading Assessment Results in Delaware

Quality Assurance Checks for the 2003 Reading Assessment Results

Potential Automated Data Checks of NAEP Student Demographic Data

Testing NAEP Full Population Estimate for Sensitivity to Violations of Model Assumptions

NAEP Charter School Questionnaire Focus Groups

Participation on NCES Technical Panel Meeting to Discuss the Implementation of Within- and Cross-grade Scaling for the 2009 NAEP Reading Assessment

Tracking of List Submission process and improvement

Final Initial NAEP Process Review Report

Literature Review from Ongoing Motivation Study

Appendix G-9: Pearson Educational Measurement (PEM)

Site Visit Team: Barbara Plake and Jim Impara, Buros Center for Testing
Date of Site Visit: Sept. 13, 2005

Audit Summary

Staff

Connie Smith – Account Manager for NAEP
Steve Kromer – General Manager
Mary Schulte - Information Technology Project Manager
Carolyn Loew - Lead Software Analyst, Processing and Scoring
Bill Buckles - Senior Project Manager, Scoring
Erica Hlebowitsh - Director, Software Solutions
Russ Vogt - Senior Project Manager, Printing
Jim Close - Project Manager, Quality
Pat Stearns - Project Manager, Packaging and Distribution

Following an initial contact with PEM by Drew Malizio, Barbara Plake communicated directly with Connie Smith at PEM about coordination of the site visit. Prior to the site visit, Plake shared with Smith the audit dimension and responsibility matrix. In collaboration with Plake, Smith drafted an agenda for the meeting which was finalized in advance of the meeting. Buros staff reviewed some materials in advance of the meeting.

In addition to several presentations by PEM personnel, the site visit consisted of several tours of selected facilities, including tours of the printing and shipping facility in Cedar Rapids and the receiving and scanning facility in Iowa City. These tours were informative about general procedures related to materials distribution and shipping and security. The following audit dimensions were identified prior to the site visit: *Organizational characteristics, Intended scope and uses of NAEP assessments, Administer the assessment, and Score the assessment and prepare final analysis database.* Based on the site visit, it appears that the role of PEM in the NAEP assessment process is somewhat more comprehensive than originally conceived in the audit plan. In the Administer the assessment dimension, PEM has responsibilities that go beyond the components that were originally identified, indicating a need for an expansion of that dimension to include printing the assessment, preparing the assessment bundles (including spiraling), and monitoring the assessments throughout the shipping and receiving processes. In addition, processes and procedures implemented by PEM have implications for renew and improve the assessment (13) dimension, which therefore should be added to the scope of PEM's involvement in NAEP.

Organizational characteristics

An organizational chart for PEM Assessment and Testing–U.S. was provided at the site visit. This chart was discussed early in the site visit, showing how NAEP responsibilities are situated in PEM's organizational structure. The majority of NAEP functions are organizationally located in the Publisher Services division. Key personnel for NAEP activities attended the meeting. Some of these staff members made

presentations regarding the role of their unit in NAEP procedures and processes and then responded to questions. Staff members seem to be well qualified for their respective positions.

Communications within the PEM NAEP staff members appears to be strong, consisting of weekly meetings and the use of peer involvement in many critical components of NAEP processes and procedures. Planning and organizational features are very strong, strengthened by systemic procedures for detailed specifications, documentation and record keeping. This is also evidenced by PEM's having several different types of organizational certifications (e.g., ISO 9002).

Communications with Alliance members, especially ETS and Westat are also strong. These communications are maintained through weekly conference calls with ETS and Westat, weekly calls with PEM and the NCES COR (Drew), participation in the weekly conference calls supported by HumRROs QCT and the periodic QAC meetings. The NAEP IMS system provides a portal for password-protected communications and the posting of the Process Improvement Log and issues identification and follow through.

Deadlines and handoffs are coordinated through meetings with the Alliance partners and close monitoring is maintained to ensure that the partners are in compliance. Systems supported by internally developed software keep track of target timelines and successful completion of target dates. Tensions were identified with pressure points and decisions that can put Alliance partners in stressful timeline situations. For example, NAGB must approve and sign off on all cognitive items (and background questions) and OMB must also issue an approval. The printed assessment documents must have the OMB release identification on the documents. Print runs cannot occur before approved information is secured. Because administration dates are fixed, any delay in receiving this approval information can put stress on the timeline for printing, packaging, and shipping.

Although not a direct responsibility of PEM (but rather that of the Alliance) a Quality Control plan is updated annually for the handoffs between Alliance partners. It would have been desirable that attention to this critical QA component would have occurred earlier in the contract life (2003–07; this handoff QC plan was first delivered for NAEP 2005 and will be updated for NAEP 2006, near the end of the contract period).

Intended scope and uses of NAEP assessment

Although PEM does not see a direct involvement of its processes and procedures in the articulation of the intended scope and uses of these assessments, there is a clear relationship between functions it performs and the quality and integrity of NAEP results. Quality is an overriding consideration in the institutional activities within PEM. Software systems, scanning technology, and other technological processes are in place to ensure that program specifications are honored throughout the printing, packaging, and shipping procedures. Several checks and security components support these activities. Once the test booklets are received, additional checks and validation efforts support the quality and integrity of the data capturing systems (image scanning, OMR scanning, Intelligent Character Recognition, data validation, and editing). PEM also serves a critical role in providing scores for open-ended responses. As pointed out in the section on Scoring the assessment, some areas were identified where additional validation efforts might strengthen the technical quality of the scores that result from these scoring procedures.

PEM has as its slogan, "What we do today will affect their tomorrow." In the spirit of this slogan, PEM recognizes that the procedures it uses in preparing the test

booklets for administration and how it captures the student responses has a critical affect on the validity of score interpretations. The old adage of “garbage in, garbage out” is directly relevant here; PEM takes seriously its responsibility in assurance that what it does today in preparing the assessments and capturing the data will affect the integrity results that are derived “tomorrow.”

Administer the assessment

As indicated earlier in this report, based on the information obtained through the audit process, it was determined that the scope of this component was not fully articulated in the Audit Dimensions document. PEM’s involvement in preparations of the assessments for administration is more comprehensive than originally conceived. Therefore, Burors augmented this component to more fully reflect the complex and comprehensive role that PEM plays in the preparation of the assessment materials for administration. These additional activities include printing, packaging, shipping, and receiving of the assessment materials.

Based on printing specifications received from ETS, PEM has the responsibility for printing the multiple test booklets and ensuring their quality. The integrity of this process is supported by several procedures including dedication of time for reviews of mock-ups that involve multiple review teams within PEM, ETS, NESSI, and AIR. The goal is to catch any printing issues early in the printing process when corrections can be achieved in an efficient and less costly manner. Once the mock-ups have been approved (and relevant green lights have been provided by governmental agencies), print runs are completed and delivered to PEM’s Cedar Rapids facility. At that facility, specifications are used to prepare the booklets for shipping, including the fulfillment of bundling specifications for packaging the materials for delivery to Westat test coordinators in the field. Several systems are in place to ensure that these specifications are fully complied with, including the use of scanning technology to check for a match with the specifications for booklet spiraling. These specifications are complex and the procedures appear to be effective in monitoring compliance with the specifications.

Use of several communications systems help support assessment administrators once the materials are in the field, including customer hotline support and fax communications. Communication linkages with Westat are maintained when the assessments are in the field to keep both partners fully informed of issues related to assessment receipt and delivery. PEM has put into place several “customer friendly” procedures to help ensure that the administrator in the field can achieve the intended administration procedures, maintain accurate assessment records, and return the materials in an efficient manner.

Once the assessments have been returned to PEM, additional systems are in place to monitor receipt control and security. PEM attempts to protect the security of the assessment through inventory systems to track receipt of all materials that were shipped. Materials are held in an “alert” area until there is a resolution of receipt control issues.

Once prepared for scanning, several checks are in place to protect the integrity of the scanned capture of the student responses. Multiple choice responses are captured electronically and prepared for transmittal to scoring procedures that are completed at ETS. Open-ended responses are captured by proprietary scanning software and prepared for use in human scoring under the direction of PEM scoring processes.

Following the completion of these data capturing procedures, data files are prepared to industry specifications and made available to ETS, Westat, AIR, and NCES.

Following this handoff, PEM's role in the life cycle of this NAEP assessment comes to an end. PEM warehouses student test booklets and ancillary materials used in NAEP assessments for an indeterminate period.

Scoring the assessment and prepare the final analysis database

PEM's role in this dimension principally lies in its preparation of the scorers for responses to open-ended NAEP prompts. The responsibility for training of the scorers switches from ETS (the item and rubric developers) to PEM as the open-ended questions move from pilot (when they are still in development) to operational, post calibration. In the scoring procedures, different issues are in place depending on whether the open-ended questions serve a trend or non-trend role. ETS has the responsibility for identifying and developing the training sets, and depending on the status of the questions (pilot/operational pre-calibration or not) ETS may or may not have additional training responsibilities. Regardless of whether PEM or ETS conducts the training, the scorers are recruited by PEM to meet PEM scorer eligibility and scoring is conducted in PEM's scoring facilities.

Current research studies are in place to explore alternative strategies for scoring procedures for trend responses. In the past, trend question scoring occurred as preplanned (and nontransparent) events in the scoring procedures. A stronger psychometric design for scoring of trend questions would be that they occur without knowledge of their "trend" status, integrated within the open-ended questions assigned to the scorers.

Procedures for gathering validity and reliability evidence involve the use of "backreading" by the scoring supervisor and randomly obtaining a second score for a percentage of the papers (either 5 percent or 25 percent depending on the volume of responses). Backreading is implemented as a mechanism for monitoring the calibration of scorers with intervention strategies in place for a scoring supervisor to take different actions depending on the severity of the problem. Supervisors may simply communicate (directly via face-to-face conference or indirectly via e-mail) with the scorer to alert him or her to concerns about score decisions or the supervisor may make a decision to "reset" a question and reseed it into the scorers' scoring set, perhaps following a retraining of one or a group of scorers.

Several issues were raised through the discussion about open-ended scoring. First, there does not appear to be a systematic use of "validity" papers, either for the non-trend or trend questions. For non-trend questions, it would be highly desirable to include validity check papers in the papers seeded to scorers. This is common practice in the scoring of performance assessments. Monitoring of scores on these validity papers would provide additional information to the scoring supervisor regarding the need for retraining or disqualification of a scorer. The issues are more complex with trend papers due to the changes that have occurred over time regarding the scoring of these papers and the need to replicate whatever idiosyncrasies might have been in place in the prior scoring procedures.

Second, the decisions regarding how the results from second scoring and supervisor backreading score results are used should be reconsidered. These results are used only for quantifying inter-rater reliability and for identification of scorer drift; these score values, regardless of whether they bring into question the accuracy of the first scorer's score value, do not alter the first score even when evidence might suggest they are inaccurate (unless the supervisory decides to disqualify, i.e., "reset", this question,

retrain, and then have the question reentered into the scorers' set of questions to score). Although, it could be perceived that it is PEM's responsibility only to provide the obtained score records to ETS for use with its scoring algorithms (which would be analogous to how ETS uses the scanned responses from the multiple choice questions), another perspective is that it is PEM's responsibility to ensure the validity of these open-ended scores that are transmitted to ETS for processing. This would be similar to the steps that PEM now carries out to ensure the validity of the scanned images for both the multiple-choice responses and the open-ended responses. Additional attention to the validity of the scores provided for the open-ended responses is desirable.

Through the use of standard confidentiality and nondisclosure procedures and through the intense and highly technical implementation of its scanning and security control systems, PEM appears to provide serious attention to the need to maintain security throughout its roles in the assessment process.

Renew and improve the assessment

Although not originally identified as a responsibility of PEM in the NAEP processes, it appears that PEM has implemented several software and technological innovations that provide support for the ongoing integrity and quality of NAEP assessments. These include systemic software and documentation systems, clear articulation of specifications for NAEP activities under the auspices of PEM, and the development and implementation of technological solutions to ensure compliance with packaging specifications, shipment and document receipt, and scanning methodologies. Due to the complexities of the NAEP assessment design, and the enhanced need for ensuring tracking of document shipping and receiving, these systems become more essential.

Findings and Recommendations

In most functions, PEM appears to be providing excellent service to the quality and integrity of the NAEP assessments. Strengths include the clear attention to systems approach to the development of specifications, software, and technical solutions to the preparation of the assessments for administration and scoring. There is a high sensitivity to its role in protecting the quality and security of the assessments. The one area in which additional attention may be needed involves the scoring of open-ended assessments. The role of validity papers and the decision rules about second scores (either by the scoring supervisor or peer panelists) should be reconsidered. Further, continued exploration of methods to improve the procedures used for scoring trend questions is recommended.

Materials Reviewed:

NAEP (2005). *NAEP Quality Assurance Procedures: Pearson Educational Measurement.*

NCS Pearson. *NAEP alliance Technical Proposal for Solicitation No. ED-02-R-0015: The National Assessment of Educational Progress (NAEP) 2003-2007, Task 7.1: MPS: Project Oversight and Management.*

This page intentionally left blank

Appendix G-10: Westat

Site Visit Team: Chad Buckendahl, Buros Center for Testing and Ed Wiley, University of Colorado

Date of Site Visit: July 11, 2005

Audit Summary

Staff

Nancy Caldwell - Project Director

Debbie Vivari – Director of Systems and Programming

David Morganstein – Vice President and Director of Statistical Staff

Diane Walsh - Deputy Project Director

Keith Rust – Vice President and Associate Director of statistical staff

Catrina Williams – Web Content Manager, NAEP State Service Center

In advance of the visit, Buros shared with Holly Spurlock, COR for AIR, information regarding the purpose of the audit, the comprehensive plan for the audit, and primary audit dimensions relevant to Westat. Using this information, Nancy Caldwell coordinated the preparations by Westat for the site visit. She communicated directly with Chad Buckendahl. Caldwell was given a preliminary agenda prior to the visit.

In addition, Buros was provided some documents prior to the site visit for review. Following the meeting Buros requested and received access through the IMS system to draft information regarding operational sampling procedures for the 2000–03 administrations.

Following introductions and a brief overview of Westat’s contract responsibilities to NCES and a brief summary of the audit goals, interviews with Westat staff were conducted. These interviews were organized around the six dimensions of the matrix that were identified as relevant to Westat. These interviews were the primary source of information for this preliminary summary. The six audit dimensions identified for Westat are *Organizational Characteristics, Conduct Field Tests, Sample Schools and Students, Administer the Assessment, Renew and Improve the Assessment, and Examination Security*.

Organizational characteristics

Westat has a long history of experience in the areas of sampling and large scale data collection. It is involved in a range of projects including studies in the health sciences, social services, education, and environmental services. Approximately 20–25 percent of its work is in education. Of the education projects, NAEP represents about half of the workload. Westat’s primary responsibilities for NAEP are in the areas of sampling and administration. Some key personnel for the project have been with Westat since it began its work with NAEP in 1983 (e.g., Cadell and Caldwell). Many have been involved in the project for a number of years and have played key roles in the evolution of the studies.

As a key leader in the NAEP Alliance (along with ETS and PEM), there is evidence of systematic meetings with internal staff through the quality control process

that documents these discussions. There are also weekly meetings with Westat's COR from NCES, Holly Spurlock, and frequent meetings among contractors involved in the alliance to address questions and challenges in operations.

Within the operations, there is the challenge of recruiting, hiring, and training as many as 5,000 field staff to administer NAEP. Although the pool has remained fairly stable, this is a nontrivial activity that requires exceptional coordination and training to ensure standardized administration nationally. The systems that Westat has developed to respond to this challenge have allowed them to expand to meet current demands but may be nearing a critical point in field personnel if NAEP continues to expand its data collection needs. Given *NCLB*'s focus, the desire to move to external administrators for greater independence in the data collection contributes to this challenge.

NAGB's interpretation of the 2002 legislation has led to greater involvement (e.g., 12th-grade assessment, special studies). This involvement has created some confusion about the decision-making process for NAEP activities. For example, Westat prepared for a fall field test for 12th-grade reading that was abruptly cancelled the week before our visit. It was unclear what the reasons were for the cancellation, but it impacts activity scheduling and prioritization for Westat.

The new NAEP Alliance contract has made it difficult to adhere to an agreed upon schedule among the contractors because there are a number of dependent components that require certain activities to occur before others. If there is a delay in one of these activities, it automatically challenges subsequent activities to meet original timelines. For example, delays in the CCD data pushed the 2006 sampling activities two months later than is typical. Although it is beyond the control of Westat, it has the potential to impact how quickly data can be handed off to PEM to create the shipping materials needed for the administration.

The review process and production of technical reports were also discussed. Westat noted that technical reports are not included in its contract as deliverables, but Westat believes that they are important as documented evidence of what was done. These reports serve as evidence of its processes and are also important as a knowledge transfer mechanism internally. NCES's strategy to put the technical reports for NAEP on the Web site has challenged Westat because the information that it contributes to the process is unique to a given study. There is also a concern that because the intent of the Web site report is to break down the large technical reports into sections; it may be difficult to integrate the full report if someone were to review it. Westat also mentioned that by "Web-izing" the technical report there may be edits that occur in the process that do not make it back into the full technical report that is referenced.

A number of problems were noted related to NCES's divisional and center review processes. First, when a report goes through adjudication, changes are made and then seen by different people who may recommend changes that were consistent with the original draft. Second, because there is a large turnover in the staff that reviews these documents, there appears to be little consistency from one round of review to the next creating additional delays in responding to reviewers' comments. Third, the timelines for these reviews often extend well beyond the agreed upon scope of the contract. For example, technical reports from 2000–01 are still in the divisional review process even though these contracts were completed a few years ago. More importantly, because the NCES Statistical Standards have changed, more recent reviews have commented on compliance with standards that were not in place when the work was originally completed. Because NCES is an important client, Westat is committed to finalizing the

technical reports from this previous work; however, time spent on these activities may also detract from responsibilities to ongoing projects.

Another example of extended timelines in the feedback process is that Westat has not yet received feedback on the 2003 draft technical report. When Westat has received feedback on previous technical reports it often occurred at times during the NAEP production process that did not allow it to pull staff members off projects to respond to comments on the report leading to further extensions of the review process. Comments that were not unique to Westat included a concern about the communicative skills and the technical competency of its reviewers. ETS has worked with Westat to provide much of the stylistic editing that is needed for the reports, so these changes were not as concerning as some of the substantive comments they have received. There was also a perception that some of the comments that were provided by reviewers did not address what they actually did operationally but that the reviewers commented on what they had hoped Westat would have done.

NAEP State Support Center

As a separate contract Westat operates a support center for the NAEP state coordinators. This effort began as a broader vision of having people in the states help recruit schools for participation, communicate NAEP information, conduct state data analyses, and write and disseminate reports. Although the state coordinators are contracted through NCES, they are supported for their activities through this contract with Westat. Westat provides professional development and training workshops on relevant topics, some of which are requested by the state coordinators. Many of these training sessions are offered via online meeting software (e.g., WebEx) to help control costs for participation. For example, one of the coordinators' training activities during the 2005 calendar year was a workshop on the basics of item response theory (IRT) presented by David Thissen from the University of North Carolina, Chapel Hill.

Another key element of the State Support Center is a secure Web site (NAEP Network) that serves as a link between the states and operations. NAGB also provides information from its periodic board meetings with the state coordinators to keep them in the loop of the board's policy considerations. Currently, there is not a formal curriculum for training activities for the state coordinators. However, this may change in time as the coordinators' role becomes more defined. The Network also allows coordinators to submit reports to the Web site for feedback from their coach or ambassador before submitting it to NCES for review. The home page and certain interior pages on the site are currently tracked, but a revision of the site will be able to track how users are accessing each page. Usage reports for the coordinators are provided to NCES so it can monitor the information that is being accessed by its contractors.

Conduct field trials

See the detailed report on NAEP sampling at the end of this report.

Sample schools and students

NAGB has paid particular attention to response rates and sample sizes as its responsibilities have shifted regarding the initial release of the reports. However, because NCES also has policies regarding appropriate sampling characteristics (through the NCES Statistical Standards), there may be some overlap or differences in the expectations. Taking these reports through NCES's divisional and center reviews and

then NAGB review with competing expectations have contributed to perceived delays in dissemination.

The recent decision to combine samples for state and national Main NAEP represents a significant change to the NAEP sampling design. Until *NCLB* effectively mandated state participation, an augmentation sample was required to measure students in states which declined to participate in State NAEP. There still appears to be separate samples collected to gather information for because of challenges with using a combined sample. The sample is augmented in many ways (e.g., minority, ELL, charter school, department of defense schools, etc.).

See the detailed report on NAEP sampling at the end of this report.

Administer the assessment

As mentioned earlier, the staffing needs for administering the assessment are great. In 2005 there were 5,000 field staff compared to 3,500 in 2003. Most are retired educators (approximately 90 percent) and there is relatively small turnover in the group (attrition was estimated to be 15 percent). Before training begins potential administrators undergo a background check and complete a home study course. There are a series of training activities that highlight the key elements of the administration process, particularly the ones that have the greatest chance to impact the validity of scores. These are documented in the training manuals for the assessment coordinators (ACs) and assessment administrators (AAs). Including information in the manual that points out these potential threats to validity is a novel approach to training and it also helps with quality control because administrators are more aware of the potential problems.

The ACs are responsible for assembling packages for the schools and are familiar with the forms, supervisors, and school questions. They also conduct the pre-assessment visits in January. There is a Quality Control Booklet that provides a scripted protocol for the pre-assessment visit to ensure standardization. As part of the quality control procedures, there is a QC log and information gathered from debriefing interviews that may impact the process.

Because of the detail-oriented nature of the six-week administration period, it adds another layer of challenge when special studies are included. For example, NAGB requested three special studies during the 2005 administration making the logistics to include these more difficult. Operations are given an opportunity to provide input on the design of some of the special studies (e.g., arts–clay, dance sequence; foreign language–performance tasks; science–manipulatives). There appears to be some tension between efforts by NAGB to be cutting edge versus what is practically and economically feasible. Some of these efforts are viewed as “piggybacked” onto administrations because it’s an opportunity to collect information while they are already in the schools.

As part of the ongoing monitoring of the assessment administration, supervisors visit each administration team one to two times during the administration. Following administration, Westat conducts callbacks to 25 percent of the schools. If something negative arises from the callback, it will contact all of the schools of the individual who was responsible for the administration. PEM also plays a role in the process by monitoring the delivery, receipt, and return of materials through the PEM Alert System. HumRRO also conducts more limited site visits (approximately 15 schools) and submits

observation reports to Janis Brown as an external quality check on the administration procedures.

Feedback helps the team make changes to the administration system. Debriefing forms and meetings with staff members, state coordinators, and NCES are all part of the process to learn about what worked and what could be improved about the administration process. This information is then integrated into the feedback loop when changes are suggested. Westat provided two examples of such changes. First, there is a policy that precludes administrators from opening bundles of booklets until one hour before the assessment. For large schools that may be administering multiple subjects, the administration team likely needs more time to prepare. Second, the timing of the pre-assessment visits currently occurs two to three weeks in advance of the assessment so there is a standardized amount of time before each administration. There has been a request to move all pre-assessment visits to January to make it easier to manage some of the logistics involved in the administration.

Westat expressed concerns about the burden of testing nationally and the potential impact on the operational administration. Particularly at the high school level, educators and students are becoming savvy about the tests that are more important versus ones that are voluntary without consequences. If this becomes more common, it has the potential to impact the recruitment, sampling, administration, and score interpretation. NAGB's change to measuring "preparedness" at the 12th grade will also require evaluating NAEP broadly to determine the impact of the directional shift. Expansion of the TUDA project would also impact the project.

Because 2007 will be a big year in the administration schedule, it will be important to stay on the critical path and not include a number of special studies that could interfere with the primary purpose of the assessment. NAGB is encouraged to consider special studies in the context of the assessment schedule as opposed to the relatively short notice of the more recent studies. This is especially important during administration years that include a third subject (e.g., science; writing–2007). The additional subject areas require large increases in staffing. Security procedures for administering the assessment are thorough and well-documented in the AC and AA manuals that were provided for review.

Renew and improve the assessment

Contractors in the NAEP process are generally required to be reactive rather than proactive because they are responding to a scope of work that is predefined with some flexibility expected. Therefore, it is often difficult for them to know when they can provide input on a proposed change in the process. NAGB's policy changes have led to its greater involvement in the details of the project rather than just at the policy level. It is often challenging for the operational staff to respond to requests for changes or special studies when a particular committee (e.g., COSDAM) or board members recommending these changes may not appreciate the operational difficulties of the request.

With the new online data tool (NAEP Data Explorer), there were concerns about confidentiality of scores given the opportunities for specificity of some searches. There has been a push for a data-swapping strategy (first discussed in the summer, 2004) that would be applied to a small number of variables for a small number of schools. Although describing the technique may present a public relations challenge, Westat does not expect that the method would impact decisions.

One of the challenges to changes or improvements in NAEP's methodologies is that ETS has used the rationale of needing to maintain trend as a reason to retain the status quo. If there are changes to the assessment, the interpretation of the trend data (short or long term) may be questioned. *NCLB* has helped facilitate some changes, but reading and mathematics are being kept together because of their role in the legislation. Another change that has been seen as positive is shifting to accommodations that are determined appropriate under the *IDEA*. Because accommodations are not the same nationally, there are some state-specific requirements and training for accommodations.

Changes in technology have allowed for systems that were not possible earlier. NCES has encouraged new innovations using technology, but then will often question budgets and timelines for implementation. One challenge to dramatically changing technology would be to provide training to the large contingent of field staff, many of whom may not be as familiar with current technology.

Westat also mentioned the need for NAEP to look more closely at computer-based assessment. Although historically the hurdles have been perceived as great, as barriers to access and computer literacy are reduced, this is a direction for the program to strongly consider. Some of the challenges to integrating technology into NAEP would include the logistics of computer administration (number), student verification (e.g., biometric screening), standardization of the testing environment, technology literacy of field staff, and systems for technology (e.g., security, firewalls).

Although Westat recommended pursuing computer-based and Web-based testing more aggressively (specifically for 12th grade and also for the writing subject area), it recognizes that because NAGB's frameworks do not currently include technology as a testing mode, NAGB may not be able to consider the recommendation. Because of the expansion over the past few years, it may be important to contain or rethink the growth of the current system to ensure that it is still meeting its primary mission.

Findings and Recommendations

Westat has a long history with NAEP and has used that experience to develop excellent internal systems and processes for how it administers NAEP. The challenge of recruiting, selecting, training, and evaluating the performance of the number of administrators is a daunting task that Westat has been able to perform admirably. In addition to this historical role, Westat has added contractual responsibilities in its leadership role in the State Support Center. Prior to our visit we were unaware that Westat played this additional role. There are only a few areas, pending the results of the sampling methodologies review, where recommendations for improved procedures and practices may be warranted:

1. We agree with Westat's recommendation to further explore computer or Web-based strategies for gathering information. For NAEP to be seen as a leader within the measurement community and for Westat to be able to continue to provide oversight of the data collection, more efficient approaches should be considered.
2. We encourage continued communication with NAGB through NCES about how policy decisions interact with operational decisions that then impact Westat's ability to effectively complete its responsibilities.
3. We recommend a greater degree of documented infrastructure and services for the State Service Center. Because we only learned about this additional role during our site visit, we were not adequately prepared to gather information about the

services it offers. We may be able to collect additional evidence about the service center from our visit with NAEP State coordinators at a later date.

Materials Reviewed:

Westat (2005). Assessment Coordinator Manual

Westat (2005). Assessment Administrator Manual

Westat (2004). Process Flowchart

Westat (2005). Quality Control Plan

Supplemental Report to Westat Site Visit: Review NAEP Sampling and Weighting

Prepared by: Edward W. Wiley, University of Colorado at Boulder

NAEP sampling and weighting are accomplished through multiple stages that occur throughout each year of assessment administration. The 2003 NAEP administration (the most recent one for which technical documentation is available, although only in draft form) included national (“Main NAEP”), state (“State NAEP”), and urban (“Trial Urban District Assessment” or “TUDA”) assessments in mathematics and reading. Westat is generally responsible for all aspects of sampling, weighting, and field operations (including data collection) employed in the NAEP program; the processes used by Westat for Main and State NAEP⁸ in 2003 are detailed below (some technical documentation for 2003 was omitted from the NAEP Web site; when no 2003 information is available this report draws upon documentation from the 2002 administration instead).

This report contains two main sections. The first section details the procedures used for sampling and weighting in NAEP and provides selected results from these procedures. The second section provides an evaluation of the procedures discussed in the first section and raises several unanswered questions that should be given additional attention. Following these two sections are a list of sources used for this report and a summary outline of major steps involved in NAEP sampling and weighting.

Sample Design

The NAEP sample design is revised annually through a collaborative effort led by Westat and involving all members of the NAEP Alliance. ETS and AIR specify school and student sample sizes required to support robust analysis. Westat develops school and student sample designs based on these specifications. All NAEP Alliance members are provided the opportunity to review and comment on the preliminary sample design in order that Alliance-wide consensus on the design may be reached. Alliance member comments and suggestions serve as checks in the quality control process; revisions and corrections are incorporated until Alliance-wide consensus is achieved. Once available the final specifications are posted to IMS.

State NAEP results are estimated using representative probability samples of students in public schools within each state; since 2002, state samples have been aggregated and augmented with a sample of students from non-public schools to serve as the national sample used for Main NAEP results. As such, this section will first describe sampling processes for public schools selected for 2003 State NAEP and then will discuss processes for augmenting the aggregate public sample for 2003 Main NAEP.

Sampling Design: 2003 State NAEP

State results are reported by jurisdiction; these include individual states, U.S. territories, Bureau of Indian Affairs schools (though BIA school declined to participate in 2003), Department of Defense schools, and school districts chosen for the Trial Urban District Assessment. In 2003, samples specific to each jurisdiction were targeted at 6,150 students, generally comprised of 62 students sampled from each of about 105 schools. The constant target of 6,150 students per jurisdiction was intended to provide aggregate estimates similar in precision and facilitate subgroup estimates as well.

⁸ Long Term Trend was not administered in 2005.

Public School Sample

A comprehensive list of public schools in each jurisdiction was needed to draw the 2003 school samples. To obtain this list, Westat first obtained the CCD file corresponding to the 2000–01 school year, selecting from this list all public schools operating during that year. Because of the timing of the CCD releases, Westat receives preliminary (rather than final) CCD files; these preliminary files are checked against the most recently adjudicated CCD files. Westat completes range and consistency checks to ensure (a) that it contains at least the minimum number of public schools required for sampling, (b) that the required information data fields are correctly displayed, and (c) that current school information data fields are consistent with data fields from previous files. School locale codes, student enrollment, percent minority student enrollment, and other key variables are compared across the current preliminary CCD and the most recently adjudicated CCD. Westat provides to NCES a summary of all schools whose current preliminary CCD data differ from the most recently adjudicated CCD data.

New Public Schools

The public school list based on CCD was augmented with schools newly eligible because they had opened or restructured between the 2000–01 school year (reflected in CCD) and 2002–03 school year (the year of assessment). In small districts new schools were identified during school recruitment. In a sample of larger districts new schools were identified via direct inquiry; weights for schools identified by this district sample were adjusted to reflect the use of a sample of districts rather than all districts.

To evaluate the quality of the new school survey, replies from public school districts are tracked to identify nonrespondents. Reply information is reviewed to determine whether schools identified as newly opened may in fact represent existing schools that have recently been renamed. This is accomplished by searching the CCD file by school name to see if a school with the same name already exists in the district being reviewed. In cases in which it is difficult to determine whether a school is in fact “new,” district or state department of education Web sites are checked to assist in making a final determination. District response rates and information about newly opened schools are summarized and questionable cases are identified. In cases for which corrective action is needed, follow-ups are conducted with nonresponding districts until either (a) a 100 percent response rate is obtained, or (b) the sampling deadline date is reached. All revisions and corrections are incorporated into CCD files as appropriate.

New school data for 2003 cannot currently be accessed via the NAEP Web site; however, data from 2002 suggests that new schools comprise 1–2 percent of the overall school samples.

Public School Stratification

State samples are selected on the basis of a stratified two-stage design (reflecting sampling of both schools and students within those schools). Schools are selected with probability proportional to a measure based on estimated enrollment in assessed grades. Schools with large numbers of minority students are sampled at twice the rate of other schools. Within each jurisdiction schools are stratified by the combination of charter status, urbanization, and minority class.⁹ Within each stratum schools are sorted by either state achievement data (when such data can be provided by jurisdictions) or by median income¹⁰ of households sharing the same ZIP Code as the school (in the absence of achievement data).

⁹ Data for these three measures is taken from the most recent preliminary files from NCES Common Core of Data

¹⁰ From the 1990 U.S. Census

From this sorted, stratified frame schools were selected via systematic random sampling (this is known as “implicit stratification”).

The goal of stratification is to minimize sampling error. In other words, stratification is carried out in the complex manner described above to best match each jurisdiction’s school sample to that jurisdiction’s population. Comparing sample distributions (that is, the characteristics of schools sampled through the stratified design) and population distributions (represented by the original frame) provides an indication of how well stratification did, in fact, minimize sampling error. Comparisons based on school characteristics and levels of achievement are not available for 2003 via the NAEP Web site; however, results of comparisons carried out by Westat for the 2002 assessment are summarized on the NAEP Web site as follows:

“...aggregations were computed for percent Blacks, Hispanics, Asians, and American Indians, and for mean median income and type of location, by jurisdiction. These aggregations were also computed for state achievement data, in those states for which we had data. Two-sided p-values were calculated to test the null hypothesis that the difference between sample and frame is zero, using the jackknife standard error of the sample aggregation (note that the frame aggregation is treated as having no sampling error, as there is no sampling process in developing the frame, except for the very limited area portion of the Private School Survey). It should be expected that many of the p-values would be small simply from randomness, as so many p-values were calculated. The results are summarized as follows:

- Of the differences that are significant, all but four absolute differences involving percentages are less than a percentage point, with most being near zero.
- Of 96 total differences that were calculated for median income, only 12 differences reached the nominal 5 percent level of significance.
- Of 68 total differences that were calculated for achievement scores, only five differences reached the nominal 5 percent level of significance.

Ineligible Schools

Some schools sampled were subsequently found to be ineligible for participation. These schools fell into one of two broad classes: schools that had closed since 2000–01 or no longer offered the grade of interest; and special schools not eligible for the NAEP assessment. In such cases, sampled schools were coded as ineligible. Numbers and percentages of schools identified as ineligible are reported in the two tables that follow. In many ways these results are not surprising; the states with the greatest proportion of ineligible schools tend to be states with many small rural schools. These small schools are more likely to be impacted by the two most common factors leading to school ineligibility—school closing and the lack of students in assessed grades.

Table 1. Eligibility of Sampled Schools By Jurisdiction

Jurisdiction	Fourth grade			Eighth grade		
	CCD school sample	Ineligible schools	Eligible school sample	CCD school sample	Ineligible schools	Eligible school sample
Total	7,618	381	7,237	6,272	480	5,792
Alabama	120	8	112	118	14	104
Alaska	188	26	162	140	30	110
Arizona	128	6	122	129	11	118
Arkansas	124	5	119	117	8	109
California	265	9	256	203	13	190
Colorado	125	1	124	118	3	115
Connecticut	113	1	112	106	2	104
Delaware	106	17	89	54	17	37
Florida	112	6	106	114	16	98
Georgia	162	6	156	123	6	117
Hawaii	109	2	107	71	3	68
Idaho	132	7	125	99	8	91
Illinois	181	7	174	177	7	170
Indiana	116	5	111	108	9	99
Iowa	141	4	137	119	1	118
Kansas	150	12	138	129	3	126
Kentucky	127	6	121	119	6	113
Louisiana	119	9	110	122	26	96
Maine	159	7	152	114	4	110
Maryland	109	1	108	107	3	104
Massachusetts	170	5	165	136	3	133
Michigan	140	4	136	114	3	111
Minnesota	123	9	114	121	14	107
Mississippi	117	6	111	118	10	108
Missouri	128	2	126	120	2	118
Montana	206	16	190	153	16	137
Nebraska	209	47	162	167	37	130
Nevada	114	3	111	73	6	67
New Hampshire	126	2	124	84	0	84

Continues next page

Table 1. Eligibility of Sampled Schools By Jurisdiction (Continued)

Jurisdiction	Fourth Grade			Eighth Grade		
	CCD school sample	Ineligible schools	Eligible school sample	CCD school sample	Ineligible schools	Eligible school sample
New Jersey	116	5	111	110	2	108
New Mexico	123	3	120	109	12	97
New York	155	6	149	160	12	148
North Carolina	158	5	153	136	3	133
North Dakota	216	5	211	158	12	146
Ohio	174	6	168	147	18	129
Oklahoma	140	3	137	130	1	129
Oregon	133	7	126	119	9	110
Pennsylvania	115	1	114	106	3	103
Rhode Island	118	4	114	59	4	55
South Carolina	113	7	106	108	10	98
South Dakota	209	15	194	159	17	142
Tennessee	118	2	116	114	6	108
Texas	204	7	197	155	9	146
Utah	114	0	114	99	3	96
Vermont	183	2	181	109	3	106
Virginia	117	1	116	111	4	107
Washington	118	9	109	114	11	103
West Virginia	149	12	137	106	11	95
Wisconsin	131	4	127	116	11	105
Wyoming	193	20	173	113	21	92
American Samoa	†	†	†	22	0	22
Bureau of Indian Affairs	2	0	2	2	1	1
District of Columbia	126	8	118	44	6	38
DDESS ¹	41	1	40	15	0	15
DoDDS ²	99	9	90	64	8	56
Puerto Rico	110	0	110	106	2	104
Virgin Islands	24	0	24	8	0	8

† Not applicable.

¹ Department of Defense Domestic Dependent Elementary and Secondary Schools.² Department of Defense Dependents Schools (Overseas).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Table 2. Number of ineligible sampled schools, grades 4 and 8: By ineligibility type, 2003

Eligibility status	Grade 4		Grade 8	
	# Schools	% of Sample	# Schools	% of Sample
Total Sampled Schools	7,618	100	6,272	100
Closed	108	1.4	84	1.3
Not a regular school	71	0.9	126	2
Does not offer sampled grade	159	2.1	209	3.3
No eligible students in sampled grade	36	0.5	55	0.9
Duplicate on sampling frame	3	0	1	0
Other ineligible	4	0.1	5	0.1
Eligible schools	7,237	95	5,792	92.4

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Public Schools: Sample Sufficiency Check

The number of sampled schools and the implied number of sampled students are compared to sample size requirements specified in the annual sample design. Westat statisticians review tabulation reports showing sample counts by selected characteristics specified in the annual sample design. Any samples that do not yield at least the minimum number of students specified in the annual sample design are redrawn.

Eligible sampled schools were assigned assessment sessions on the basis of enrollment of students eligible for assessment at the appropriate grades. Although larger schools were assigned more than one assessment sessions, most schools were assigned a single session.

Public School Substitutes

The refusal to participate of sampled schools introduces a potential bias into NAEP estimates; the magnitude of such bias is related to the degree to which schools that refuse are systematically different from those that agree to participate. Two strategies can be employed to deal with school refusals—replacement with substitute schools and weight adjustment for school nonresponse. The decision of whether to recruit substitutes falls to the NAEP state coordinator. Substitute schools were rarely activated for 2003 State NAEP; only a single school (in the BIA jurisdiction) was used. (Substitutes were used more frequently for the private schools sampled as part of the national assessment.) Substitute schools were selected on the basis of a distance measure generated to identify substitute candidates within the same state and urbanicity and most similar in terms of minority percentage, grade enrollment, and average achievement or median income. Several sampled schools did not have available substitutes. New schools were not assigned substitutes.

Westat produces materials containing information for sampled schools and substitutes. These materials are used by NAEP state coordinators, field staff, home office staff, and other Alliance members. They include listings of schools by selected categories, address labels, and activity summary sheets. Manual reviews are conducted to match the information contained in the generated lists to the master school information file. Discrepancies between data contained in the lists and the master file are resolved

through an iterative process of revision of programming specifications and generation of new lists.

Public School Response

Sampled schools eligible for assessment are recruited to participate in mathematics and reading. The target for participation, established by NCES standards, is 85 percent or greater weighted response. This rate was achieved in most cases in 2003, alleviating the need to recruit substitute schools because of failure to meet NCES standards. Even when this target is met, however, assessment results may be subject to nonresponse bias; this is discussed in greater detail in the final section of this report.

The following four tables detail weighted and unweighted response rates by jurisdiction for the 2003 State NAEP assessments in fourth and eighth grade.

Table 3. Counts and response rates of eligible sampled schools and recruited substitute schools, grade 4: By participating jurisdiction, 2003

Jurisdiction	Total eligible school sample (number)	Non-responding schools (number)	Responding schools (number)	Unweighted response rate before substitution (percent)	Recruited cooperating substitutes (number)	Unweighted response rate after substitution (percent)
Alabama	112	0	112	100	0	100
Alaska	162	3	159	98.1	0	98.1
Arizona	122	1	121	99.2	0	99.2
Arkansas	119	0	119	100	0	100
CA–Los Angeles	83	0	83	100	0	100
CA–San Diego	55	0	55	100	0	100
California	256	2	254	99.2	0	99.2
Colorado	124	0	124	100	0	100
Connecticut	112	1	111	99.1	0	99.1
Delaware	89	1	88	98.9	0	98.9
Florida	106	0	106	100	0	100
GA–Atlanta	50	0	50	100	0	100
Georgia	156	0	156	100	0	100
Hawaii	107	0	107	100	0	100
Idaho	125	0	125	100	0	100
IL–Chicago	83	0	83	100	0	100
Illinois	174	0	174	100	0	100
Indiana	111	0	111	100	0	100
Iowa	137	1	136	99.3	0	99.3
Kansas	138	0	138	100	0	100
Kentucky	121	0	121	100	0	100
Louisiana	110	0	110	100	0	100
Maine	152	0	152	100	0	100
Maryland	108	0	108	100	0	100
MA–Boston	59	0	59	100	0	100
Massachusetts	165	0	165	100	0	100
Michigan	136	0	136	100	0	100

Continues next page

Table 3. Counts and response rates of eligible sampled schools and recruited substitute schools, grade 4: By participating jurisdiction, 2003 (Continued)

Jurisdiction	Total eligible school sample (number)	Non-responding schools (number)	Responding schools (number)	Unweighted response rate before substitution (percent)	Recruited cooperating substitutes (number)	Unweighted response rate after substitution (percent)
Minnesota	114	1	113	99.1	0	99.1
Mississippi	111	0	111	100	0	100
Missouri	126	0	126	100	0	100
Montana	190	3	187	98.4	0	98.4
Nebraska	162	3	159	98.1	0	98.1
Nevada	111	0	111	100	0	100
New Hampshire	124	1	123	99.2	0	99.2
New Jersey	111	1	110	99.1	0	99.1
New Mexico	120	1	119	99.2	0	99.2
NY–New York City	79	0	79	100	0	100
New York	149	0	149	100	0	100
NC–Charlotte	51	0	51	100	0	100
North Carolina	153	0	153	100	0	100
North Dakota	211	0	211	100	0	100
OH–Cleveland	56	0	56	100	0	100
Ohio	168	0	168	100	0	100
Oklahoma	137	0	137	100	0	100
Oregon	126	1	125	99.2	0	99.2
Pennsylvania	114	0	114	100	0	100
Rhode Island	114	0	114	100	0	100
South Carolina	106	0	106	100	0	100
South Dakota	194	3	191	98.5	0	98.5
Tennessee	116	0	116	100	0	100
TX–Houston	80	0	80	100	0	100
Texas	197	0	197	100	0	100
Utah	114	1	113	99.1	0	99.1
Vermont	181	2	179	98.9	0	98.9
Virginia	116	0	116	100	0	100
Washington	109	0	109	100	0	100
West Virginia	137	0	137	100	0	100
Wisconsin	127	0	127	100	0	100
Wyoming	173	1	172	99.4	0	99.4
District of Columbia	118	0	118	100	0	100
DDESS ¹	40	1	39	97.5	0	97.5
DoDDS ²	90	2	88	97.8	0	97.8
Puerto Rico	110	0	110	100	0	100
Virgin Islands	24	0	24	100	0	100

¹ Department of Defense Domestic Dependent Elementary and Secondary Schools.

² Department of Defense Dependents Schools (Overseas).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, NAEP 2003 Assessment.

Table 4. Counts and response rates of eligible sampled schools and recruited substitute schools, grade 8: By participating jurisdiction, 2003

Jurisdiction	Total eligible school sample (number)	Non-responding schools (number)	Responding schools (number)	Unweighted response rate before substitution (percent)	Recruited cooperating substitutes (number)	Unweighted response rate after substitution (percent)
Alabama	104	0	104	100	0	100
Alaska	110	5	105	95.5	0	95.5
Arizona	118	0	118	100	0	100
Arkansas	109	0	109	100	0	100
CA–Los Angeles	67	0	67	100	0	100
CA–San Diego	28	0	28	100	0	100
California	190	1	189	99.5	0	99.5
Colorado	115	0	115	100	0	100
Connecticut	104	0	104	100	0	100
Delaware	37	0	37	100	0	100
Florida	98	1	97	99	0	99
GA–Atlanta	16	0	16	100	0	100
Georgia	117	0	117	100	0	100
Hawaii	68	1	67	98.5	0	98.5
Idaho	91	0	91	100	0	100
IL–Chicago	83	0	83	100	0	100
Illinois	170	0	170	100	0	100
Indiana	99	0	99	100	0	100
Iowa	118	2	116	98.3	0	98.3
Kansas	126	0	126	100	0	100
Kentucky	113	0	113	100	0	100
Louisiana	96	0	96	100	0	100
Maine	110	0	110	100	0	100
Maryland	104	8	96	92.3	0	92.3
MA–Boston	34	0	34	100	0	100
Massachusetts	133	1	132	99.2	0	99.2
Michigan	111	0	111	100	0	100
Minnesota	107	0	107	100	0	100
Mississippi	108	0	108	100	0	100
Missouri	118	0	118	100	0	100
Montana	137	4	133	97.1	0	97.1
Nebraska	130	1	129	99.2	0	99.2
Nevada	67	0	67	100	0	100

Continues next page

Table 4. Counts and response rates of eligible sampled schools and recruited substitute schools, grade 8: By participating jurisdiction, 2003 (Continued)

Jurisdiction	Total eligible school sample (number)	Non-responding schools (number)	Responding schools (number)	Unweighted response rate before substitution (percent)	Recruited cooperating substitutes (number)	Unweighted response rate after substitution (percent)
New Hampshire	84	0	84	100	0	100
New Jersey	108	1	107	99.1	0	99.1
New Mexico	97	0	97	100	0	100
NY–New York City	77	0	77	100	0	100
New York	148	0	148	100	0	100
NC–Charlotte	29	0	29	100	0	100
North Carolina	133	0	133	100	0	100
North Dakota	146	0	146	100	0	100
OH–Cleveland	35	0	35	100	0	100
Ohio	129	0	129	100	0	100
Oklahoma	129	0	129	100	0	100
Oregon	110	0	110	100	0	100
Pennsylvania	103	0	103	100	0	100
Rhode Island	55	0	55	100	0	100
South Carolina	98	0	98	100	0	100
South Dakota	142	0	142	100	0	100
Tennessee	108	0	108	100	0	100
TX–Houston	38	0	38	100	0	100
Texas	146	0	146	100	0	100
Utah	96	1	95	99	0	99
Vermont	106	2	104	98.1	0	98.1
Virginia	107	0	107	100	0	100
Washington	103	0	103	100	0	100
West Virginia	95	0	95	100	0	100
Wisconsin	105	0	105	100	0	100
Wyoming	92	0	92	100	0	100
American Samoa	22	0	22	100	0	100
District of Columbia	38	0	38	100	0	100
DDESS ¹	15	1	14	93.3	0	93.3
DoDDS ²	56	2	54	96.4	0	96.4
Puerto Rico	104	0	104	100	0	100
Virgin Islands	8	0	8	100	0	100

¹ Department of Defense Domestic Dependent Elementary and Secondary Schools.

² Department of Defense Dependents Schools (Overseas).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Table 5. Weighted counts and response rates of eligible sampled schools and recruited substitute schools, grade 4: By participating jurisdiction, 2003

Jurisdiction	Weighted number of eligible schools	Weighted number of nonresponding schools	Weighted number of responding schools	Weighted response rate (percent)
Alabama	60,365	0	60,365	100
Alaska	9,251	64	9,187	99.3
Arizona	78,464	228	78,236	99.7
Arkansas	36,509	0	36,509	100
CA–Los Angeles	62,229	0	62,229	100
CA–San Diego	12,069	0	12,069	100
California	498,992	4,849	494,142	99
Colorado	57,436	0	57,436	100
Connecticut	44,476	428	44,048	99
Delaware	10,994	147	10,847	98.7
Florida	190,736	0	190,736	100
GA–Atlanta	5,480	0	5,480	100
Georgia	118,669	0	118,669	100
Hawaii	14,799	0	14,799	100
Idaho	18,454	0	18,454	100
IL–Chicago	35,976	0	35,976	100
Illinois	155,923	0	155,923	100
Indiana	81,531	0	81,531	100
Iowa	34,812	67	34,745	99.8
Kansas	33,286	0	33,286	100
Kentucky	47,120	0	47,120	100
Louisiana	58,570	0	58,570	100
Maine	15,406	0	15,406	100
Maryland	66,972	0	66,972	100
MA–Boston	5,020	0	5,020	100
Massachusetts	74,181	0	74,181	100
Michigan	134,727	0	134,727	100
Minnesota	60,412	20	60,393	100
Mississippi	39,448	0	39,448	100
Missouri	72,863	0	72,863	100
Montana	11,117	41	11,076	99.6
Nebraska	21,027	141	20,885	99.3
Nevada	28,342	0	28,342	100
New Hampshire	16,912	26	16,886	99.8
New Jersey	99,124	976	98,148	99
New Mexico	25,414	243	25,171	99
NY–New York City	80,552	0	80,552	100
New York	216,892	0	216,892	100
NC–Charlotte	8,293	0	8,293	100

Continues next page

Table 5. Weighted counts and response rates of eligible sampled schools and recruited substitute schools, grade 4: By participating jurisdiction, 2003 (Continued)

Jurisdiction	Weighted number of eligible schools	Weighted number of non-responding schools	Weighted number of responding schools	Weighted response rate (percent)
North Carolina	105,428	0	105,428	100
North Dakota	8,048	0	8,048	100
OH–Cleveland	6,948	0	6,948	100
Ohio	149,651	0	149,651	100
Oklahoma	46,476	0	46,476	100
Oregon	40,432	13	40,419	100
Pennsylvania	138,931	0	138,931	100
Rhode Island	12,367	0	12,367	100
South Carolina	51,794	0	51,794	100
South Dakota	9,323	17	9,306	99.8
Tennessee	74,771	0	74,771	100
TX–Houston	17,956	0	17,956	100
Texas	331,644	0	331,644	100
Utah	36,674	33	36,641	99.9
Vermont	8,122	58	8,064	99.3
Virginia	98,082	0	98,082	100
Washington	74,278	0	74,278	100
West Virginia	20,364	0	20,364	100
Wisconsin	62,669	0	62,669	100
Wyoming	6,364	4	6,360	99.9
Charter Schools ¹	1,130	8	1,122	99.3
District of Columbia	6,348	0	6,348	100
DDESS ²	3,182	25	3,157	99.2
DoDDS ³	6,464	81	6,383	98.7
Puerto Rico	51,343	0	51,343	100
Virgin Islands	1,495	0	1,495	100

¹ The special charter school study was conducted only in fourth grade in the 2003 assessment.

² Department of Defense Domestic Dependent Elementary and Secondary Schools.

³ Department of Defense Dependents Schools (Overseas).

NOTE: The weighted number of recruited cooperating substitutes at the grade 4 level was zero for all jurisdictions in the 2003 state assessment.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Table 6. Weighted counts and response rates of eligible sampled schools and recruited substitute schools, grade 8: By participating jurisdiction, 2003

Jurisdiction	Weighted number of eligible schools	Weighted number of nonresponding schools	Weighted number of responding schools	Weighted response rate (percent)
Alabama	54,248	0	54,248	100
Alaska	9,517	72	9,445	99.2
Arizona	72,365	0	72,365	100
Arkansas	32,693	0	32,693	100
CA—Los Angeles	47,959	0	47,959	100
CA—San Diego	9,818	0	9,818	100
California	445,095	4,325	440,770	99
Colorado	58,348	0	58,348	100
Connecticut	41,626	0	41,626	100
Delaware	9,005	0	9,005	100
Florida	177,800	1,831	175,970	99
GA—Atlanta	4,286	0	4,286	100
Georgia	113,615	0	113,615	100
Hawaii	13,237	4	13,233	100
Idaho	19,041	0	19,041	100
IL—Chicago	34,810	0	34,810	100
Illinois	154,918	0	154,918	100
Indiana	73,614	0	73,614	100
Iowa	37,609	424	37,185	98.9
Kansas	36,975	0	36,975	100
Kentucky	50,132	0	50,132	100
Louisiana	54,349	0	54,349	100
Maine	17,052	0	17,052	100
Maryland	64,435	4,980	59,456	92.3
MA—Boston	5,264	0	5,264	100
Massachusetts	74,428	711	73,716	99
Michigan	132,950	0	132,950	100
Minnesota	64,066	0	64,066	100
Mississippi	38,334	0	38,334	100
Missouri	69,393	0	69,393	100
Montana	12,350	236	12,114	98.1
Nebraska	21,719	14	21,705	99.9
Nevada	26,718	0	26,718	100
New Hampshire	16,932	0	16,932	100
New Jersey	95,447	919	94,528	99
New Mexico	24,520	0	24,520	100
NY—New York City	67,380	0	67,380	100
New York	205,850	0	205,850	100
NC—Charlotte	7,831	0	7,831	100

Continues next page

Table 6. Weighted counts and response rates of eligible sampled schools and recruited substitute schools, grade 8: By participating jurisdiction, 2003 (Continued)

Jurisdiction	Weighted number of eligible schools	Weighted number of nonresponding schools	Weighted number of responding schools	Weighted response rate (percent)
North Carolina	107,103	0	107,103	100
North Dakota	8,524	0	8,524	100
OH–Cleveland	5,830	0	5,830	100
Ohio	140,976	0	140,976	100
Oklahoma	48,378	0	48,378	100
Oregon	40,524	0	40,524	100
Pennsylvania	140,209	0	140,209	100
Rhode Island	12,100	0	12,100	100
South Carolina	52,362	0	52,362	100
South Dakota	10,055	0	10,055	100
Tennessee	66,036	0	66,036	100
TX–Houston	12,798	0	12,798	100
Texas	324,436	0	324,436	100
Utah	35,153	115	35,038	99.7
Vermont	7,749	189	7,560	97.6
Virginia	94,110	0	94,110	100
Washington	75,548	0	75,548	100
West Virginia	20,277	0	20,277	100
Wisconsin	64,824	0	64,824	100
Wyoming	7,307	0	7,307	100
American Samoa	1,179	0	1,179	100
District of Columbia	3,755	0	3,755	100
DDESS ¹	1,797	25	1,772	98.6
DoDDS ²	4,884	54	4,830	98.9
Puerto Rico	44,602	0	44,602	100
Virgin Islands	1,618	0	1,618	100

¹ Department of Defense Domestic Dependent Elementary and Secondary Schools.

² Department of Defense Dependents Schools (Overseas).

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Sample Design: 2003 Main NAEP

Since 2002 State NAEP samples have included fourth and eighth grades in public schools in participating jurisdictions. In choosing to use combined state samples rather than a single national sample NAEP has traded efficiency (combined state samples are roughly ten times the size of a single national sample) for precision (greater samples allow more precise measurement). If national assessment was the only goal this tradeoff may not be considered worthwhile; however, because precision at the individual state level is also required, there is little reason to prefer a separate national sample solely in terms of the efficiency tradeoff. ETS research has detailed the additional precision of

combined state samples, only slight discrepancies between combined and national estimates, smaller standard errors associated with combined estimates, and a reduced need for poststratification adjustments in using combined samples. The use of combined samples appears to be a change for the better for Main NAEP.

To obtain a nationally representative sample for Main NAEP, state samples must be supplemented with public school samples for those jurisdictions which ultimately did not participate in State NAEP as well as a nationally representative private school sample. (Main NAEP also included a pilot study at 12th grade; this pilot study is not directly covered in this report.)

Public school sample augmentation is relatively straightforward. Jurisdiction school samples were established before it was known exactly which jurisdictions would ultimately participate in the state program. School samples were drawn from all jurisdictions as part of State NAEP—including those jurisdictions that did not ultimately participate in State NAEP—to ensure that the Main NAEP sample was representative. In the state sampling process probabilities of selection were calculated for each school based on jurisdiction. For Main NAEP these probabilities were recomputed to represent likelihood of selection as part of a national sample (rather than within each jurisdiction).

Private School Sample

Adding private schools to the national sample involved a separate sample selection process. This was similar to the public school sampling process used for State NAEP. The 1999–2000 NCES Private School Survey (PSS) provided the basis for the private school frame (this was the same frame used for the 2002 assessment). The PSS file is abstracted to obtain a comprehensive listing of nonpublic schools eligible for sample inclusion. Similar to construction of the public school frame, Westat must ensure that the PSS list is comprehensive for sampling purposes. Range and consistency checks are run on the PSS abstraction to ensure (a) that it contains at least the minimum number of nonpublic schools required for sampling, (b) that required school information data fields are correctly displayed, and (c) that current school information data fields are consistent with data from previous files. Key variables are compared across the current preliminary PSS and the most recently adjudicated PSS; cases requiring corrective action are summarized by Westat and provided to NCES.

The private school list must be augmented with newly opened nonpublic schools not appearing on the PSS. This step is critical to ensuring that the school sampling frame contains all eligible schools. Newly opened nonpublic schools are identified by inquiring of a sample of Catholic dioceses about new Catholic schools within each diocese. To evaluate the quality of these data, replies from dioceses sampled for this activity are tracked to identify nonrespondents. Reply information is reviewed to determine whether schools identified as newly opened may in fact represent existing schools that have recently been renamed. In cases in which it is difficult to determine whether a school is in fact “new,” district or state department of education Web sites are checked to assist in making a final determination. Diocese response rates and information about newly opened schools are summarized and questionable cases are identified. In cases for which corrective action is needed, follow-ups are conducted with nonresponding dioceses until either (a) a 100 percent response rate is obtained, or (b) the sampling deadline date is reached. All revisions and corrections are incorporated into PSS files as appropriate.

In 2003 PSS schools were explicitly stratified by school type (Roman Catholic, Lutheran, and Conservative Christian schools, other private schools with known

affiliation, and private school with unknown affiliation). Schools were also implicitly stratified by Census division, urbanization, and minority status (percent black, Hispanic, American Indian enrollment).

Private schools were selected with same procedure as public schools (with probability proportional to a stepped measure of size based on eligible enrollment). Private schools were sampled at three times the rate of public schools in 2003. Targets for each school type were based on participation rates from the 2002 private school sample; these targets were adjusted upwards by 5 percent in anticipation of additional sample attrition due to such factors as school ineligibility, student exclusion. The final targets used to determine school sample sizes and sampling rates are listed in the table below.

Table 7. Target private school student sample sizes, national main assessment: By grade and private school stratum, 2003

Grade	Private school stratum	National main assessment target	NAEP 2002 school yield rate	NAEP 2002 student yield rate	Attrition-adjusted target
	All private	25,800	†	†	40,252
4	Total private	12,600	†	†	18,531
	Catholic	6,300	0.90	0.94	7,820
	Lutheran	1,575	0.86	0.93	2,068
	Conservative Christian	1,575	0.68	0.93	2,615
	Other private	3,150	0.59	0.93	6,028
8	Total private	12,600	†	†	20,363
	Catholic	6,300	0.87	0.94	8,089
	Lutheran	1,575	0.83	0.94	2,120
	Conservative Christian	1,575	0.60	0.93	2,964
	Other private	3,150	0.50	0.92	7,190

† Not applicable.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

The private school sampling frame also included schools without a known affiliation. These schools were sampled in a separate stratum to make sure that the private school samples were fully representative. The target for this group was set at 25 schools for each grade.

Refusal to participate was much more common among the private school sample than it was for public schools. Substitutes for private schools that refuse to participate were assigned in a process similar to that used in State NAEP. No ineligible schools were found among the private school sample in 2003.

School Response: 2003 Main NAEP

Sampled schools eligible for assessment are recruited to participate in mathematics and reading. The target for participation, established by NCES standards, is

85 percent or greater weighted response. This rate was achieved in most cases, although in some cases substitutes were required. The following four tables detail weighted and unweighted response rates for public and private schools in the 2003 Main NAEP assessments in fourth and eighth grade.

Table 8. Public school response experience (unweighted), national main assessment: By grade and Census region, 2003

Grade	Census region	Eligible school sample	Nonresponding originally sampled schools	Responding originally sampled schools	Unweighted response rate before substitution	Recruited cooperating substitutes	Unweighted response rate after substitution
4	Total	6,971	27	6,944	99.6	1	99.6
	Northeast	1,222	5	1,217	99.6	0	99.6
	Midwest	1,798	8	1,790	99.6	0	99.6
	South	2,112	1	2,111	99.9	0	99.9
	West	1,839	13	1,826	99.3	1	99.3
8	Total	5,586	28	5,558	99.5	0	99.5
	Northeast	951	4	947	99.6	0	99.6
	Midwest	1,501	3	1,498	99.8	0	99.8
	South	1,740	9	1,731	99.5	0	99.5
	West	1,394	12	1,382	99.1	0	99.1
12	Total	118	6	112	94.9	0	94.9
	Northeast	19	0	19	100	0	100
	Midwest	33	0	33	100	0	100
	South	39	0	39	100	0	100
	West	27	6	21	77.8	0	77.8

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Table 9. Public school response experience (weighted), national main assessment: By grade and Census region, 2003

Grade	Census region	Eligible school sample weighted aggregation	Nonresponding originally sampled schools, weighted aggregation	Responding originally sampled schools, weighted aggregation	Weighted response rate before and after substitution
4	Total	3,714,988	7,356	3,707,632	99.8
	Northeast	626,412	1,488	624,924	99.8
	Midwest	824,270	244	824,025	100
	South	1,364,290	147	1,364,143	100
	West	900,016	5,477	894,540	99.4
8	Total	3,577,804	13,820	3,563,984	99.6
	Northeast	611,391	1,820	609,571	99.7
	Midwest	815,623	438	815,185	99.9
	South	1,311,067	6,810	1,304,256	99.5
	West	839,723	4,752	834,971	99.4

NOTE: Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Table 10. Private school response experience (unweighted), national main assessment: By grade and Census region, 2003

Grade	Private school stratum	Eligible school sample	Non-responding originally sampled schools	Responding originally sampled schools	Unweighted response rate before substitution	Recruited cooperating substitutes	Unweighted response rate after substitution
4	Total	696	163	533	76.6	15	78.7
	Roman Catholic	234	20	214	91.5	2	92.3
	Lutheran	100	11	89	89	1	90
	Conservative Christian	114	35	79	69.3	2	71.1
	Other private and unknown	248	97	151	60.9	10	64.9
8	Total	739	183	556	75.2	17	77.5
	Roman Catholic	252	35	217	86.1	7	88.9
	Lutheran	109	7	102	93.6	0	93.6
	Conservative Christian	118	27	91	77.1	2	78.8
	Other private and unknown	260	114	146	56.2	8	59.2
12	Total	31	13	18	58.1	1	61.3
	Roman Catholic	9	6	3	33.3	1	44.4
	Lutheran	1	1	0	0	0	0
	Conservative Christian	6	0	6	100	0	100
	Other private and unknown	15	6	9	60	0	60

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Table 11. Private school response experience (weighted), national main assessment: By grade and Census region, 2003

Grade	Private school stratum	Eligible school sample weighted aggregation	Nonresponding originally sampled schools, weighted aggregation	Responding originally sampled schools, weighted aggregation	Weighted response rate before substitution	Recruited cooperating substitutes, weighted aggregation	Weighted response rate after substitution
4	Total	398,436	85,396	313,040	78.6	5,919	80.1
	Roman Catholic	208,083	19,757	188,326	90.5	1,135	91.1
	Lutheran	22,199	2,406	19,794	89.2	265	90.4
	Conservative Christian	60,015	18,907	41,107	68.5	552	69.4
	Other private and unknown	108,139	44,326	63,813	59	3,967	62.7
8	Total	354,588	92,677	261,911	73.9	8,601	76.3
	Roman Catholic	184,173	27,449	156,724	85.1	4,504	87.5
	Lutheran	16,299	1,012	15,287	93.8	0	93.8
	Conservative Christian	47,295	12,005	35,291	74.6	546	75.8
	Other private and unknown	106,821	52,212	54,609	51.1	3,551	54.4

NOTE: Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Student sample

Schools submit to Westat lists including student names and demographic information; roughly 70 percent of schools provide this information in electronic format with the other 30 percent providing it via hard copy. Westat's Data Processing group maintains and supports the systems used for these submissions. Student lists are checked to determine whether data is complete, whether variable names and value labels are accurate, and whether potential data problems may exist. Online checks and offline reports provide feedback to school filing lists electronically. These inform schools about the progress of their student list submissions and alert them to potential data problems.

Student sampling is carried out via separate procedures specific to schools' methods of student list submission. The School Data System (SDS) sampling procedure is used to sample students attending schools that prepare hard-copy student lists. The SDS is a laptop-based software package used by field supervisors to manually enter students' demographic information, select student samples, and check the sampling results against an external data source—in this case, the CCD.

The SDS sampling algorithm is initially tested by using it to generate student samples for all sampled schools, against which Statistics staff baseline projections (based on CCD) are compared. In cases in which corrective action is needed, mismatches between the SDS sample and baseline projections are investigated and the SDS student sample algorithm is reviewed and revised accordingly. The testing process is repeated iteratively until matches between SDS samples and baseline projections are achieved.

A second procedure—E-Sampling—is used to sample students attending schools who submitted student lists electronically. The E-Sampling algorithm generates student samples for all sampled schools; these samples are then compared to both baseline sample projects and the SDS samples drawn in the procedure described above. All mismatches are investigated and the E-Sampling algorithm is reviewed and revised until matches are achieved between E-Sampling samples, SDS samples, and baseline projections.

Inclusion and Accommodation

The target population for 2003 Main NAEP included all students in public or private schools who were enrolled in the fourth or eighth grades in the 50 states and the District of Columbia. Because NAEP is intended to provide achievement estimates representative of all students in state and national populations, every effort is made to include every student capable of participating. Inclusion of students for whom regular NAEP assessments may not be appropriate has represented one of the major challenges to NAEP. Starting in 2002 NAEP required states to use the same standard rules for including students with disabilities (SWD) and English language learners (ELL); these rules were designed to lower the rate of students excluded from NAEP participation. Based on these new rules, the majority of students participating in NAEP completed assessments under standard conditions; the only exceptions to this were students with disabilities (i.e., students with an IEP developed under the *IDEA* or those with an accommodation plan under Section 504 and the *ADA*) and students identified by school personnel as having limited English proficiency (with fewer than three years of English instruction). Differential participation, whether due to exclusion or other factors such as absenteeism, could substantially impact comparability of state results. Although the procedures adopted in 2002 were designed to increase participation and improve the consistency of inclusion across states, whether these goals were accomplished remains an open question. The state-level student participation rates reported in the tables following the next paragraph vary substantially. Fourth-grade participation is generally greater than eighth-grade participation; however, differences among states—from a high of 97 percent participation of North Dakota fourth-graders (in both math and reading) to a low of 85 percent of New York eighth-graders participating in mathematics—remain substantial. It is well known that participation in assessments such as NAEP is related to student characteristics, the degree of interstate variability in participation could impact the state-by-state comparability of NAEP scores. This issue is covered in more detail in the final section of this report.

Table 12. Weighted student response rates and exclusion rates, state reading assessment:
By participating jurisdiction, 2003

Jurisdiction	Fourth grade			Eighth grade		
	Weighted student response rate	Weighted student exclusion rate: SWD students	Weighted student exclusion rate: ELL students	Weighted student response rate	Weighted student exclusion rate: SWD students	Weighted student exclusion rate: ELL students
Alabama	94.70	1.92	0.36	92.18	2.41	0.58
Alaska	93.61	2.22	1.00	90.24	2.00	0.47
Arizona	90.79	4.92	4.21	88.63	4.76	3.59
Arkansas	95.50	4.82	1.22	93.26	4.04	1.26
CA-Los Angeles	95.92	3.37	5.39	90.47	2.84	2.93
CA-San Diego	91.80	2.92	3.78	88.72	1.41	2.32
California	93.88	2.51	4.01	91.21	2.48	2.08
Connecticut	95.34	2.17	1.87	91.27	1.89	1.81
Colorado	94.74	3.52	1.46	90.97	3.15	1.03
Delaware	94.12	10.38	1.07	89.94	8.02	1.15
Florida	92.85	3.00	2.62	91.35	4.27	2.37
GA-Atlanta	94.38	1.65	0.51	92.71	3.19	1.13
Georgia	95.49	3.11	1.33	93.30	2.23	0.73
Hawaii	96.45	2.80	2.05	92.02	3.45	1.68
Idaho	95.24	2.76	1.32	92.57	3.15	0.86
IL-Chicago	92.09	5.53	5.91	93.08	5.03	3.01
Illinois	93.97	5.16	4.11	92.68	3.88	1.89
Indiana	94.44	3.58	0.38	93.16	3.26	0.76
Iowa	96.29	6.55	0.93	94.14	4.28	0.50
Kansas	95.26	2.49	1.11	93.39	2.67	1.45
Kentucky	95.59	8.26	0.50	92.75	6.83	0.47
Louisiana	95.77	5.91	0.74	92.00	5.48	0.44
Maine	93.40	6.86	0.54	92.15	4.68	0.16
Maryland	93.92	5.99	1.99	88.79	2.97	0.70
MA-Boston	94.53	4.08	5.66	93.15	4.63	7.23
Massachusetts	93.64	2.79	1.97	90.94	2.89	1.76
Michigan	94.54	6.19	1.51	90.70	5.92	0.53
Minnesota	93.67	2.63	0.95	90.44	2.83	0.81
Mississippi	94.41	5.85	0.50	92.70	4.70	0.37
Missouri	94.74	7.32	1.24	93.85	7.77	0.79
Montana	94.47	4.61	0.52	92.83	4.68	0.42
New Hampshire	94.90	4.13	1.53	93.54	3.99	1.51
New Jersey	92.86	4.85	5.02	88.25	2.26	1.90
Nebraska	93.88	3.31	0.84	91.55	2.82	0.42
Nevada	94.56	3.39	1.95	91.49	2.20	0.80
New Mexico	94.91	4.44	5.13	92.51	4.64	5.22

Continues next page

Table 12. Weighted student response rates and exclusion rates, state reading assessment: By participating jurisdiction, 2003 (Continued)

Jurisdiction	Fourth grade			Eighth grade		
	Weighted student response rate	Weighted student exclusion rate: SWD students	Weighted student exclusion rate: ELL students	Weighted student response rate	Weighted student exclusion rate: SWD students	Weighted student exclusion rate: ELL students
NY–New York City	91.72	1.87	4.94	81.04	2.04	4.12
New York	91.41	5.10	3.48	85.58	5.05	2.14
NC–Charlotte	94.73	3.56	2.63	91.70	3.24	1.31
North Carolina	95.76	6.26	2.13	93.01	6.31	1.65
North Dakota	96.61	3.67	0.73	95.17	4.39	0.42
OH–Cleveland	90.62	10.79	1.57	76.42	11.88	4.73
Ohio	92.13	5.72	0.79	90.55	5.39	0.46
Oklahoma	95.78	5.01	1.14	92.76	3.65	0.88
Oregon	94.07	6.40	3.86	90.48	3.88	2.58
Pennsylvania	95.78	3.20	0.93	92.47	2.02	0.19
Rhode Island	93.56	3.17	2.36	88.45	2.84	1.97
South Carolina	94.61	7.20	1.05	91.98	8.13	0.45
South Dakota	95.28	4.04	0.55	94.86	3.28	0.28
Tennessee	93.80	3.98	0.73	92.51	2.49	0.30
TX–Houston	93.06	9.19	19.4	90.48	6.94	6.25
Texas	95.43	7.30	5.10	92.68	6.68	3.20
Utah	94.61	3.22	2.85	91.56	2.41	1.40
Vermont	94.20	5.88	0.54	89.51	4.29	0.35
Virgin Islands	96.31	2.16	2.03	96.91	4.17	1.76
Virginia	94.81	7.72	3.49	92.47	7.68	1.91
Washington	95.24	4.26	1.66	92.12	2.78	1.45
West Virginia	94.11	9.08	0.20	92.44	8.92	0.29
Wisconsin	95.29	4.44	1.89	92.06	4.83	1.27
Wyoming	93.81	1.63	0.43	92.26	2.01	0.21
Charter Schools ¹	91.86	3.25	1.88	†	†	†
District of Columbia	94.07	4.93	1.24	88.78	6.46	1.87
DDESS ²	95.58	3.62	0.94	95.87	1.63	1.65
DoDDS ³	95.26	1.46	1.26	95.55	0.54	0.62
Puerto Rico ⁴	†	†	†	†	†	†

† Not applicable.

¹ The special charter school study was conducted only in fourth grade in NAEP 2003.

² Department of Defense Domestic Dependent Elementary and Secondary Schools.

³ Department of Defense Dependents Schools (Overseas).

⁴ Puerto Rico did not participate in the reading assessments in NAEP 2003.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Assessment.

Once school and student samples are selected, Westat delivers to PEM files containing school, grade, session, student, and shipping information. PEM uses these files to prepare preprinted Administration Schedules and to assign and track assessment booklets. Prior to delivery, the content of files prepared for PEM is compared to a master file. To determine whether transmission was successful, PEM returns the files and they are compared to the master file. If summary counts and frequencies suggest discrepancies between files sent to PEM and files received from PEM, the system is reviewed for possible programming errors. The process is repeated until returned files match those transmitted.

Weighting

NAEP weighting programs are updated annually to account for changes in state and national populations. Student weights for the National sample contained three components—a base weight, an adjustment for school nonparticipation, and an adjustment for student nonparticipation. Weights may also be scaled (poststratified) so that sums of weights for appropriate subgroup estimates are consistent with known national totals of assessable students across the nation. Weights for students sampled but excluded from assessment are estimated in a similar manner.

In addition to overall estimation weights, replicate weights—used to estimate sampling variability of NAEP estimates—are also provided for each student, excluded student, and school. Replicate weights are key to the jackknife variance procedure currently used to generate approximately unbiased estimates of sampling variance result. These weights are based on “replicate groups” created by dividing sample elements to reflect the sampling design of the assessment; the same replicate groups are used for Taylor Series alternatives to the jackknife variance procedure.

Weights are created for several assessment samples:

- State NAEP jurisdictions
- National public schools
- National private schools
- Trial Urban District Assessment (TUDA) sites
- Grade 4 students in charter schools in Calif., Texas, and Mich.

These samples are not mutually exclusive; individual students may be included in more than one sample. As such, students are given an individual weight for each sample in which they are included. These individual weights reflect several components:

- base weights reflecting school sampling
- base weights reflecting student sampling
- base weight factors reflecting assignment to reading or mathematics booklets
- adjustments for school nonresponse
- adjustments for student nonresponse
- trimming of school base weights to reduce variability
- trimming of student weights to reduce variability

Replicate weights used to estimate sampling variability of NAEP scores are also estimated for students in each assessment sample.

School Weights

Base weights are given separately by grade and reflect the nature of sampling of schools. In general these weights represent the reciprocal of the probability of school selection. However, for both new schools and substitute schools additional adjustments are incorporated to base weights. New school base weights reflect the probability of selection of districts into the new school district sample and selection of schools into the new schools sample. In many cases these joint probabilities are very small, yielding base weights of great magnitude. To reduce outlying values new school base weights are “trimmed” so that they are no larger than three times the weight that would have resulted had the school been selected from the original school sampling frame. Substitute schools—those recruited to replace schools that refuse to participate—are assigned base weights specific to the schools they replaced rather than unique to themselves. In most cases the number of students sampled within substitute schools is not equal to the number that would have been assessed in the schools they replaced; as such, student base weights for these schools are adjusted to reflect differences in size between substitute and original schools. In addition, similar to new schools base weights, substitute base weights are trimmed so that they do not exceed three times the weight that would have resulted had the school been selected from the original school sampling frame.

Student Weights

Student base weights are similar in that they reflect the inverse of the probability that a given student is selected for assessment, given that that student’s school has been selected as part of the school sample.

Nonresponse Adjustment

NAEP weights included nonresponse adjustments at both the school and the student levels. Weights of responding schools are adjusted upward to compensate for nonresponding schools; similar, responding student weights are adjusted upwards for nonresponding students.

School Nonresponse Adjustment

NAEP uses a “quasi-randomization” approach to adjust for school nonresponse. Schools are assigned to “response cells” similar to their initial sampling stratum. Public school response cells are based on the combination of the following classifications:

- Trial Urban District Assessment (TUDA) district vs. the balance of the state for states with TUDA districts
- charter school status (grade 4 only)
- urbanicity classification
- minority classification, or achievement level or median income, or grade enrollment

Private school response cells are based on the following school characteristics:

- reporting group (e.g., region, gender)
- census division stratum
- school location stratum
- minority status stratum

Within each response cell weights of responding schools are increased to represent the full set of schools originally sampled (that is, those responding and not responding). This assumes that responding schools within each response cell represent a simple random sample from the full set of responding and nonresponding schools. The degree to which bias remains even after nonresponse adjustment is a function of the homogeneity of achievement within the cell. That is, bias will remain a problem if schools that respond within a given response cell are systematically different in achievement than those that do not respond.

Nonresponse adjustments may be unstable for cases in which few schools exist within a given response cell. To avoid such instability, Westat limits cell sizes and adjustment factors in such cases by collapsing cells with few schools with cells reflecting similar characteristics. All school weights adjusted for nonresponse are compared to base weights to identify cases needing attention; discrepant cases are checked for potential bias.

NCLB now requires NAEP participation as a condition for receiving Title I funding. As such, nonresponse is less frequent for schools dependent on Title I funding. In fact, nearly every jurisdiction approached full public school participation in 2003. Because private schools remain unaffected by *NCLB*, school refusal remains a significant problem.

Student Nonresponse Adjustment

Student nonresponse adjustment procedures are similar to school adjustments—they are meant to compensate for eligible sampled students who did not participate in the assessment. They use a strategy similar to that used for schools—students are assigned to cells based on student characteristics, and weights of responders within each cell are inflated to account for those cell students that did not respond.

Cells vary for public and private schools. Public school student cells are formed within grade, jurisdiction, Trial Urban District Assessment (TUDA) district, and charter school status domains using the following structure:

- SWD and ELL status by subject (inclusion rules vary by subject) versus not SWD and not ELL
- school nonresponse cell
- age ("older" vs. "normal age" vs. "younger" student)
- gender
- eligibility for free or reduced-price lunch

Private school student nonresponse cells are based on the following characteristics:

- school nonresponse cell
- age
- gender
- race/ethnicity.

Similar to the school nonresponse adjustment, sparsely populated cells are collapsed to avoid adjustment instability. All adjusted student weights are compared to student base weights to identify discrepancies; weights identified as discrepant are checked for potential bias.

Again, similar to school nonresponse adjustment, student nonresponse adjustment relies on the assumption that, within each cell, students responding and students nonresponding are similar in achievement. The degree to which this is not the case—that

is, responders and nonresponders differ in achievement—reflects a remaining threat of bias due to differential nonresponse.

While excluded students are not included in NAEP score estimation (and therefore receive no nonresponse adjustment), weights for these students are provided so their characteristics can be analyzed.

Student Weight Trimming

Unusually large student weights may result from compounding weighting adjustments; common statistical sampling practice is to “trim” these weights in order to avoid large sampling variability in statistical estimates that might otherwise result. In 2003, NAEP student weights were reduced to no greater than 3.5 times the median of comparison group for public schools and 4.5 times the comparison median for private schools. Student weight trimming was performed within jurisdictions and private school reporting subgroups, and was carried out separately by grade, subject, and school type (public or private). In the 2003 assessment only 636 out of 740,947 students had weights that required trimming—less than 0.1 percent.

Replicate Weights

School and student sampling weights allow the estimation of statistically sound, nationally representative estimates based on the 2003 assessment results. As with any statistical estimate, these estimates are subject to some sampling variability, so it is necessary to interpret them in the context of their uncertainty due to sampling, as reflected in their estimated standard errors.

Simple cases (such as estimated means from a simple random sample) have exact formulas for estimating standard errors; the same is not true in cases characterized by stratified sampling. In these more complex cases alternative methods are needed to estimate sampling variability. Jackknife replication is currently used for this purpose. In general the jackknife process involves iterations in which the statistic of interest is estimated on selected portions of the sample; the variability of this statistic over repeated iterations is taken to reflect the statistic’s uncertainty due to sampling.

In NAEP the drawing of repeated samples is accomplished through the estimation of multiple sets of replicate weights, each of which represents a single replicate sample. Schools are assigned to one or more of 62 replicate strata. For each replicate (each of which corresponds to one of the replicate strata), a random subset of schools (or, in some cases, students within schools) is excluded; the remaining subset is reweighted to reflect this exclusion and is added to schools in the other 61 replicate strata to represent one of the 62 replicates.

The computation of replicate weights requires five steps:

1. Defining replicate strata and forming replicates
2. Computing school-level replicate weights (for noncertainty schools)
3. Adjusting school replicate weights for nonresponse and trimming
4. Computing student-level replicate weights (for certainty schools)
5. Adjusting student replicate weights for nonresponse and trimming

Schools are assigned to replicate strata separately by units representing the combination of grade (fourth and eighth), private or public status, and jurisdiction (for public schools) or affiliation (for private schools). In each case all sampled schools—including ineligible schools and those that refused to participate—are assigned to replicate strata.

Within each unit (such as a state), noncertainty schools are paired on the basis of similarity into one of a maximum of 62 replicate strata (units with odd numbers of noncertainty schools have one three-school “triplet” in addition to its paired schools).

For each pair of schools in a given replicate (each of which corresponds to one of the replicate strata), one of the two schools is excluded (i.e., its weights are set to zero), the second is weighted upward to compensate (i.e., its weights are doubled), and the remaining schools (i.e., those in other replicate strata) maintain their original weights. The statistical estimate for that replicate is calculated based on the new weights. Over repeated iterations (one for each replicate stratum) the variability of the statistical estimate is taken to represent that estimates sampling variability.

In noncertainty schools student-level weighting factors remain constant; replicate weighting takes place only at the school level. This is not the case with certainty schools; in these schools students (rather than schools) were assigned to one of up to 62 replicate strata. Students are the object of replication reweighting and student weights—rather than school weights—are adjusted during the replication process.

Both school replicate weights and student replicate weights are adjusted for nonresponse and trimmed in processes similar to base weights in order to maintain the impact of these factors.

Quality Control Procedures

Westat has well-established algorithms to check the accuracy of weighting programs. Weighting programs are run using test data that will produce known outcomes if the programs work properly. Test-generated weighting values are compared with known weighting values as a quality check; deviations are flagged for further review. Weighting programs are adjusted as appropriate and the testing process is repeated until differences fall within a specified tolerance range.

In 2003 NAEP weighting processes included several additional quality control (QC) procedures. The NAEP Web site reports the following results from these quality control checks:

Weighting

There was no evidence of any problems with the 2003 assessment weighting and adequate evidence that there were no problems with the weighting. The more simplified procedures introduced in 2003 resulted in reduced opportunities for the occurrence of problems, and greater opportunities for verifying that problems had not occurred.

External Checks of the Weighting Process

1. Comparison of the original school sample with the frame was favorable. A problem was noted with the proportion of black students enrolled in grade 8 in the national main public sample (frame-16.51 percent; sample, 17.05 percent; $p=0.0254$). All individual state p -values exceed 0.1, except in Idaho, where the difference is 0.07 percent lower in the sample than in the frame.

In connection with the school nonresponse (NR) adjustment, a problem was detected in the imputation of achievement and income data for 11 schools. Examination ascertained the problem had no effect on the final NR adjustments.

2. Comparison of characteristics from the original public school sample and the participating public school sample showed no differences, a finding which was

- ascribed to the high response rates in the participating school sample. The same comparison for private schools showed that the responding sample reported more Black students at both the fourth and eighth grades.
3. The comparison of the participating school sample to the student sample is difficult to evaluate, because there are real differences in the data, especially due to time and for the percent Hispanic students enrolled in school. Investigation of these findings were conducted; some of the subgroups that were studied to see if the differences were due to new enrollees included
 - at grade 4: Atlanta Trial Urban District Assessment (TUDA), Department of Defense overseas and domestic schools for dependents (DoD schools), Florida, Mississippi, South Dakota, and
 - at grade 8: Atlanta TUDA, Department of Defense overseas and domestic schools for dependents (DoD schools), Mississippi.
 4. Comparison of the participating student sample to the full student sample found very small differences, attributable, for the most part, to sampling error. Because of the design of the weighting process, no differences were found in the percent of students excluded.
 5. Comparison of the mathematics and reading samples found some differences, most of which were attributable to sampling error. In order to reduce clustering in future NAEP efforts, a revision in the booklet spiraling procedure was suggested.

Participation, Exclusion, and Accommodation Rates

Final rates were presented in quality control tables for each grade, subject, and jurisdiction. School rates were calculated as they had been calculated for previous assessment and also were calculated according to National Center for Education Statistics (NCES) standards.

The rates were below 85 percent for certain kinds of private schools at both grades. Rates were below 85 percent for students at grade 8 in Cleveland and New York City TUDA jurisdictions. An NR bias analysis was completed, as required by NCES standards.

Title I Data

Whereas all missing Title I data in 2002 were imputed as "no" at data entry, they were treated as "missing" in 2003. Cases of inconsistency in percentages between 2002 and 2003 were noted, but explanations as to the cause are still lacking at this time, since the true value of how much was "missing" in 2002 is unknown. Further, it is unclear as to whether "missing" was more likely to mean "yes" or "no." Large variations also existed

from state to state in the percentages. At best, extreme caution is advised in the use of Title I data as a trend variable from 2002 to 2003.

National Student Lunch Program (NSLP) Data

Some inconsistencies in these data between 2002 and 2003 have been noted. These inconsistencies appear to reflect a high degree of "status unascertained." Accordingly, it is suggested that the use of NSLP as a trend variable be limited to those cases in which the amount "not ascertained" does not exceed 10 percent in either year. (This problem was addressed for NAEP 2004.)

It further appeared that the mixing up of codes for "Free" and "Reduced-price" lunch was relatively common; this problem was also addressed for the 2004 assessment.

Race/Ethnicity Data

Within states, many changes over time were found to be attributable to sampling error. A few differences appeared to be due to school NR bias in 2002. Otherwise, no problems were detected at the state level. However, at eighth-grade, a 2 percent increase at the national level was noted in the percentage of black students. This appeared to trace back to the original school sample in both years.

The presence of strong evidence led to the suggestion to the NAEP data analysis contractor that school race or ethnicity data not be used for 14 schools, or about 0.1 percent of the sample. No single state contained more than two of these schools. The data indicated that codes were confused. This situation, incidentally, was unrelated to e-filing.

Type of Location

A few "unusual" changes in the data between 2002 and 2003 were noted but were not attributable to inconsistencies in the codes on the Common Core of Data (CCD) for the two years (i.e., such inconsistencies were found to be quite rare). Some of these changes may be related to the large changes which occur in the school frames from year-to-year, with many schools added and many dropped. Some may be due to NR bias in 2002. This problem is under further examination.

Response Rates

Public school response rates for 2003 held at a very high level; private school rates improved somewhat over the previous year, but continued to lag outside of Catholic and Lutheran private schools. Overall, student response rates remained similar to those recorded in 2002. A number of public schools in which response rate differences were noted between 2002 and 2003 were found to have been caused by school NR bias in 2002, and the state was notified of this fact in 2002.

An NR bias analysis has been undertaken for private schools at both grades, as well as for grade 8 students in two TUDA jurisdictions.

Exclusion Rates

Reading exclusions were found to be much higher than exclusions for mathematics (balanced by higher accommodation in mathematics). Nonetheless, with a few exceptions, reading exclusion was generally less than in 2002. Some exclusion outliers were noted among TUDA jurisdictions.

Final trimmed weights must be delivered to ETS for use in NAEP score estimation. Prior to delivery the content of files is compared to a master file. To determine whether file transmission was successful, ETS returns the files and they are compared to the master file. Discrepancies in summary counts and frequencies trigger a review of the system for possible programming errors; this process is repeated iteratively until returned files match those transmitted.

Evaluation

As its name implies, NAEP is designed to measure educational progress. As such, NAEP is faced with conflicting demands of maintaining comparability across time and employing an evolving and ever-improving technology of survey assessment. The ability to improve NAEP's design has been constrained even further by *NCLB*'s aggressive reporting requirements. By cutting available processing time by two-thirds (from the traditional 18 month turnaround of results to a required turnaround of six months), *NCLB* has shifted the balance of effort devoted to NAEP even more heavily toward production at the expense of technological improvement. The shift in format of NAEP technical reports from a deliverable publication to a changing online compendium has caused emphasis to shift even further from technical improvement to production. Although much of NAEP sampling and weighting is in good shape, certain questions regarding how best to carry out NAEP sampling and weighting in the changing context of NAEP remain unanswered.

The first part of this section details many of the technological strengths of NAEP's sampling and weighting procedures (most of which refer to work of Westat). Questions that remain unanswered and require additional exploration are covered in the second part of this section.

Strengths of Current Sampling and Weighting Procedures

In general Westat's sampling and weighting processes are excellent; Westat is a recognized leader in survey methodology and its leadership shows in its work with NAEP. Most of the processes used for establishing NAEP samples and weights reflect best practices in survey research and by and large little change is needed for many procedures.

Sampling frames are largely representative of the populations of interest. Westat takes major steps to compile exhaustive lists of schools from which to sample and supplement these lists with new schools. Some new schools are identified by a separate sampling procedure so special additional attention should be paid to the representativeness and completeness of the results from this procedure. Sampling frame construction will likely improve as technology for tracking schools and students improves. It's interesting to note that the frame incompleteness was one of the major

criticisms of early survey efforts (e.g., Coleman report); this is much less of an issue today.

School sampling is carried out in a quality manner. The combination of “take-all” jurisdictions, the “sparse state” option, the need for representative samples across multiple jurisdictions and strata, the augmentation of state samples to create a national sample, and the presence of many different types of schools (both eligible and ineligible) all make school sampling extremely challenging. Westat’s processes are strong in this area and the extensive checks it employs to gauge the match between sample and population characteristics go a long way toward minimizing sampling error.

Student sampling is to be commended as well. Westat has done a nice job of implementing programs that are mindful of the needs of schools yet still able to provide adequate measurement. For example, “almost-all” and similar provisions that allow entire classes to participate help to ease the logistical burden of implementing the assessment within a given school.

Finally, the calculation of sample weights is another area in which Westat has implemented well-thought-out procedures to deal with the complexity of stratified sampling. Estimating base weights for both schools and students, adjusting for nonresponse at each of these levels, trimming outlying weights to reduce estimate sampling variability, and the calculation of replicate weights are all carried out using well established methods accepted by the survey research community.

At each step of the way Westat implements several quality assurance processes to verify the accuracy of generated results. Westat’s processes are based on its many years as an industry leader in survey methodology across several sectors, of which educational assessment is only one example (representing around a quarter of its work; health care is the largest, representing 30–40 percent.). Although the NAEP program is currently faced with several challenges—many of which require additional investigation (as detailed below)—under Westat’s leadership NAEP is in good hands for evolving and improving the sampling and weighting designs to meet these challenges.

Unanswered Questions that Require Additional Exploration

NAEP has a long tradition of driving cutting edge advances in educational assessment and survey technology. Much of what our profession currently considers best practice has its roots in the national assessment. Although many of the survey and weighting procedures are in good shape, in order to maintain NAEP’s ability to provide technically sound measurement, potential changes to several aspects of the national assessment should be given thorough consideration. Some areas needing additional inquiry reflect technology that was once state-of-the-art but has now given way to improvements, whereas others are due to the changing policy context in which NAEP operates. Detailed below, the following areas merit additional exploration:

- Inclusion and accommodation for special needs students
- Accounting for school and student nonresponse and refusal to participate
- Ensuring adequacy of state samples
- Impact of repeated sampling of schools and districts across multiple assessment administrations
- Methods for estimating sampling variability of NAEP estimates

Inclusion/exclusion and accommodations for special needs students

Appropriate accommodations are expected to be provided to students who require them. Two subgroups of students are most affected by this ruling—students with disabilities (SWD) and English language learners (ELL).

On the surface new regulations would appear to lead to an increase in the overall percentage of students included in assessment as well as consistency across states in student inclusion. Greater inclusion and cross-state consistency remain a problem, however. Although rates of exclusion have dropped in recent years, a highly publicized GAO report recently revealed that the rate of SWD student exclusion in fourth grade reading had improved little—from 40 percent to 35 percent—from 2002 to 2005. Furthermore, states vary tremendously in exclusion rates—in 2005, for example, Delaware excluded more than 13 percent of its fourth-grade students sampled for the reading assessment, whereas Alabama excluded fewer than 3 percent. Exclusion rates in individual states vary over time as well—Louisiana, for example, excluded 6 percent in fourth-grade reading in 2003 and 14 percent in the same assessment in 2005. Louisiana’s much heralded rise in state-level reading achievement estimates over these two years is confounded by this dramatic change in exclusion. Finally, states not only differ in their rates of exclusion but also in the accommodations they provide to special needs students who were not excluded—so even included students may have had incomparable experiences in different states.

NCLB has brought greater attention to state-by-state comparisons, yet differential exclusion threatens such comparisons. At the same time, the instrument that gives the most information about special needs students—the SWD/ELL student questionnaire—is undergoing revision to include far fewer items. Each SWD/ELL questionnaire has traditionally asked the student’s teacher about the student and the special programs in which he or she participated; it generally took approximately three minutes to complete. The shortening of the questionnaire will limit the information available about these important subgroups just as more attention must be paid to them.

Further study must address the impact of differential exclusion and accommodation of special needs students across states. Strategies for estimating the impact of exclusion—including full population estimation work done at AIR—appear promising as ways to improve the comparability of State NAEP scores; these strategies should be further explored as well.

Accounting for school and student nonresponse and refusal to participate

School and student nonresponse and refusal to participate represent one of the most significant threats to the validity of NAEP estimates. NAEP is designed to give estimates for full populations, and samples are drawn to be representative of these populations. When subsets of these samples do not participate—whether from school refusal, student absenteeism, or parental opt-out—estimates may be biased as a result.

NCLB has raised the stakes for NAEP—as such accurate (unbiased) measurement and jurisdiction-to-jurisdiction comparability are essential. At the same time, *NCLB* has changed the context in which NAEP operates and may indirectly change the nature of student and school nonresponse in NAEP assessments:

- *NCLB* ushered in high-stakes testing at the state level with mandated tests in selected subjects in certain grades; NAEP participation adds an additional testing burden to schools for whom state test participation is already compulsory.

- Participation in NAEP is now a mandatory condition for receipt of Title I funding but remains optional for non Title I schools. As a result participation may increasingly become a function of Title I status.
- *NCLB* notification requirements increased the awareness among parents of their right to opt out of NAEP.
- Interest in 12th grade NAEP is increasing, though motivational issues, greater nonresponse, and the prevalence of dropouts all introduce additional challenges to valid measurement in 12th grade.

Interviews with Westat suggest that, while student participation is trending upward, school participation is declining. School participation reported for 2003 (the latest for which technical documentation is available) was relatively high; however, Westat memo 2006-0.0S suggests that in 2004, school participation rates after substitution were 93 percent, 88.4 percent, and 83.7 percent for public school fourth-, eighth-, and twelfth-grade samples, and 75.3 percent, 78.5 percent, and 53.2 percent for private school fourth-, eighth-, and twelfth-grade samples.

Several questions remain to be answered regarding nonresponse in NAEP:

- What is the impact of nonresponse on NAEP estimates? How does nonresponse threaten the validity and cross-state comparability of estimates?
- Two methods can be used for school nonresponse—school substitution and nonresponse adjustment. Substitution is not required if weighted response rates are at least 85 percent; however, even if this target is met the nonresponse bias could be nontrivial. What is the impact of not using substitute schools when nonresponse rates exceed 85 percent but do not reach 100 percent?
- To deal with nonresponse in the absence of substitution, Westat reassigns nonparticipant weights to demographically similar participating schools; in effect, scores for nonparticipants are imputed based on the scores of demographically similar participants. If participants and nonparticipants are not exchangeable (i.e., nonparticipants are not a random sample from the nonparticipant and participant sample) this can introduce bias into NAEP estimates. To what extent are participants and nonparticipants similar in terms of achievement and characteristics?
- What is the sensitivity of NAEP results based on the use of either of these two alternatives (substitution and nonresponse adjustment)?

Ensuring adequacy of state samples

State samples must be adequate in size and representativeness to provide reliable estimation of performance. States (and other jurisdictions) represent a smaller level of aggregation for reporting than does the nation. Estimation at the state level has traditionally required sample sizes of around 2,500 students from 100 or so schools per subject area assessment. In the current context interest does not stop at the state level; reporting is also required for historically prioritized student subgroups (such as those defined by ethnicity, lunch program status, language proficiency, and student disability). As interest shifts from absolute achievement to relative subgroup achievement it becomes even more crucial that NAEP state samples be of sufficient size to allow subgroup-level

analyses. NAEP has traditionally taken steps to oversample students in some key subgroups (e.g., by sampling schools with larger representation of blacks and Hispanics at double the rate of other schools). However, as the achievement of additional subgroups becomes greater in priority, and patterns of demographics shift within schools, additional measures should be considered to help ensure adequate samples of subgroups within states.

Today many states are seeing significant demographic changes; furthermore, demographic characteristics differ substantially from state to state. At the same time, some of the most significant data problems faced by NAEP involve missing Title I data, uncertain National Student Lunch Program data, and problems with some schools' identifications of racial/ethnic status. All of these issues can affect sampling via less accurate sampling frames and the incomparability of results over time.

Beyond sampling are problems of differential response at the subgroup within state level. Nonresponse was noted as a major issue above; however, its impact on smaller samples of students within subgroups within states/jurisdictions can be even greater. At this point it is not clear whether NAEP state samples are sufficient to support robust estimation of subgroup performance within states. The ability of state samples to provide accurate, valid estimates of subgroup performance in the face of challenges and demographic changes mentioned above should be examined in greater depth.

Impact of repeated sampling of schools and districts across multiple assessment administrations

Several schools and districts are sampled with certainty or near certainty across multiple NAEP sessions. For example, in "take-all" jurisdictions all schools are selected for the sample with certainty. As such, what appears to be a random sample in a given year may be more systematic when considered over multiple administrations. Even though the student sample in certainty schools is refreshed annually, students in these schools may share characteristics that are not shared with students in non-certainty schools. Several systematic factors may threaten the validity and comparability of results from these units. As school professionals become familiar with the NAEP assessment, scores of their students may improve in ways that may not be shared with students in districts for which NAEP is a more novel experience. On the other hand, districts repeatedly selected for NAEP participation may experience some fatigue with and resistance to the assessment, adding another potential threat to the validity of these results.

Additional analysis must estimate the impact of repeated administration in units often (or always) selected for NAEP. Furthermore, the prevalence of "certainty" schools and districts is uneven across states; the degree to which this calls into question state-by-state comparisons also needs additional study.

Methods for estimating sampling variability of NAEP estimates

As with any statistical estimate, NAEP estimates are not exact; because they are based on samples of students (rather than entire populations) they are subject to some uncertainty due to sampling variability. The estimation of uncertainty due to sampling variability is crucial to the interpretation of any statistical estimate, NAEP estimates included. NAEP should take steps to ensure that the methods employed for estimating sampling variability are the most accurate available.

The accuracy of standard errors of NAEP estimates is particularly important for several reasons:

- Analyses of achievement gaps—one of the primary areas of emphasis of recent federal accountability policy—require accurate variance and standard error estimation.
- Standard error estimates inform the state-level sample sizes needed to provide accurate estimates of high-priority subgroups (e.g., groups based on ethnicity, English proficiency, lunch program eligibility) within each state
- The use of open-ended (constructed response) items introduces uncertainty attributable to sampling of specific items as well as the assignment of individual scorers to rate assessment responses. In mathematics, more than 50 percent of student assessment time tends to be devoted to constructed-response questions; in reading, individual student assessment booklets contained an average of 9 to 13 multiple-choice questions, 8 to 10 short constructed response questions, and one (for fourth-grade) or two (for eighth-grade) extended constructed-response questions. Optimal strategies must be employed to accurately estimate the sampling variability associated with constructed response items.
- The more widespread analysis of NAEP data has led to the use of design effects rather than replication methods for taking into account variability associated with complex stratified sampling. By simply requiring an adjustment to standard errors based on simple random sampling, the use of design effects allows more general analysis of NAEP data in linear models as well as in more advanced inferential procedures. The uncertainty of design effects is rarely taken into account; however, just like any other NAEP statistic, a particular design effect has uncertainty due to sampling variability associated with it. Estimates of the uncertainty of design effects due to sampling variability should be taken into consideration when using these statistics.

The mid-1980s saw the incorporation of replication methods—specifically, the jackknife—into the estimation of sampling variability of NAEP estimates. Jackknife replication has since been used as the primary method for estimating standard errors of NAEP estimates, although Taylor Series methods based on the jackknife replication design has been used as a less computational alternative. Although the jackknife has been used as the NAEP standard procedure, the technology of statistical replication methods has advanced a great deal since the 1980s. In particular, the bootstrap has been shown in many situations to provide more efficient, more asymptotically accurate estimates to which the jackknife only approximates.

Additional attention should be given to alternatives to the traditional methods for assessing sampling variability. Westat has done some work in this area—see, for example, a 2000 paper from Brick, Morganstein, and Valliant on replication methods. This paper did not address the traditional bootstrap, however, limiting its focus only to jackknife and balanced repeated replication methods. The bootstrap has been examined in other studies (for example, the work of NCES’s Steve Kaufman presented at the JSM in the early 2000s), but these studies tend to either focus directly on Balanced Half Replication or do not take into account the adjustments that are needed when the bootstrap is used in certain situations (as described in Wiley, 2001). As such these studies are not able to provide an adequate demonstration of the bootstrap’s ability to generate more accurate estimates of sampling variability than those provided by the jackknife.

Materials Reviewed:

NCES/NAEP Documentation

NAEP Online Technical Documentation:

- Sampling (2000–02)
- Weighting (2000–03)

NAEP Contractors Statement of Work

NVS: An Agenda for NAEP Validity Research

Data Companion: NAEP 2003 Mathematics and Reading Assessments Secondary-Use Data Files

NAEP Report 83-1, “A New Design for a New Era”, Messick, Beaton, and Lord

NCES Handbook of Survey Methods, Ch. 20 (NAEP)

Steve Kaufman, NCES:

- Kaufman, S. (2001). “Using the Bootstrap in a Two-Stage Nested Complex Sample Design,” Proceedings for the Section on Survey Methods, American Statistical Association, Alexandria, Va.
- Kaufman, S. (2001). “A New Model for Estimating the Variance under Systematic Sampling,” Proceedings for the Section on Survey Methods, American Statistical Association, Alexandria, Va.
- Kaufman, S. (2000). “Using the Bootstrap to Estimate the Variance in a Very Complex Sample Design,” Proceedings for the Section on Survey Methods, American Statistical Association, Alexandria, Va.
- Kaufman, S. (1999). “Using the Bootstrap to Estimate the Variance from a Single Systematic PPS Sample,” Proceedings for the Section on Survey Methods, American Statistical Association, Alexandria, Va.

NAEP Contractor and NVS studies

RTI (primarily Jim Chromy):

- Effects of Finite Sampling Corrections on State Assessment Sample Requirements
- Federal Sample Sizes for Confirmation of State Tests in *NCLB*
- Participation Standards for 12th grade NAEP

AIR (primarily Don McLaughlin):

- Evaluation of the Precision of Estimates from the NAEP Using a Two-Dimensional Jackknife Procedure
- Evaluation of Bias Correction Methods for “Worst-Case” Selective Non-Participation in NAEP
- Properties of NAEP Full Population Estimates
- NAEP Full Population Estimates Data Files
- Participation of and Accommodations for English Language Learners.
- Participation of and Accommodations for Students with Disabilities: How to Compare NAEP and State Assessment Results (CCSSO)

ETS (primarily Jiahe Qian)

- Statistical Power Analysis and Empirical Results for NAEP Combined National and State Samples
- Analysis of NAEP Combined National and State Samples

NAEP Testing for 12th-Graders—Motivational Issues (Jere Brophy and Carole Ames)

Site Visits

AIR Site Visit (June 29, 2005)

- Interview Notes

Westat Site Visit (July 11, 2005)

- Interview Notes
- Process Memos
- NAEP 2005 Weighting Process Overview
- NAEP 2006 Frame Building and Sampling Process Overview
- Quality Control Plans and Flowcharts

Published and unpublished academic research

JEM Summer 1992 Special Issue

JES Summer 1992 Special Issue

Gene Johnson—Considerations and Techniques for the Analysis of NAEP Data, *Journal of Education Statistics*, Winter 1989, 14(4), pp. 303–334.

Wiley, E. W. (2001). *Bootstrap strategies for variance component estimation: Theoretical and empirical results*. Unpublished Doctoral Dissertation. Palo Alto, Calif.: Stanford University.

Major Steps in Sampling and Weighting

STATE NAEP: SCHOOL SAMPLING

- Establish jurisdictions
- Build public school frame within each jurisdiction
 - Start with 2000–01 schools listed in CCD
 - Add new schools
 - Small districts—during recruitment
 - Large districts—via sample survey
- Select schools within each frame
 - Some jurisdictions are “take-all”
 - Others – selected with probability proportional to a stepped measure of size (“MOS”) based on eligible enrollment:
 - 1–5 students: MOS=15.5
 - 5–20 students: MOS = 3.1*enrollment
 - 20–69 students: MOS = 62
 - >69 students: MOS=enrollment
 - Large schools can be selected multiple times
 - Stratified
 - Explicitly stratified by charter status, urbanization, minority class
 - Implicitly stratified by state-level achievement (where available by jurisdiction) or median income (where achievement data not available)
 - Final probability is scaled so each jurisdiction sample approximates target 6,510 as closely as possible
 - “Take-All” option—Available to jurisdictions with small number of schools (schools selected with certainty)
 - “Sparse State” option—Available to jurisdictions in which student populations tended to be spread over a large number of small schools
- Remove sample schools subsequently identified as ineligible
 - Schools closed or found to have zero enrollment in grade of interest
 - Special schools (ungraded schools, zero-enrollment vocational schools, special education schools, and schools serving as parts of prisons and hospitals)
- Check sample characteristics against population characteristics to gauge sampling error.

NATIONAL (MAIN) NAEP: SCHOOL SAMPLING

- Select sample for National NAEP by augmenting the aggregation of state samples with nationally-representative sample of private schools as well as public schools from jurisdictions not participating in State NAEP.
- Public School Augmentation for National NAEP
 - Jurisdiction sample targets originally established for all jurisdictions through State NAEP process (targets were developed before participation and refusal were known)
 - Recalculate probabilities of selection for each school to represent likelihood of selection as part of a national sample (rather than within each jurisdiction)
- Private School Augmentation for National NAEP

- Private School Frame from PSS
- Explicit stratification by school type
 - Roman Catholic schools,
 - Lutheran schools,
 - Conservative Christian schools
 - other private schools with known affiliation
 - private school with unknown affiliation
- Implicit stratification hierarchically by Census division, urbanization, and minority status (percent black, Hispanic, or American Indian enrollment)
- Schools within each stratum selected with same procedure as public schools (with probability proportional to a stepped measure of size based on eligible enrollment)
- No ineligible schools found in private school sample in 2003

STUDENT SAMPLING

- Assign sessions to sampled schools
 - According to eligible enrollment; most receive single session.
- Select substitute schools for schools selected
- Recruit schools to participate
- Account for school nonparticipation if necessary through substitution or nonresponse reweighting
- Establish target number of students to be sampled (up to 62 students for each time school is selected plus “Almost-All” provision that prevents assessing all but a handful of students)
 - Fourth-grade schools were allowed to have all students assessed if their enrollment was between 70 and 120.
- Assign students randomly (but evenly through spiraling) to either mathematics or reading assessment.
- Assess Students
- Determine Student Response
 - Students assessed in initial session (with or without accommodation);
 - Students assessed in makeup session;
 - Students absent from both sessions (not excluded but not assessed)
 - Withdrawn students;
 - Disabled (SWD) excluded students;
 - English Language Learner (ELL) excluded students;

WEIGHTING

- Weights created for several assessment samples:
 - State NAEP jurisdictions
 - National public schools
 - National private schools
 - Trial Urban District Assessment (TUDA) sites
 - Grade 4 students in charter schools in Calif., Texas, and Mich.
- Individual weights reflect several components:
 - Base weights reflecting school sampling (reciprocal of school selection probability, given separately by grade)
 - Originally selected schools
 - Some selected with certainty (weight = 1)

- New schools (Two components)
 - Probability of selection of their district into the new school district sample
 - Probability of selection of school into new school sample
 - Substitute schools
 - Substitutes inherit selection probability of original school
 - Student base weight adjusted to reflect difference in size between substitute and original school
 - base weights reflecting student sampling and assignment to reading or mathematics booklets
 - Adjustments for school nonresponse
 - Increase weights of schools “similar” to nonresponders
 - Assumes homogeneity of achievement across schools that respond and those that do not (within each cell)
 - Adjustments for student nonresponse
 - Increase weights of students “similar” to nonresponders
 - Assumes homogeneity of achievement across students who respond and those who do not (within each cell)
 - Trimming of school base weights to reduce variability
 - No trimming of base weights for schools originally sampled
 - New school base weights trimmed to not exceed three times the weight that would have resulted had the school been selected from the original school sampling frame
 - Substitute base weights trimmed to not exceed three times the weight that would have resulted had the school been selected from the original school sampling frame.
 - Trimming of student weights to reduce variability
 - Student weights reduced to multiple of median of comparison group
 - Multiple for public schools = 3.5
 - Multiple for private schools = 4.5
 - Performed within jurisdictions and private school reporting subgroups
 - Carried out separately by grade, subject, and school type (public or private).
 - In 2003, 636 out of 740,947 student weights required trimming

REPLICATE WEIGHTS

- Jackknife Replication
 - Define replicate strata and form replicates
 - Assign schools separately by grade (fourth and eighth), private/public status, and jurisdiction (for public schools) or affiliation (for private schools)
 - Include all sampled schools—including ineligible and nonresponders
 - Compute school-level replicate weights (for noncertainty schools)
 - Pair schools in terms of similarity
 - Iteratively exclude half of schools and double weights of other half
 - Each iteration generates a set of school replicate weights
 - Adjust school replicate weights for nonresponse and trimming

- Similar to school base weight processes
 - Compute student-level replicate weights (for certainty schools)
 - Pair students within each school
 - Iteratively exclude half of students and double weights of other half
 - Each iteration generates a set of student replicate weights
 - Adjust student replicate weights for nonresponse and trimming
 - Similar to student base weight processes
- Final sets of replicate weights represent the joint contribution of school weights and student weights
 - Noncertainty schools: Replicate weights for schools and original student weights
 - Certainty schools: Replicate weights for students with school weights equal to 1.0

QUALITY CONTROL

- Internal checks performed during the weighting process
- External (before and after) checks of the weighting process
- Review of participation and exclusion rates
- Check of individual school demographic data
- Comparisons with 2002 demographic data for public schools by state

This page intentionally left blank

Appendix G-11: NAEP State Coordinators

Site Visit Team: Chad Buckendahl and Susan Davis, Buros Center for Testing; and April Zenisky Laguilles, University of Massachusetts–Amherst
Site Visit Date: Sept. 26, 2005

Audit Summary

Staff

Rima Zobayan - NCES
Marcie Hickman - North Carolina NAEP State Coordinator
Robert Hillier – Hawaii NAEP State Coordinator
Wendy Geiger – Virginia NAEP State Coordinator
John Kennedy – Maine NAEP State Coordinator
Kathryn Sprigg – Washington NAEP State Coordinator
Barbara Smey-Richman – New Jersey NAEP State Coordinator
Dianne Chadwick – Iowa NAEP State Coordinator

As a part of the evaluation of NAEP, Chad Buckendahl and Susan Davis from the Buros Center for Testing met with seven of the NAEP state coordinators on Sept. 26, 2005, prior to the NAEP state service center prerelease meeting. The purpose of this meeting was to gain an understanding of the responsibilities of state coordinators and the types of activities they undertook to meet each of the five goals for state coordinators as defined by NCES (*Data Analysis, Reporting, Training and Professional Development, Promote the Understanding of NAEP, Coordinate the Administration of NAEP, Quality Assurance*). The comments from the state coordinators are organized below within five of the 14 audit dimensions used by Buros in its evaluation of NAEP. Unlike other site visits where the audit team went to the primary work sites, a focus group with these state coordinators was organized as part of a previously scheduled meeting to maximize efficiency in the data collection.

Organizational characteristics

The state coordinators reported a variety of backgrounds including teaching in both K–12 and secondary education settings. In addition, some coordinators served as administrators in the education field or worked in state assessment offices. One coordinator reported experience working with a testing contractor and also working as a private testing consultant. Other types of experience reported included the private sector, program evaluation, and educational research.

Panelists were asked about their communication with others involved in the NAEP system. Many coordinators reported frequent communication with their NAEP coach. The NAEP coaches serve as intermediaries between the state coordinators and the NAEP State Service Center (NSSC). The NSSC is run through a separate contract with Westat. The coaches, who are all former employees of state education agencies, serve as mentors to the coordinators and provide assistance with some of their day-to-day tasks and address any questions or problems that arise. The NAEP coaches also offer training opportunities through meetings with their coordinators. The coaches often post questions

they receive on the NAEP state service center Web site for reference by other coordinators.

The state coordinators also indicated they communicated extensively with the NSSC directly or through the NSSC with NCES staff. The coordinators indicated the NSSC was helpful in answering questions and addressing problems. In addition to the NSSC, state coordinators also receive training and guidance from NCES via the WebEx meetings and training sessions. One of the largest communication networks is among the state coordinators who frequently contact each other for questions and guidance. Coordinators also communicate via NAEP discussion boards and meet at least once a year at NAEP meetings and at professional conferences (e.g., Large Scale Assessment Conference). One problem noted by the state coordinators related to communication is with the administration field staff. Although not in all cases, many noted poor communication with field staff leading to some problems during the administration.

State coordinators are required to submit work plans twice a year to NCES detailing how they intend to meet each of their goals (*Data Analysis, Reporting, Training and Professional Development, Promote the Understanding of NAEP, Coordinate the Administration of NAEP, Quality Assurance*). These work plans are submitted in September and March and detail the work for the next year and provide a summary of their progress for the previous six months. Coordinators frame their work plans for the year around their responsibilities related to the assessments. The amount of work they plan for the year depends on the level of involvement of their state in NAEP assessment (i.e., participation in field trials, number of schools selected). In addition, the coordinators consider the goals of the state (e.g., lowering exclusion rates, integration of NAEP results in state assessment system) and the areas in which they would like to develop their skills (e.g., data analysis, exploration of alignment). The coordinators reported consulting with other state coordinators or their NAEP coach in creating their work plan. The coordinators' supervisors and NAEP coach typically review the work plans before they are submitted to NCES. Rima Zobayan provides feedback on the work plans. Throughout the year, the work plans guide the state coordinators' tasks. Several of the state coordinators reported that their progress is evaluated within by their state agency by assessing their accomplishment of the work plan. NCES began an evaluation process that included site visits with NAEP state coordinators in August 2005, approximately a month prior to this meeting.

The panelists were asked about their efforts to meet their *Training and Professional Development* goal. There appear to be two facets within this goal: coordinators seek professional development opportunities for themselves and also provide professional development opportunities for teachers and administrators within their state. To accomplish the first aspect of this goal, state coordinators reported attending training opportunities provided by NCES (e.g., the prerelease training sessions, WebEx meetings), attending the linking and scaling conference at ETS, participating in NAEP research, attending professional conferences (e.g., AERA, CCSSO), engaging in self-study (e.g., books, articles), and participating in discussion and research efforts with colleagues. To accomplish the second part of this goal, state coordinators offer workshops across the state on topics such as using NAEP resources (e.g., NAEP question took) and interpreting NAEP results.

Intended scope and use of NAEP assessments

One of the state coordinator goals is to *Promote the Intended Use of NAEP*. Several coordinators have approached this goal by trying to promote awareness of NAEP within the state. This is accomplished by educating administrators and teachers about NAEP and including a link to the NAEP Web site from the state education Web sites. This goal also includes ensuring the proper use or interpretation of NAEP results. The state coordinators noted the intended use of NAEP data and results was to evaluate progress of students in this country. The state coordinators cited several common misuses of NAEP data they had observed from various stakeholders. First, NAEP assessments are often used to compare performance across states without considering the necessary precautions before doing so. Second, many states also use NAEP data to confirm trends found in state assessment data, which may be problematic when it involves direct comparisons of achievement levels. Third, many stakeholders misinterpret change in NAEP scores, as they are unaware of the meaning of a small shift in the NAEP scale.

State coordinators reported several strategies used to discourage problematic misuses. First many of the state coordinators hold meeting throughout the year across the state within regions, counties, districts, and schools to discuss current NAEP activities (e.g., what tests are going to be given or reported that year) and familiarize individuals with NAEP tools and resources. Such meetings are also held at universities with preservice teachers. Second, coordinators stay in continual contact with school administrators via newsletters, e-mail, and phone calls to keep them up to date on NAEP activities. This also serves to familiarize stakeholders with their State NAEP coordinator in case they have any questions on how to interpret NAEP data. Third, the NAEP state coordinator and public information officer monitor the press after a NAEP release as many reports within their state include misinterpretation of NAEP results. By closely monitoring what is being reported about NAEP, the coordinators can refute incorrect interpretations and be prepared to address questions related to these interpretations.

Administer the assessment

State coordinators are responsible for several activities during the NAEP administration as a part of the *Coordinate the Administration of NAEP* goal. The amount of time required by this activity depends on several factors (e.g., if the state was selected to participate in a pilot study, how many schools in their state were selected to participate in NAEP, the type or number of assessments being conducted that year, and if there is a state mandate for NAEP participation). Some states have legislation requiring participation in NAEP for any school that is selected, however, this is inconsistent across states. Without such legislation to assist the process, the NAEP state coordinator must spend time recruiting schools which may involve several forms of personal communication (e.g., letters, phone calls, visits) which can be quite extensive. After recruitment, state coordinators are responsible for entering information about participating schools into the school control system. Coordinators expressed frustration with this system because the information cannot be uploaded electronically. As the administration date approaches, state coordinators commonly serve as a liaison between schools and the NAEP field staff in making preparations. During the day of administration, state coordinators often observe as many administrations as possible and try to intervene with any administration problems.

The state coordinators noted several problems with the administration of NAEP. First, some of the coordinators suggested that there were not enough field staff available during the administration to help with things such as accommodations for special needs students [Note: This concern may be particularly related to years when there are larger samples needed]. Some of the state coordinators indicated that many of the field staff in some states were unprepared and quit (in some cases a third) during the administration. They speculated this was due to poor recruitment, low pay, and unrealistic workloads. The second problem noted was that the NAEP questionnaires for students with disabilities were too long and required extensive time to complete. In addition, many school assessment coordinators were faced with reviewing the individualized education programs (IEPs) and related forms for students with disabilities (SD) and all English language learners (ELL) for NAEP assessments.

Write, review, issue, and disseminate reports and data

One of the state coordinator goals is *Data Analysis*; however, their responsibilities here are not related to the operations of NAEP, but rather analyses that relate to the dissemination of information. Many of the state coordinators complete the *Data Analysis* goal by reformatting NAEP reports to make them understandable by stakeholders within their state. These reports are designed to highlight findings and data that are important to the state. In addition, several state coordinators reported conducting specific types of analyses such as strand analysis, sub-group exploration, gap analysis, and trend analysis.

NAEP reports are typically provided without interpretation or opinion and the state coordinators are commonly asked by stakeholders within their state to provide meaning of the NAEP results. States want to know the worth of the data to schools and educators. State coordinators mentioned this being a very interesting aspect of their job; however, some do not often have adequate time to address this goal. Several state coordinators reported addressing this goal by developing special reports to be shared at conferences around the state.

Renew and improve the assessment

To meet the goal of *Quality Assurance*, the state coordinators report conducting several types of evaluations including checking data, observing administrators, and monitoring the assessment process. Following each administration, the coordinators participate in a WebEx with NCES where they can report any problems that occurred during the administration.

Findings and Recommendations

Overall, we would like to commend the NAEP state coordinators for the work they are doing in their states. As indicated in this summary, state coordinators are serving a variety of functions including: serving as an information center in their state for NAEP, promoting the understanding of NAEP throughout their state, and finding meaning in NAEP results for stakeholders within their state. To accomplish these tasks the state coordinators are provided with training and guidance from a support network including the NAEP coaches, the NSSC, and NCES. These organizations providing service to the NAEP state coordinators appear to be addressing the needs of the state coordinators

through personal communication, discussion boards, training sessions, and regular meetings.

Based on our observations we would also like to offer a few recommendations. First, we recommend additional administrative support for the state coordinators. Several of the state coordinators felt they had inadequate time to address some of the more important goals (e.g., data analysis, reporting) during years when their responsibilities included a greater amount of clerical work (e.g., data entry). The state coordinators felt their skills and abilities were not being maximized with this work taking up so much time. This support would afford the coordinators time to focus on providing services that are more consistent with the skill set (e.g., data analysis, reporting, communication) for which they were hired.

Second, we recommend additional preparation of the field administration staff prior to operational administration. This additional preparation would include advance meetings with the state coordinator to help the staff understand any contextual information that may be necessary in a given state. The state coordinators expressed frustration with their relations with the NAEP administration field staff—there was a lack of communication with the field staff and many apparently quit during the administration. By bringing the state coordinators into the planning process earlier, there is an opportunity to proactively address questions that might otherwise arise during the course of administration. This additional time with the field staff will allow the state coordinators to ask questions and take care of their organizational responsibilities in advance of the administration day.

Our third recommendation concerns the evaluation of work conducted by the state coordinators. The funding for these positions comes from the U.S. Department of Education through the state education agencies (SEAs). The coordinators submit proposed updates on progress and proposed work plans to NCEES (the COR). However, the SEA conducts the direct oversight of the coordinator's work. Our concern is a possible disconnect between the coordinators serving as agents for the NAEP program under the supervision of individuals NOT involved in the NAEP program. For example, many of the coordinators indicated they spent time and effort preparing reports of NAEP results that would be useful for different constituencies across their state. Because the work products, such as reports like these, are not reviewed before dissemination, it is unclear if the coordinators are conveying the NAEP results in a manner that matches the intended uses of NAEP data. Therefore, our recommendation is that there be a structured evaluation program by which the work products of coordinators are reviewed and evaluated by someone involved in the NAEP system. Ultimately this responsibility would likely fall to ED as it is the primary contractor for the state coordinators.

Fourth, we recommend a training curriculum for the state coordinators. The current strategy for professional development includes only minimal short-term structure (e.g., state service center provided a summer curriculum of training to prepare for release of data). Although state coordinators may require different levels of training given the needs within their states, some common elements will help to ensure equitable service. The current training opportunities appear to be available for state coordinators who choose to participate. A more structured program of training would ensure equitable skills in areas important to the coordinators accomplishing the NCEES goals.

This page intentionally left blank

Appendix G-12: Hager Sharp

Buros Reviewer: Brett Foley, Buros Center for Testing
 Dates of material review: June–August 2005

Audit Summary

Materials submitted by:

Siobhan Mueller and Debra Silimeo

As the utility studies within the evaluation of NAEP are exploring the reporting aspect of NAEP in depth, our review of the work conducted by Hager Sharp was limited to a review of materials submitted by the organization. The audit dimension identified for Hager Sharp is “Write, review, issue, disseminate reports and data.” See the utility study reports for a more in-depth review of NAEP reporting.

Write, review, issue, disseminate reports and data

Audiences

According to a NAEP Dissemination and Outreach, and Meeting Logistics Support report (2003 (b)), the primary audiences for targeted for outreach and dissemination of major NAEP reports include the media, NAEP state coordinators, state education officials, local education officials, parents, national policymakers, state policy makers, and education organizations and associations.

Evaluating Stakeholder Appropriateness/Utility

According to a working proposal submitted by Debra Silimeo (2004), Hager Sharp planned to conduct

...Market research, also known as a “Customer Research Agenda,” [that] will consist of focus groups and interviews with NAEP users to learn about how they are using the data, the usefulness of the reports and other information that they believe would be valuable for them to receive. These findings will also help determine how to make the best use of briefings and workshops (p. 1).

This Customer Research Agenda will focus on distributing materials to the media, parents and the general public, associations and education groups, state coordinators, and teachers, principals, and school administrators. To ensure the successful releases of reports Hager Sharp plans to develop and distribute materials to address communication challenges, develop a customer research agenda, expand NCES’ research agenda, explore ways to better assist state and district efforts, develop and communicate a formal data distribution plan, consider release logistics and timing, and explore expanded use of technology (Silimeo, 2004, p. 1).

Distribution to Appropriate Audiences: Major Reports

Hager Sharp identified a number of outlets by which it would distribute NAEP reports. First is the State Service Center which will be used to distribute material to NAEP state coordinators as well as educators and policymakers. Second are educational organizations to reach school officials. Third is through online mediums such as the NAEP Web site, e-mail, and

listservs. Fourth is through the use of Media Build, which affords distribution to targeted media outlets. Finally, Hager Sharp will distribute material to the Educational Writers Association for distribution through educational publications (NAEP Dissemination and Outreach, and Meeting Logistics Support report, 2002).

Distribution to Appropriate Audiences: Special Studies and Secondary Analyses

Hager Sharp included in its materials specific plans for distributing materials related to special studies and secondary analyses. It noted:

Each release will be evaluated for its newsworthiness and appropriate audiences, and Hager Sharp will employ effective outreach and dissemination tactics. “Including Special Needs Students” will serve as an example of the methods we could use. (NAEP Dissemination and Outreach, and Meeting Logistics Support report (2003 (a), p.10).

The report referenced above (Including Special Needs Students) was used in as a model by which Hager Sharp could assess the users of this type of information. Specifically, it identified eight groups of consumers of this information: research and academic groups, policymakers, educators, the testing community, disabilities groups, multicultural groups, medial, and NAEP state coordinators. For each group, it noted specific means by which this type of information will be distributed to interested parties.

Procedures for Timely Reporting of Results

Given the shortened preparation time for many NAEP reports, Hager Sharp outlined specific steps it will use in preparing for the release of each NAEP report. The timeline below was documented in the NAEP Dissemination and Outreach, and Meeting Logistics Support report (2003 (b)):

3 Months before release

- Create media lists, update and incorporate new media where necessary (ongoing)
- Begin development of press kit contents (Speaker bios, FAQs, Fact Sheets)
- Begin preparation for CD-ROM press kits
- Develop list of press conference speakers
- Select and secure venue, conduct walk-through

Two Months before release

- Draft and approve briefings invitations for press, associations and content groups
- Approve list of speakers, issue invite to speakers
- Develop press conference agenda.
- Draft press release announcing results
- Continue preparing press kit materials
- Draft press release template for state coordinators
- Pre release workshop for state coordinators.
- Plan/Begin production of VNF

One Month before release

- Editorial board meetings in cities participating in trial urban assessment
- Arrange for Web chat

- Finalize press conference agenda and speakers
- Finalize pre-briefing agendas, send invitations for press pre-briefings
- Draft, review and approve media alert for press conference
- Review and approve press release
- Approve press kit components, begin production of kits, stuff kits
- Produce PowerPoint presentation and pre-event briefing materials
- Pitch key media.
- Data briefing to NAGB
- Data briefing to secretary of education
- Send media alert to daybooks.
- Final editing of VNF

Three to five days before release

- Conduct prerelease briefings with press, education associations and content groups, Hill staffers
- Venue walk-through
- Pitch story to media

Day of Release

Media

- Send out press release
- Conduct proactive story pitching
- Post all media materials to Web site
- Serve as media liaison, manage “day of” media inquiries and interviews
- Distribute VNF

Press Conference

- Staff for sign-in table
- Display of NCES and Report banners
- Production of Web chat
- Management of press conference site technical requirements.

Briefings

- Briefing to Hill
- Briefing to governors’ aides
- Briefing to education associations and organizations
- Briefing to content groups

Web chat

- Produce Web chat.
- Support Web chat speaker as necessary with information to answer questions from participants

Post-Release Follow-up

- Follow-up media outreach, respond to and manage media inquiries
- Media monitoring
- Media reporting and analysis
- Final media report
- Additional briefings (pp. 22–23)

Appropriate use of data

Hager Sharp conducts a post-release follow-up in which it monitors the quantity and accuracy of media reporting and analysis. It uses evidence of misrepresentations of NAEP (or how NAEP scores are used) to better understand the public perceptions of NAEP and how well its education efforts are working.

Materials reviewed:

Hager Sharp submitted two “Deliverables CDs.” These CDs contained various files relating to tasks performed as well as monthly and annual reports. The reports utilized to construct this summary are listed below:

NAEP Dissemination and Outreach, and Meeting Logistics Support. (2002). *Strategic plan for Part A, NAEP Dissemination and Outreach Task 2* (Deliverables CD 2, Contract No. ED-02-PO-2738) [CD-ROM]. Washington, D.C.: Hager Sharp Inc.

NAEP Dissemination and Outreach, and Meeting Logistics Support. (2003 (1)). *Final strategic plan for Part A, NAEP Dissemination and Outreach Task 1, Dissemination of Special Studies and Secondary Analyses* (Deliverables CD 1, Contract No. ED-02-PO-2738) [CD-ROM]. Washington, D.C.: Hager Sharp Inc.

NAEP Dissemination and Outreach, and Meeting Logistics Support. (2003 (2)). *Final strategic plan for Part A, NAEP Dissemination and Outreach Task 1, Release of Major Reports* (Deliverables CD 1, Contract No. ED-02-PO-2738) [CD-ROM]. Washington, D.C.: Hager Sharp Inc.

National Assessment of Educational Progress Reading and Mathematics 2003 National, State, and TUDA Releases. (2004). *Media Coverage Debriefing and Data Release Activity Analysis* (Deliverables CD 1, Contract No. ED-02-PO-2738) [CD-ROM]. Washington, D.C.: Hager Sharp Inc.

Silimeo, D. (2004). *Working proposal customer research agenda* (Deliverables CD 2, Contract No. ED-02-PO-2738) [CD-ROM]. Washington, D.C.: Hager Sharp Inc.

Chapter 2:
**Evaluation of the Standard Setting on the 2005 Grade 12 National Assessment of
Educational Progress Mathematics Test**

Stephen G. Sireci, Jeffrey Hauger, Christine Lewis,
Craig Wells, April L. Zenisky, and Jill Delton
Center for Educational Assessment
University of Massachusetts, Amherst

This page intentionally left blank

Contents

List of Figures and Tables.....	2-v
Abstract.....	2-vii
Introduction.....	2-1
The 2005 Grade 12 NAEP Mathematics Assessment.....	2-2
A New Method for Setting Standards on the Grade 12 Math Assessment.....	2-3
The Current Evaluation.....	2-3
A Brief Description of Standard Setting.....	2-4
Standards for Standard Setting.....	2-5
Method.....	2-11
Meetings Attended.....	2-11
Data Analyzed.....	2-12
Description of the Mapmark Method.....	2-13
Evaluation Criteria.....	2-19
Results.....	2-23
Procedural Evidence.....	2-23
Internal Evidence.....	2-34
External Evidence.....	2-45
Summary and Conclusions.....	2-53
Criticism of Earlier NAEP Standard Setting and Its Relevance to the Current Study.....	2-55
Limitations of the Mapmark Method.....	2-58
Limitations of the Evaluation.....	2-58
Recommendations.....	2-59
References.....	2-61

This page intentionally left blank

Figures and Tables

Figures

Figure 1: Example of Domain Characteristic Curve Information Provided to Panelists.....	2-16
Figure 2: Sample “Primary Item Map” Illustrating NAEP Subscales.....	2-17
Figure 3: Sample Percent Correct Table.....	2-19
Figure 4: Item Density for 2005 Grade 12 NAEP Mathematics Items.....	2-27
Figure 5: Sample Domain Score Chart.....	2-30
Figure 6: Variability in Panelists’ Median Cut Scores Across Rounds and Groups.....	2-39
Figure 7: Variability in Panelists’ Median Cut Scores Across Rounds and Tables.....	2-40
Figure 8: Variability in Panelists’ Median Cut Scores Across Rounds and Ethnic Groups.....	2-41
Figure 9: Variability in Panelists’ Median Cut Scores Across Rounds by Panelist Type.....	2-42
Figure 10: Variability in Panelists’ Median Cut Scores Across Rounds by Sex of.....	2-43
Figure 11: Variability in Panelists’ Cut Scores Across Rounds by Geographic Region.....	2-44

Tables

Table 1: NAEP Achievement Level Descriptions (Generic).....	2-1
Table 2: Content Area Weights on the 12th-grade NAEP Mathematics Assessments: 1990 and 2005.....	2-3
Table 3: Excerpts from the <i>Standards for Educational and Psychological Testing</i> Relevant to Standard Setting.....	2-6
Table 4: Standard Setting Meetings Observed by University of Massachusetts Staff.....	2-12
Table 5: Major ACT Reports Related to 2005 Grade 12 Mathematics Standard Setting.....	2-12
Table 6: Sample Feedback Using Teacher Domains.....	2-15
Table 7: Summary of Criteria for Evaluating 2005 Grade 12 NAEP Math Standard Setting.....	2-21
Table 8: Recommended Criteria for Selecting NAEP Standard Settings Panelists.....	2-23
Table 9: NAGB’s Recommended Criteria for NAEP Standard Setting Panels.....	2-24
Table 10: Summary of Panelists’ Round 4 Survey Data.....	2-31
Table 11: Summary of Panelists’ Exit Survey Data: Sufficient Time.....	2-32
Table 12: Summary of Panelists’ Exit Survey Data: Confidence in Process.....	2-33
Table 13: Average Cut Scores Across Rounds.....	2-34
Table 14: Standard Errors for Mean Cut.....	2-35
Table 15: Statistical Comparison of Operational and Pilot Study Final Cut Scores.....	2-36
Table 16: Summary of Estimates of Standard Errors of (Round 4) Cut Scores.....	2-37
Table 17: Sample Sizes for Panelist Subgroups.....	2-37
Table 18: Summary of Consistency of Cut Scores Across Item Rating and Mapmark Methods.....	2-45
Table 19: 2005 Grade 12 NAEP Math Achievement Level Descriptions and Results.....	2-47
Table 20: AP Calculus Exam Results for 2005 High School Seniors.....	2-49
Table 21: 2005 ACT Results: High School Seniors.....	2-50
Table 22: Comparison of Achievement Level Results: 2004 Pilot and 2005 Assessment.....	2-52
Table 23: Summary of Results Regarding Evaluation Criteria.....	2-54

This page intentionally left blank

Abstract

This report represents an independent evaluation of the process used to set achievement level standards on the 2005 Grade 12 NAEP Math test. The data used in this evaluation included observations of the standard setting meeting, observations of advisory committee meetings in which the results were discussed, review of documentation associated with the standard setting study, analysis of the standard setting data, and analysis of other data related to the mathematics proficiency of 2005 Grade 12 students. The evaluation framework used criteria for evaluating standards contained in the *Standards for Educational and Psychological Testing* (AERA et al., 1999) and other suggestions from the literature (e.g., Kane, 1994, 2001). The process was found to have adequate procedural and internal evidence of validity. Using external data to evaluate the standards provided more equivocal results. In considering all evidence and data reviewed, we concluded the process used to set achievement level standards on the 2005 Grade 12 NAEP Math test was sound and the standards set are valid for the purpose of reporting achievement level results on this test.

This page intentionally left blank

Introduction

Since 1990, one of the primary means by which the results from the National Assessment of Educational Progress (NAEP) are reported is in terms of the estimated percentages of our nation's students who fall into different achievement level categories. For all NAEP assessments, three achievement levels are defined: Basic, Proficient, and Advanced. To establish these achievement levels, cut scores must be set on NAEP exams. The process of setting cut scores on tests is called *standard setting*, which is one of the most difficult and controversial activities in educational testing (Cizek, 2001a). The degree to which these cut scores are appropriately set is one of the most critical validity issues associated with NAEP, because the inferences that are made from these results have important consequences for how the academic achievement of our nation's students is interpreted.

The achievement levels on NAEP exams are established by the National Assessment Governing Board (NAGB).¹¹ NAGB establishes both generic achievement level descriptors that cut across all NAEP exams as well as specific descriptions of what students at different achievement levels are expected to know and do in each subject area in grades 4, 8, and 12. NAGB describes the generic achievement level descriptors as representing "an informed judgment of 'how good is good enough' on NAEP....The three levels are used as the primary means of reporting what students should know and be able to do on the National Assessment."¹² The specific definitions of each achievement level are presented in Table 1.

NAEP achievement level results are reported for the nation, for states, and for subgroups of students defined by sex, ethnicity, socioeconomic status, and other important demographic variables. The validity of these achievement level results is critical because their intent is to describe the proficiencies of our nation's students with respect to well-defined categories of performance.

Table 1. NAEP Achievement Level Descriptions (Generic)

Achievement Level	Description
Basic	This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
Proficient	This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.
Advanced	This level signifies superior performance.

Source: National Assessment Governing Board (2007). Downloaded from <http://www.nagb.org/> on Feb. 14, 2007.

Setting standards on NAEP has been controversial since the idea was originally proposed (Vinovskis, 1998). It has been criticized both on logistical grounds and with respect to its technical defensibility (e.g., Linn, Koretz, Baker, and Burstein, 1991; Stufflebeam, Jaeger, and Scriven, 1991; U.S. General Accounting Office, 1993), but it has also been staunchly defended

¹¹ For the history of NAGB's development of NAEP achievement levels, see Vinovskis (1998).

¹² National Assessment Governing Board (2007). Downloaded from the World Wide Web from <http://www.nagb.org/> on Feb. 14, 2007.

(Hambleton et al., 2000). At this juncture, one thing is clear—NAEP achievement level results are one of the most widely used and interpreted indicators of the academic achievement of U.S. students in grades 4, 8, and 12 (Jaeger, 2003; Zenisky, Hambleton, and Sireci, 2007).

This report focuses on a recent and important activity related to setting achievement levels on a NAEP exam—the setting of standards on the 2005 Grade 12 NAEP Mathematics Test. The method used to set the achievement level standards on this assessment was a new method, established in part to address criticisms of how NAGB set these standards in the past. In this report, we evaluate the process used to set the achievement level standards on this assessment, drawing from the psychometric literature regarding technical and quality control issues in setting and evaluating standards on educational tests. We do not, however, address the policy issue of whether standards *should* be set on NAEP assessments. Rather, we comprehensively evaluate the specific process NAGB used to set the achievement level standards on the 2005 Grade 12 NAEP Mathematics assessment.

The 2005 Grade 12 NAEP Mathematics Assessment

During the 1990s, NAGB’s Assessment Development Committee worked on revising the NAEP mathematics frameworks (test specifications). Through public meetings and recommendations of the Assessment Development Committee, NAGB decided to make minimal revisions to the fourth and eighth grade assessments, but to substantially revise the framework for the twelfth grade mathematics test. This revision was motivated by two factors: (a) a desire to reflect the three-year mathematical curriculum common in many high schools, and (b) a curriculum study that found 79 percent of 12th grade students take two years of Algebra and one year of Geometry (National Assessment Governing Board, 2004). Table 2 illustrates the content weights for the Grade 12 NAEP mathematics assessment in 1990 (which were also used in the last assessment in 2000) alongside the revised specifications that went into effect for the 2005 assessment. The proportion of test content devoted to Numbers and Operations decreased in 2005, whereas the content area Data Analysis, Statistics, and Probability increased, as did the content area Algebra. Also in 2005, Measurement and Geometry were merged into one subscale, and the proportion of content devoted to those areas decreased.

A few other modifications to the 12th grade NAEP assessment were also implemented. First, students were able to use their own calculators on items that required them instead of a standard calculator NAEP provided. Second, the length of test time per block was increased from 15 to 25 minutes, which may alter the difficulty of certain items. Third, new items were created to cover new content. Although some content overlap between certain areas still existed, NAGB decided to break the long-term trend line for the grade 12 math assessment because the creation of new items and the rearrangement of item blocks made the assessment too different from those in the past. Therefore, it was decided to establish a new trend line for 12th grade mathematics.

Table 2. Content Weights, 12th Grade NAEP Mathematics Assessment:1990 and 2005

Content Area	Weight	
	1990	2005
Numbers and Operations	.20	.10
Measurement	.15	.30
Geometry	.20	
Data Analysis, Statistics, and Probability	.20	.25
Algebra	.25	.35
Total	1.00	1.00

A New Method for Setting Standards on the Grade 12 Math Assessment

Given the significant changes to the 12th grade math assessment, new standards needed to be established for the assessment. Prior to this time, the standard setting procedure used for all NAEP assessments was a modification of the Angoff method (Loomis and Bourque, 2001). In the summer of 2004, NAGB awarded the contract for setting the achievement level cut scores on the 2005 Grade 12 Mathematics assessment to ACT. Due to criticisms of the previous standard setting methodology, ACT decided to explore an alternate methodology known as the Mapmark methodology (ACT, 2005a; 2005b; 2005c; 2005d). Before implementing this new methodology, NAGB first asked for work that assessed the impact of using the Mapmark method to set the achievement level standards. It was suggested that this new method be compared to the Angoff method (also known as the “item rating” method in the context of NAEP) using the eighth grade math assessment. ACT conducted several pilot studies (described later) to evaluate the use of Mapmark and compared it to the Angoff method for setting standards on NAEP assessments.

The results of the pilot studies indicated that the two methods were comparable with respect to results and defensibility (ACT, 2005b, 2005c). Following deliberation by ACT’s Technical Advisory Committee for Standard Setting and NAGB’s Committee on Study Design and Methodology, the NAGB Board voted to implement the Mapmark procedure to set the Achievement Levels for the 2005 Grade 12 Mathematics assessment, presumably to address criticisms of previous standard-setting studies (e.g., Pellegrino, Jones, and Mitchell, 1999).

Implementation of the New Standard-Setting Procedure

A comprehensive standard-setting study, using the Mapmark method, was carried out in November 2004 to set the new standards on the 2005 Grade 12 NAEP Mathematics assessment. This standard-setting study was commissioned by NAGB and implemented by their contractor ACT. In the remainder of this report, we describe the new method and evaluate this standard-setting activity using both observational procedures and analysis of the data gathered during the study.

The Current Evaluation

The purpose of our evaluation is to critically evaluate the standard-setting processes on the 2005 Grade 12 NAEP Mathematics assessment to determine whether the standards are reasonable and defensible. Our evaluation criteria rely heavily on Kane’s (1994, 2001) framework for validating and evaluating standard-setting studies (i.e., procedural evidence, internal evidence, external evidence) as well as on guidelines provided by the *Standards for*

Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education, 1999; hereafter referred to as the *Standards*). Our evaluation includes a review of all documentation related to standard-setting, observations of the standard-setting itself and of the discussion of the results at various NAGB and ACT committee meetings, and reanalysis of the data gathered from standard-setting panelists.

Terminology

The nomenclature used in standard-setting can be a bit confusing and so we define some important terms before proceeding further. *Achievement levels* refer to the score reporting categories used on NAEP assessments that describe “what students should know and be able to do.” As described earlier, there are three achievement levels—Basic, Proficient, and Advanced. Students who are not considered at or above Basic fall into a fourth, unofficial and undefined category referred to as “Below Basic.” The specific scores on the NAEP score scales that are used to distinguish between these achievement levels are called *cut scores*. These cut scores represent *standards* of student performance on a NAEP test that are thought to characterize the threshold performance for each achievement level category. A *standard-setting study* is the study used to determine or recommend the cut scores to be used on a particular NAEP test to distinguish between the achievement levels. Thus, “recommended” cut scores are the end products of a standard setting study. A *standard-setting method* is the specific process used to determine performance standards on a particular exam. In some cases, the method may be a combination of different methods that are often used alone.

A Brief Description of Standard Setting

Standard-setting is the process of dividing a continuous variable, such as a test score scale, into a discrete variable with two or more categories (sometimes referred to as performance or achievement levels). The demarcations between these categories are characterized by *cut scores*, which are points along the score scale continuum that divide one category from another. Setting cut scores on a continuous score scale may lead to loss of information (because there are fewer score categories to differentiate examinees), but provides categories that may be more meaningful and understandable to policymakers and others who are unfamiliar with (or confused by) scale scores. Kane (2001) acknowledged that the standard-setting process results in an ordinal scale superimposed onto what is typically a continuous test score scale:

The adoption of cut scores to assign examinees to performance levels introduces a new, ordinal scale of performance levels, and thereby adds a new layer to the existing interpretation. The use of an ordered set of performance levels with evaluative labels clearly suggests that there are substantial differences between the performance levels. Examinees who are assigned to a particular performance level based on their score are assumed to have met the general requirement for that level. (p. 54).

Many different methods exist for setting cut scores (or standards) on educational tests (see Cizek, 1996a and 2001b, for descriptions of a variety of these methods). However, all methods are inherently subjective because there is no “true” standard to discover—that is, the optimal cut score is not simply a parameter to be estimated. Hence, setting standards on educational tests is essentially the establishment of a policy, albeit one that is informed by data. These data are typically in the form of judgments from subject matter experts (standard-setting panelists)

regarding the probability that examinees who score near the desired achievement levels will have success on specific items.

The subjectivity of standard setting is frustrating for the primarily quantitative field of psychometrics. Cizek (2001a) stated “standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other” (p.5). McGinty (2005) also acknowledged the subjectivity in setting standards, but he emphasized the need to better understand standard-setting studies and how to evaluate them:

As frustrating as these concerns may be, they are understandable when standard setting is recognized for what it is: an exercise in human judgment, elusive and fraught with subjectivity, characterized by many features that are not amenable to psychometric analysis. Nevertheless, the high-stakes nature of standard setting makes it imperative that researchers forge onward toward improved ways of evaluating the quality of standard setting judgments. (p. 270)

Calls like McGinty’s are one reason why there has been a great deal of research on setting standards on NAEP exams (e.g., ACT, 1995, 2005b; Hambleton et al., 2000; Loomis and Bourque, 2001).

Standards for Standard Setting

The most recent version of the *Standards* (AERA et al., 1999) noted the increasing importance of standard setting by incorporating additional standards and guidance related to setting cut scores. For example, the *Standards* state, “... In some situations the validity of test interpretations may hinge on the cut scores” (p. 53). They also pointed out “Cut scores embody value judgments as well as technical and empirical considerations” (p. 54).

The *Standards* define validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). When standards are set on tests, the evidence and theory used to defend the appropriateness of the cut scores are critical for evaluating the validity of interpretations based on test scores. For this reason, the *Standards* provide several guidelines for conducting and evaluating standard-setting studies. The guidelines that are most relevant to standard setting are presented in Table 3. The standards (guidelines) are presented alongside abbreviated comments, also taken from the *Standards*.

A review of these specific standards and their associated comments emphasizes the importance of (a) having a strong rationale for the standard-setting method used, (b) selecting appropriate standard-setting panelists, (c) ensuring panelists understand their tasks and are competent to perform them, (d) implementing the standard-setting method appropriately, and (e) documenting the entire process. In addition, the *Standards* call for estimates of the reliability of classification decisions such as conditional standard errors of measurement around cut scores and estimates of decision consistency. Computation of these estimates is separate from the process of standard setting, but they are important for evaluating the validity of the cut scores.

The guidance provided in the *Standards* reflects the best practices in standard setting found throughout the literature (see for example, Cizek, 1996b, 2001b; Cizek, Bunch, and Koons, 2004; Hambleton, 2001; Hambleton and Powell, 1990; Jaeger, 1990; Kane 1994, 2001; and Meara, Hambleton, and Sireci, 2001). Kane (1994, 2001) provided a comprehensive discussion of the difficulty in validating cut scores as well as a framework for evaluating them. This framework is congruent with the spirit of and specific guidelines suggested in the *Standards*. We turn now to a description of this framework, which we used to evaluate the 2005 Grade 12 NAEP Mathematics test.

Table 3. Excerpts from the *Standards for Educational and Psychological Testing* (AERA et al., 1999) Relevant to Standard Setting

Standard	Selected Comments from <i>Standards</i>
<p>1.7: When validation rests in part on the opinions or decisions of expert judges... procedures for selecting such experts and for eliciting judgments...should be fully described. The qualifications, and experience, of the judges should be presented. The description of the procedures should include any training and instructions provided...indicate whether participants reached their decisions independently, and...report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth. (p. 18)</p>	<p>Systematic collection of judgments may occur...in formulating rules for test score interpretation (e.g., in setting cut scores)...Whenever such procedures are employed, the quality of the resulting judgments is important to the validation. It may be entirely appropriate to have experts work together to reach consensus, but it would not then be appropriate to treat their respective judgments as statistically independent. (p. 19)</p>
<p>2.14: ...Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.” (p. 35)</p>	
<p>2.15: When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms...(p. 35)</p>	
<p>4.9: When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained. (p. 56)</p>	<p>Serious efforts should be made whenever possible to obtain independent evidence concerning the soundness of such score interpretations (pp. 56–57).</p>

Continues next page

Table 3. Excerpts from the *Standards for Educational and Psychological Testing* (AERA et al., 1999) Relevant to Standard Setting (Continued)

Standard	Selected Comments from <i>Standards</i>
<p>4.19: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented. (p. 59)</p>	<p>Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories...If a judgmental standard-setting process is followed, the method employed should be clearly described, and the precise nature of the judgments called for should be presented...Documentation should also include the selection and qualification of judges, training provided, any feedback to judges concerning the implications of their provisional judgments, and any opportunities for judges to confer with one another. Where applicable, variability over judges should be reported. Where feasible, an estimate should be provided of the amount of variation in cut scores that might be expected if the standard-setting procedure were replicated. (pp. 59–60)</p>
<p>4.20: When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.</p>	<p>...It is highly desirable, when appropriate and feasible, to investigate the relation between test scores and performance in relevant practical settings...Professional judgment is required to determine an appropriate standard-setting approach (or combination of approaches) in any given situation. In general, one would not expect a sharp difference in levels of the criterion variable between those just below versus just above the cut score, but evidence should be provided where feasible of a relationship between test and criterion performance over a score interval that includes or approaches the cut score. (p. 60)</p>
<p>4.21: When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way. (p. 60)</p>	<p>The procedures used...should result in reasonable, defensible, standards that accurately reflect the judges' values and intentions...Special care must be taken to assure that judges have a sound basis for making the judgments requested. Thorough familiarity with descriptions of different proficiency categories, practice in judging task difficulty with feedback on accuracy, the experience of actually taking a form of the test, feedback on the failure rates entailed by provisional standards, and other forms of information may be beneficial in helping judges to reach sound and principled decisions. (p. 60)</p>

Kane's Validity Framework

Kane's (1994, 2001) framework for evaluating standard setting studies involves three general sources of validity evidence: procedural, internal, and external, as well as "overall coherence" (2001, p. 59). Although he acknowledged that no one source of evidence is sufficient for validating cut scores, when taken together, these different sources of evidence can support the "interpretive argument" that the cut scores are reasonable and defensible.

Throughout his writings on this topic, Kane emphasized that it is impossible to validate standards or cut scores in an absolute sense. Rather, he characterizes the task of evaluating standards as one of determining reasonableness of the process and the detection of potential fatal flaws. Kane (1994) wrote

The best that we can do in supporting the choice of a performance standard and an associated [cut] score is to show that the [cut] score is consistent with the proposed performance standard and that this standard of performance represents a reasonable choice, given the overall goals of the assessment program. In practice, however, we seldom, if ever, achieve even this goal. A more modest, but realistic goal in most cases is to assemble evidence showing that the passing score and its associated performance standard are not unreasonable. (p. 437)

To accomplish this goal, Kane (1994) suggested evaluating the three aforementioned general categories of validity evidence (procedural, internal, external) to support standards set on educational tests. Cizek et al. (2004), Hambleton (2001), and others have supported these general categories. Each of these general categories is briefly described in the next section.

Procedural evidence

Kane (2001) noted "Procedural evidence is a widely accepted basis for evaluating policy decisions" (p. 63). Procedural evidence for evaluating standard setting "focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures" (Kane, 1994, p. 437). This category of evidence includes the selection of qualified standard-setting participants (judges or panelists), appropriate training of judges, clarity in defining the tasks and goals of the procedure, appropriate data collection procedures, and proper implementation of the method.

With respect to the selection of participants, all panelists should possess sufficient knowledge of the content tested and the population of examinees who take the test. It may also be important to ensure the composition of the panel reflects key characteristics of the population of potential expert panelists. Appropriate training of panelists is also important so that all panelists understand the judgments they will make. It is important to confirm that panelists understood their tasks, had confidence in their ratings, and were able to provide independent, unbiased judgments. Surveying panelists regarding their impressions of the standard-setting session and their thoughts regarding the implementation of the method is often used to evaluate the quality of standard-setting data and the appropriateness of the processes followed.

Internal evidence

Internal evidence for evaluating standard-setting studies focuses on the expected consistency of results, if the study were replicated (see comment associated with *Standard 4.19* in Table 2). A key internal evaluation criterion is the standard error of the cut score, although calculation of this standard error is often not straightforward due to dependence among panelists' ratings (due to facilitated discussion among the panelists) and practical factors (e.g., time and expense in conducting independent replications). For this reason, evaluations of the variability across panelists within a single study, and the degree to which this variability decreases across

subsequent rounds of the study, are often analyzed as internal validity evidence. However, Kane (2001) pointed out that interpretations of the variability of panelists' ratings are not always clear:

A high level of consistency across participants is not to be expected and is not necessarily desirable; participants may have different opinions about performance standards. However, large discrepancies can undermine the process by generating unacceptably large standard errors in the cutscores and may indicate problems in the training of participants. (p. 73)

In some cases, the consistency of results across random or specific subgroups of panelists is studied. Kane (2001) noted that consistency can be evaluated across independent panels, subgroups of panelists, or assessment tasks (e.g., item formats), and he suggested the use of generalizability theory for gauging the amount of variability in panelists' ratings attributed to these different factors.

Kane (2001) also suggested an internal validity analysis that can be done after the cut scores are set. This analysis involves looking at the performance of students very close to the cut scores (borderline students) on items that panelists thought such students would do well on. If these students did poorly or extremely well on such items, the cut score is inconsistent with the panelists' predictions.

External evidence

External evidence refers to the degree to which the classifications of examinees are consistent with other performance data. Kane (2001) characterizes external evidence as being similar to convergent validity evidence. External validity evidence would include classification consistency across different standard-setting methods applied to the same test and examinees, tests of mean differences across examinees classified in different achievement levels on other construct-relevant variables, and the degree to which external ratings of examinee performance are congruent with their test-based achievement level classifications.

External validity evidence is hard to gather and the results may be hard to interpret. These data are hard to gather because valid, external criteria rarely exist (hence the need for tests and standards in the first place) and construction of such measures involves considerable time, personnel, and money. Even when these data are gathered and analyzed the results may be hard to interpret because the validity of the external data would need to be established. This problem of the validity of the criterion has been discussed for well over 60 years in the predictive validity literature (e.g., Guilford, 1946; Jenkins, 1946; Toops, 1944) and it applies in full force to the situation of gathering external evidence in standard setting.

With respect to consistency of standard-setting results across different standard-setting methods, this approach is useful, but has serious limitations. As Kane (2001) described, A lack of agreement between two standard-setting studies using different methods should not be very surprising, because the different methods ask participants to use different kinds of data...in different ways. Nevertheless, if we consider the methods to be exchangeable in the sense that the resulting cutscores are interpreted in the same way, large discrepancies tend to undermine confidence in both cutscores. (p. 75)

Hence, like the procedural and internal sources of validity evidence for evaluating standards, external evidence is not perfect. Therefore, in evaluating the validity of the standard setting conducted on the 2005 Grade 12 NAEP Mathematics test, a comprehensive approach must be taken, with careful consideration of all sources of evidence. In the next section, we describe our evaluation methods, including a description of evaluation criteria, which is drawn from Kane (1994, 2001), the *Standards* (AERA et al., 1999) and other sources found in the literature for evaluating standard setting studies (e.g., Cizek, 1993, 1996b; Cizek et al., 2004; Hambleton, 2001; Meara, et al., 2001).

Method

Our evaluation of the standards set on the 2005 Grade 12 NAEP Mathematics test involved observing as much of the process as possible and reanalyzing the data from standard-setting panelists. These data included their standard-setting judgments across rounds, as well as their responses to the comprehensive surveys they took throughout the process.

In this section, we describe the standard-setting study, and we provide an overview of the Mapmark method. We also describe the meetings we attended as observers, the data analyzed, and the procedures used to evaluate the standard-setting session.

Meetings Attended

The setting of standards on NAEP exams is complex, involving many stakeholders and organizations. Our work on this evaluation started shortly after the contract for the evaluation was officially awarded in October 2004. The operational standard-setting study for this exam occurred in November 2004. Members of our evaluation team attended this four-day study and subsequent meetings of ACT's Technical Advisory Committee for Standard Setting (TACSS) and NAGB's Committee on Standards, Design, and Methodology (COSDAM). Evaluation team members also attended NAGB Board meetings when the standard-setting activities on this exam were discussed. Table 4 documents the meeting dates and purposes for the meetings we observed. The NAGB Board meetings also included COSDAM subcommittee meetings. It should be noted that these committees had important meetings before our evaluation work started. For example, a pilot study was conducted in July 2004 (see ACT, 2005c).

In addition to these meetings, ACT produced several reports related to this study. A listing of the documents we reviewed for this report is presented in Table 5. These reports contained a variety of valuable information regarding the standard-setting study and served as the primary documentation of the process.

Table 4. Standard Setting Meetings Observed by University of Massachusetts–Amherst Staff

Meeting	Date	Purpose
Operational Standard Setting	Nov. 11–15, 2004	Set standards (cut scores) on the 2005 Grade 12 NAEP Mathematics Test
ACT Technical Advisory Committee on Standard Setting	Dec. 17–18, 2004	Present and discuss the results from the standard setting session.
NAGB Committee on Standards, Design, and Methodology	Jan. 11–12, 2005	Present and discuss the results from the standard setting session.
ACT Technical Advisory Committee on Standard Setting	Feb. 17–18, 2005	Review and comment on the final report and presentation regarding ACT's recommendations.
NAGB Board Meeting	March 3–5, 2005	Discuss ACT recommendations.
NAGB Board Meeting	May 19–21, 2005	Continue discussions related to 12th Grade Mathematics Assessment

Table 5. Major ACT Reports Related to 2005 Grade 12 Mathematics Standard Setting

Report	Date
<i>Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Executive summary</i>	April 29, 2005
<i>Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Process report</i>	April 29, 2005
<i>Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Technical report</i>	May 11, 2005
<i>Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Special studies report</i>	May 13, 2005

Data Analyzed

The critical data used to compute recommended cut scores are the bookmark placements and associated cut scores panelists provide after each round of ratings and discussion. These data were provided to us in February 2005. In addition to the panelists' provisional and final cut scores, we also received panelists' responses to surveys they took throughout the four-day meeting, and data on panelists' background characteristics. We also requested and received the item parameters (from the 2004 field test) that were used to help set standards on the 2005 exam.

Description of Panelists' Survey Data

In the Mapmark standard setting conducted on the 2005 Grade 12 NAEP math test, the panelists had the opportunity to evaluate the process multiple times. After each round and at the end of each day, a questionnaire was administered to solicit panelist feedback regarding their understanding of the methods, the perceived clarity of the processes, and their opinions of any other information covered during the study. Each survey included Likert-type and open-ended questions. A total of six questionnaires were administered during the four-day period. At the end of Round 4, panelists answered a questionnaire regarding their opinions about the cut scores that were determined by the group. In addition, panelists were asked to fill-in their cut score at each achievement level and estimate the percentage of students who would be at or above the achievement level. This method was a clever way to determine how well the panelists understood the procedure as a whole and how well they understood the information given to them when they provided their final cut score recommendations.

Description of the Mapmark Method

The Mapmark standard-setting method is a considerable extension of the Bookmark method. To understand the Mapmark method, it is helpful to first understand the Bookmark method, and so we provide a brief description of the Bookmark next. More comprehensive descriptions of the Bookmark method are provided in Lewis, Mitzel, and Green (1996); Lewis, Green, Mitzel, Baum, and Patz (1998); Lewis, Mitzel, Green, and Patz, (1999); and Mitzel, Lewis, Patz, and Green, 2001 (see also Cizek, Bunch, and Koons, 2004; and Karantonis and Sireci, 2006). However, before describing the Bookmark and Mapmark methods further, it is important to note that both methods involve many of the same critical steps as other standard-setting methods. That is, panelists are typically oriented to the purpose of the study, discuss the concept of “borderline” students, take sets of test items (without the answer key) to get an appreciation of test difficulty, and spend significant time deliberating before making their judgments.

Description of the Bookmark Method

The Bookmark method uses item response theory (IRT) to “map” items onto the score scale in which cut scores (standards) need to be set. A key feature of the Bookmark method is the *ordered item booklet* (OIB), which is a booklet of test items in which the items appear in ascending order of difficulty (as estimated using an IRT model). Panelists review the OIB and spend a significant amount of time discussing the knowledge, skills, and abilities (KSAs) students need to perform successfully on the items. This discussion involves reviewing every item in the OIB. Participants are encouraged to discuss “(a) what knowledge, skills, and abilities must be applied to correctly respond to a given item, and (b) what makes each item progressively more difficult than the previous item in the booklet” (Mitzel et al., 2001, p. 253).

Following these extensive discussions, panelists are asked to place a bookmark in the OIB where students who are at the border of a specific achievement level (e.g., borderline Basic/Proficient) are likely to have success on all items *before* the bookmark, but are not expected to have a high likelihood of success on items appearing after the bookmark. We forestall discussion of what “high likelihood of success” means for the moment. Since IRT places items and examinees on the same scale, the location of the item preceding a panelist’s bookmark can be used as the panelist’s cut score. The final cut score is calculated by taking the average (mean or median) of the panelists’ cut scores.

The OIB used in both the Bookmark and Mapmark methods typically contains one item per page. Selected-response items appear once in the OIB, but constructed-response items that

are polytomously scored (e.g., a student can get from 0 to 4 points on an item) appear several times—once for each score point. As mentioned earlier, the task required of each panelist is to place a bookmark in the OIB at a position that represents her or his best estimate of the point at which the borderline student for a particular category is likely to have mastered items before the bookmark but not items after the bookmark. For selected-response items, “mastery” is typically defined as a having at least a .67 probability of answering the item correctly. For polytomously scored items, mastery is defined as having at least a .67 probability of receiving a particular score point or higher.¹³ As originally described in Lewis et al. (1998), “the location of a [selected-response] item is defined as the point on the scale at which a student has a .67 (2/3) probability of success, with guessing factored out” (p. 3). For the polytomously scored items, each score point has a unique location on the scale, defined as the point at which a student has a .67 probability of obtaining the specific score point or higher.

As in many other standard-setting procedures, the Bookmark method proceeds in rounds. In most cases, the number of rounds is three. The rounds following the initial bookmark placement are designed to foster consensus as the study progresses. As described above, the first round ends when the panelists place their bookmarks in the OIB. It is important to note that these initial bookmark placements are done independently, without discussing their choices with other panelists. During the second round, participants are provided feedback from Round 1 cut scores (e.g., the average cut score and range of cut scores across panelists) and discuss this information. This Round 2 discussion “centers on what students should know to attain a given achievement level” (Mitzel et al., 2001, p. 254). At the end of Round 2, participants provide an updated set of cut scores (they can of course, reassert their initial bookmark placements, if they wish). New Round 2 cut scores are then calculated based on panelists’ new cut scores.

Round 3 typically begins with the presentation and discussion of impact data (percentage of students expected to fall into each performance category) estimated from the Round 2 results. At the end of Round 3, participants make their final bookmark placements.

Extending Bookmark to Mapmark

The Mapmark standard-setting method was developed to improve the process of setting standards on NAEP (ACT, 2005b). One specific criticism of the process used to set standards on other NAEP exams was a lack of correspondence between the achievement level descriptions and the types of items that students within an achievement level could answer successfully (Linn, 1998; Pellegrino et al., 1999). As described earlier, the first round of the Mapmark (and the Bookmark) method involves comprehensive reviews of items and discussions of the KSAs required to answer them, followed by panelists placing their bookmark at the point in the OIB at which items before the bookmark have a high probability of being answered correctly by the hypothetical borderline student. Hence, from the outset, the method explicitly links expected performance on specific items to the achievement levels.

The ostensible improvements in the Mapmark over the Bookmark method essentially come from the use of “teacher domains,” “domain scores,” and “item maps.” These additional features are designed to provide clarity for panelists with respect to their sense of the most appropriate locations for their recommended cut scores. As described in ACT (2005b), “ACT believed that the Bookmark method contains some very attractive features for setting standards, but that it could be improved with the use of item maps...and domain-score feedback” (p. 17).

¹³ The use of a response probability (RP) of .67 is somewhat controversial (see Karantonis and Sireci, 2006, for a discussion of research related to choice of RP) and values other than .67 are sometimes used or recommended (Kolstad et al., 1998). As described below, the customary RP of .67 was used during the first round of ratings for the 2005 Grade 12 NAEP Math standard setting. We discuss this issue further in the results section.

Teacher domains represent sets of items that are homogeneous with respect to the KSAs required to answer them. In general, domains represent a single skill or content area. They are more general than a single item, but more specific than the content domains (subscales) in the NAEP frameworks. For NAEP math, teacher domains and domain scores were created within each subscale. Schulz, Lee, and Mullen (2005) explain that creation of these domains allows content experts to focus their judgments on reliable content distinctions within a test. The groupings of items into domains are based on the judgments of content experts.

As described in the process report for the 2005 Grade 12 Math standard setting, “ACT proposed to develop for use in the Mapmark method, the kinds of domains that would be most useful for describing to educators and noneducators alike, in a clear and reliable fashion what it is that students at a given level of achievement can or cannot do, and what growth in achievement means” (ACT, 2005b, p. 17). This same report describes teacher domains as having three important features: (a) a clear definition (i.e., the domain definition consists of a brief title, brief narrative description, and up to three sample items), (b) coherence (i.e., teachers should be able to reliably classify items into domains using only the definitions), and (c) variability in difficulty (i.e., domains should differ in difficulty and cover wide range of proficiency; p. 18). This last characteristic illustrates the qualitative and statistical work that goes into creating these domains.

The domains are created so that they are distinct with respect to both content *and* difficulty. Specifically, domain characteristic curves are computed (using the IRT item parameters for items within the domain) and these characteristic curves tend to be non-overlapping and ordinal with respect to difficulty (ACT, 2005b; Schulz et al., 2005). The domain scores represent sub-scores from items within the teacher domains that are distinct in terms of difficulty.

Because domain scores and domain characteristic curves represent sets of items that make cohesive sense to standard-setting panelists, they are used to facilitate discussion of the expected performance of borderline students on items measuring the domain. An example of how teacher domains were used to provide feedback to the panelists participating in the 2005 Grade 12 NAEP Math standard setting is provided in Table 6. As is evident from this data display, panelists can discuss whether these data are sensible, given their understanding of (a) the achievement level descriptions, (b) the knowledge and skills of borderline students, and (c) the teacher domains.

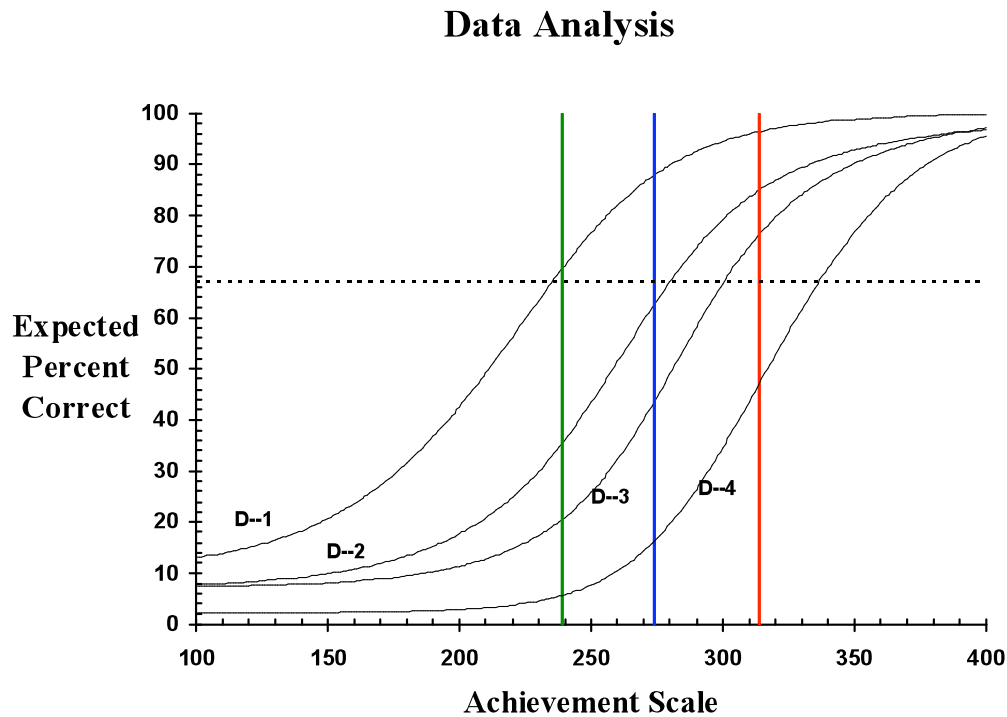
Table 6. Sample Feedback Using Teacher Domains

NAEP Subscale	Teacher Domain	Expected Percent Correct on Score Domain at Lower Borderline of		
		Basic	Proficient	Advanced
Number Properties and Operations	N1. Perform Basic Operations	81	90	96
	N2. Determine Correct Operations	59	81	95
	N3. Place Value and Notation	42	68	95
	N4. Multistep Problems	19	44	83

Source: Adapted from ACT(2005b), p. 48.

An example of domain score information provided to panelists is presented in Figure 1. As can be seen in this figure, the preliminary Round 1 cut scores suggest that borderline “Basic” students would master items from the first domain, but not the other three. Panelists would discuss such information as they consider adjustments to their cut scores after Round 1.

Figure 1. Example of Domain Characteristic Curve Information Provided to Panelists



Source: ACT (2005b), p. 47.

Notes: Vertical lines show the locations of preliminary cut scores from round 1 for Basic, Proficient, and Advanced, from left to right, respectively. Dashed horizontal line shows 67 percent mastery criterion.

Item maps are another important feature that distinguishes Mapmark from Bookmark. Item maps are graphical representations of items, arranged in order of difficulty, stratified within a domain. In the Mapmark method used to set standards on the 2005 Grade 12 NAEP math test, various item maps were used to illustrate the difficulty orderings of items (according to the NAEP score scale) stratified by the NAEP math content domains (subscales) and teacher domains. Panelists’ cut scores, or the average cut score for a group of panelists, can be placed within the map to facilitate discussion of preliminary cut scores. An example of this type of item map is presented in Figure 2, which is an item map illustrating the ordering of items according to their difficulty (expressed in terms of the NAEP score scale) and three math content areas (subscales). The horizontal lines in the item maps are the median cut scores from Round 1.

Figure 2: Sample “Primary Item Map” Illustrating NAEP Subscales
 Source: Adapted from ACT (2005b, p. 37).

Scale	Number Properties and Operations	Measurement/Geometry	Data Analysis
Above			M117 M118 M119 P36_4
354			
351			
348			
345			
342			
339			P35_4
336			M116 P34_2
333			
330			P31_2 P34_1
327		P29_2 D22 P30_2	
324		M115 P28_2 D20	D19
321	P27_4	P26_2 M114 M113	D18
318	P27_3 P24_2	P25_2	
315	P27_2	M110 P30_1 D17	
312	P27_1		M107
309	M106	P21_4	P20_2
306		M102 M103	P35_3
303		P29_1 P28_1 P18_2 D15	M100
300		M96 P18_1 P25_1 P17_2	M97
297	M95 P24_1	P16_2	
294	M86 D14	M88 M87 P14_2 M92	P13_2 P36_3
291	D11 M84		D12 P31_1
288		P10_3 P11_2 M82	P9_4 P12_2
285	M78	M74	P35_2 M75 M80
282	P7_2	M70 P21_3 M73 M72	D10
279		P5_2 P11_1 P17_1	M67 P13_1
276	D9 M66 P4_2	M65	M63
273	P4_1 M62	M58 M60	
270	M55		P20_1 P9_3
267	M49 P3_2 M52	P26_1 P5_1	M53 M54
264	P7_1	D7 M45 M44 M47	P2_4 M46
261		M41 M42	P9_2
258		P10_2 P21_2	
255	M33 M37 M38 D6	M32 M34	P36_2 M36
252	M30	M26 M28 P14_1	M27 P9_1 M29 P12_1
249			M25
246	P1_2 M22 P3_1	M23 M24	P36_1
243		M21 D4	D3
240	M20		
237		M19	P35_1
234		M17	
231		M15 M16 P10_1	
228		D2	
225		M13 M14	M12
222		M11	P2_3 D1
219		P21_1	
216	M10	M7 M9	M8
213	M5 M6	P16_1	
210			
207			
204	M4		
Below	P1_1 M2	M3	P2_1 P2_2

Another important source of feedback for panelist review and discussion after Round 1 is the “percent correct table,” which illustrates the percent correct scores within a domain for students at the preliminary cut scores (i.e., borderline students). A sample percent correct table is presented in Figure 3. In the Mapmark standard setting for this NAEP math test, panelists were asked to make judgments regarding whether these percent correct scores seemed to be “too low, OK, or too high for the borderline of each achievement level” (ACT, 2005b, p. 5). Next, they were asked to “choose a scale score for their Round 2 cut score recommendations” (p. 55).

They were instructed that if most of their ratings on the percent correct scores for the domains indicate the percentages are too high, they might want to recommend a higher cut score, and so forth. Panelists independently provided their ratings.

After the first round of standard setting (bookmark placements), preliminary cut scores were calculated for each panelist based on their bookmark placements and the average cut scores were highlighted in subsequent graphs provided to panelists. For this NAEP math test, the Mapmark method involved four rounds (which is another departure from a typical Bookmark study). In Round 2, the feedback provided to panelists using teacher domains, domain scores, and percent correct tables were used to facilitate discussion among panelists and a revised cut score for each achievement level from each panelist.

In Round 3 of this implementation of the Mapmark method, the panelists discussed revised item maps, domain score charts, and percent correct tables. These tables and figures were revised by updating the preliminary cut scores based on Round 2. The panelists were then asked to provide revised (if necessary) cut score recommendations.

In Round 4, the item maps, domain score charts, and percent correct tables were revised based on the Round 3 cut scores and redistributed to panelists. In addition, panelists were given “consequences data,” which indicate the expected percentages of students within each achievement level, the expected percentages of students at or above each level, and the expected percentage of students below the lowest cut score (basic). Panelists were asked to review these data and provide revised (Round 4) cut scores based on these new data, if necessary.

Figure 3. Sample Percent Correct Table

Subject	Function Domain	Item ID	Number of Students		Percent Correct	Proficiency
			Correct	Total		
Math	1. Problem Solving Operations	M1	400	400	100%	100%
	2. Operations and Properties	M2	176	176	100%	100%
	3. Operations and Properties	M3	176	176	100%	100%
	4. Operations and Properties	M4	176	176	100%	100%
	5. Operations and Properties	M5	176	176	100%	100%
Algebra	6. Operations and Properties	A1	476	476	100%	100%
	7. Operations and Properties	A2	476	476	100%	100%
	8. Operations and Properties	A3	476	476	100%	100%
	9. Operations and Properties	A4	476	476	100%	100%
	10. Operations and Properties	A5	476	476	100%	100%
Geometry	11. Operations and Properties	G1	396	396	100%	100%
	12. Operations and Properties	G2	396	396	100%	100%
	13. Operations and Properties	G3	396	396	100%	100%
	14. Operations and Properties	G4	396	396	100%	100%
	15. Operations and Properties	G5	396	396	100%	100%
Data Analysis	16. Operations and Properties	DA1	476	476	100%	100%
	17. Operations and Properties	DA2	476	476	100%	100%
	18. Operations and Properties	DA3	476	476	100%	100%
	19. Operations and Properties	DA4	476	476	100%	100%
	20. Operations and Properties	DA5	476	476	100%	100%
Algebra	21. Operations and Properties	A6	476	476	100%	100%
	22. Operations and Properties	A7	476	476	100%	100%
	23. Operations and Properties	A8	476	476	100%	100%
	24. Operations and Properties	A9	476	476	100%	100%
	25. Operations and Properties	A10	476	476	100%	100%

Note: Adapted from ACT (2005b, p. 48).

Evaluation Criteria

As mentioned earlier, our criteria for evaluating the 2005 Grade 12 NAEP Mathematics standard setting were drawn from the *Standards* (AERA et al., 1999) and from seminal writings in the standard setting literature (e.g., Brennan, 1995; Cizek et al., 2004; Kane, 1994, 2001; Hambleton, 2001; Linn, 1998; Pitoniak, 2003, cited in Cizek et al., 2004). Kane’s (1994, 2001) framework for validating standards involves three evaluation dimensions—procedural, internal, and external, as described earlier. Specific criteria within each dimension are presented in Table 7. This table highlights the 19 criteria we used to evaluate the 2005 Grade 12 NAEP

Mathematics standard-setting study. We created this list by synthesizing the suggestions from the literature previously cited as well as from the *Standards*.

Some of the criteria listed in Table 7 should be clear from their brief description and the review provided earlier, while others may need further explanation. For example, much has been written regarding the selection of panelists (e.g., AERA et al., 1999; Hambleton, 2001; Jaeger, 1991; Raymond and Reid, 2001; Reid, 1991). Guidelines for selecting panelists suggest including panelists from diverse backgrounds (e.g., different specialty areas, ethnicities, geographic regions, males and females, etc.) who are qualified in the subject area of interest. Qualifications may include years of teaching experience, certification and other indicators of teaching excellence, and familiarity with the types of students tested. Jaeger (1991) and Raymond and Reid (2001) emphasized that standard-setting participants must be knowledgeable in the area being tested, be able to understand and perform the required tasks, and be able to work well within a group setting. Consideration of these factors in selecting standard-setting panelists helps illustrate that the panelist selection process was carefully done and well conceived.

Selection of participants also involves selecting a sample large enough to produce reliable results. Although the literature contains examples of standards set using as few as five panelists (Livingston and Zieky, 1982), others have suggested 15–25 panelists should be used to make the standards more defensible (e.g., Hambleton, 2001; Mehrens and Popham, 1992). Jaeger (1991) suggested recruiting enough panelists so that the standard error of the cut score would be below an acceptable level (e.g., one-fourth of the standard error of measurement for the test).

With respect to panelist training, evidence that panelists understood their tasks, took samples of test items under exam-like conditions, practiced performing the required tasks, and had their questions regarding task completion sufficiently answered, suggests the training was done well. Proper training also involves adequate discussion of the achievement levels and the types of students likely to be at the borders of the achievement level categories.

Internal criteria involve analysis of panelists' data across panelists, subgroups of panelists, independent panels, rounds, item formats, and any other facets relevant for evaluating the generalizability of the results. One of the implicit assumptions in standard-setting methods that involve group discussions and several rounds of panelists' ratings is that panelists will influence each other in constructive ways that will foster convergence to consensus cut scores. If that ideal occurs, the variability across panelists would decrease from earlier to later rounds.

It should be noted that the 19 criteria listed in Table 7 represent an amalgamation of many of the suggestions found in the literature regarding the evaluation of standard-setting studies. Neither the literature nor the *Standards* mandates that a standard-setting study should satisfy all of these criteria.

Table 7. Summary of Criteria for Evaluating 2005 Grade 12 NAEP Math Standard Setting

Evaluation Dimension	Criterion	Brief Explanation
Procedural	Care in selecting participants	Qualifications, competence, and representativeness of panelists; sufficient number of panelists
	Justification of standard-setting method(s)	Degree to which methods used are logical, defensible, and congruent with testing purpose
	Panelist training	Degree to which panelists were properly oriented, prepared, and trained
	Clarity of goals/tasks	Degree to which standard-setting purposes, goals, and tasks were clearly articulated
	Appropriate data collection	Data were gathered as intended
	Proper implementation	Method implemented as intended
	Panelist confidence	Panelists understood tasks and had confidence in their ratings
	Sufficient documentation	Documentation of the entire process so (a) it is understood and (b) can be replicated
Internal	Sufficient inter-panelist consistency	Reasonable standard deviations and ranges of cut scores across panelists
	Decreasing variability across rounds	The variability across panelists' cut scores decreases across rounds—evidence of emerging consensus
	Small standard error of cut score (consistency within method)	Estimate of degree to which cut scores would change if study were replicated
	Consistency across independent panels	Estimate of degree to which cut scores would change if different panelists were used
	Consistency across panelist subgroups	Estimate of degree to which cut scores would change if specific types of panelists were used
	Consistency across item formats	Estimate of the consistency of cut scores across item formats (e.g., SR, CR items)
	Analysis of borderline students performance on specific items	Degree to which expectations of hypothetical borderline students' performance are consistent with the performance of students near the cut scores
External	Consistency across standard-setting methods	Degree to which results from different standard-setting methods yield similar results
	Consistency across other student classification data	Degree to which classifications of students based on external data are congruent with classifications based on the cut scores
	Mean differences across proficiency groups on external criteria	Degree to which students classified into different achievement levels differ on other relevant variables
	Reasonableness	Degree to which cut scores produce results that are within a sensible range of expectations

This page left intentionally blank

Results

The previous sections of this report introduced the 2005 Grade 12 NAEP Mathematics exam and described (a) problems and issues in setting standards on educational tests, (b) the data we gathered in conducting our evaluation, (c) the Mapmark standard-setting method, and (d) professionally accepted criteria for evaluating standard-setting studies. In this section, we summarize the results of our evaluation of the standard setting. This section is organized using Kane's procedural, internal, and external evaluation criteria.

Procedural Evidence

Selection of Panelists

NAGB policy stipulates that standard-setting panels have a broad representation that includes educators and the general public. According to the policy, 70 percent of the panelists should be educators (55 percent of whom are classroom teachers) and 30 percent from the general public (Loomis and Bourque, 2001). A summary of the criteria to be used in selecting panelists is presented in Table 8.

Table 8. Recommended Criteria for Selecting NAEP Standard Setting Panelists

Criterion	Target
Grade Level Classroom Teachers	55 percent
Non-teacher Educators	15 percent
General Public	30 percent
Diverse Minority/Racial Ethnic Group	30 percent
Male	Up to 50 percent
Representative of the four NAEP Regions	25 percent (each region)

The selection process of panelists used complex sampling techniques that are proportional to the demographic recommendations set by NAGB. First, stratified samples of school districts from a national database were drawn to incorporate the diversity of districts' demographics. That is, 15 percent of the sample drawn needed to be districts with enrollments over 25,000 and 15 percent of the sample drawn needed to be districts with at least 25 percent of the population considered to be in poverty. A total of 687 nominators were drawn, with 84 percent from public and 16 percent from private schools. Accordingly, 61 percent of the nominators were teachers, 7 percent were non-teacher educators, and 32 percent were from the general public.

The next step was to identify individuals to recommend panelists for the standard setting by drawing three separate samples from the districts without replacement. A separate draw for teacher panelists, non-teacher educators' panelists, and general public panelists was conducted for public school districts and repeated for private school districts. The individuals chosen as panelists could nominate up to four candidates whom they felt would be qualified. Finally, the pool of qualified nominees was selected from a computer algorithm that rated each nominee based on the information they provided.

Forty-six panelists were initially selected, but due to extraneous factors such as time conflicts, only 31 panelists agreed to participate. These 31 panelists represented 23 states. Ten panelists (32 percent) were from the Northeast, eight (26 percent) were from the Midwest, seven (23 percent) were from the West, and six (19 percent) represented the South. Thirteen (42 percent) were women and 18 (58 percent) were men. Four (13 percent) were black, three (10

percent) were Hispanic, and two (6 percent) were Asian; the remaining 22 (70 percent) were Caucasian.

Seventeen panelists (55 percent) were teachers. Three of these teachers (18 percent) received the national *Teacher of the Year* award. Nine panelists (29 percent) were classified as “general public.” Five panelists (16 percent) were classified as non-teacher educators. This group included a professor of mathematics and individuals who worked in education at the state level. Most of the panelists considered to be “general public” worked in fields that were related to mathematics (e.g., statistician, engineer, investment banker). However, two panelists did not have mathematical careers (e.g., a mayor and a detention administrator). The desired characteristics of the group of panelists are compared with the actual percentages in Table 9. The actual percentages met or came extremely close to their targets for four of the six criteria. For the other two criteria (sex and geographic region), the actual percentages were within one or two panelists of the target. It is also notable that the overall number of panelists (31) is large, relative to suggestions in the literature (e.g., Hambleton, 2001; Mehrens and Popham, 1992).

Table 9. NAGB’s Recommended Criteria for NAEP Standard Setting Panels

Criterion	Target Percentage	Actual Percentage
Grade Level Classroom Teachers	55	55
Non-teacher Educators	15	16
General Public	30	29
Diverse Minority/Racial Ethnic Group	30	29
Male	Up to 50	58
Representative of the four NAEP Regions	25 (each region)	19, 23, 26, 32

Justification of Standard Setting Method

Our earlier description of the Mapmark method illustrated the logic underlying the method and how it could address specific criticisms of prior methods used to set standards on NAEP exams (e.g., ensure the cut scores reflected differences in student performance that were congruent with the differences in the NAEP achievement level descriptors). In general, we believe the justification of the use of this method to set standards on the 2005 Grade 12 NAEP Math exam is extensive. First, the method begins with the Bookmark method, which is currently the most popular method for setting standards on state-mandated educational tests (Karantonis and Sireci, 2006) and was developed to address limitations associated with previous standard setting methods. Specifically, it is designed to set multiple cut scores in a single test, it can be used on tests comprising multiple item types (selected- and constructed-response), and it is purportedly cognitively simpler for panelists (Lewis et al., 1998). Proponents of the method also claim the OIB helps standard setting panelists understand the relative difficulty of the items and helps focus their attention on the types of items that are likely to be answered correctly by examinees at different proficiency levels.

The Mapmark method was designed to address limitations of the Bookmark method by introducing groupings of items into meaningful domains, and adding instructive graphs to inform their judgments. Specifically, the use of item maps and domain scores aims to facilitate panelists’ discussions and focus their judgments on how borderline students *will* perform on test content located around the cut scores. A criticism of prior methods used to set standards on NAEP exams is that there was no evidence that achievement level definitions actually described how students within the achievement levels actually performed on the exam (Linn, 1998; Pellegrino et al., 1999). Over the years, researchers have questioned the degree to which

students classified into specific achievement levels actually exhibit the skills included in the achievement level descriptions. The purpose of producing domains is to create a clear and internally consistent description of achievement levels of what students can and cannot do at certain ability levels (Schulz et al., 2005). ACT (2005b) claims domain scores are also very helpful in articulating the skills possessed by students in the various achievement categories and can also be used to facilitate the selection of exemplar items for different achievement levels.

In addition to theoretical justification, four studies were done to evaluate the Mapmark method before it was approved to set standards on this exam. Two of these studies were characterized as “field trials.” These field trials used items from the 2003 Grade 8 Mathematics test. The purposes of the studies were to evaluate the degree to which panelists understood the item maps and domain score information and to provide information regarding choice of response probability (RP) value for ordering items in the OIB. The third study also used items from the 2003 Grade 8 NAEP Math test. The purpose of that study was to compare the results of Mapmark-derived cut scores on that exam to those derived using the item rating (modified Angoff) method that was used to set the actual standards on the exam. The fourth study was a pilot study where all 2005 Grade 12 NAEP math items were used to set cut scores. The pilot study implemented and compared two methods—the item rating method and Mapmark (Yin and Schulz, 2005). The results of these studies supported the use of the Mapmark method to set standards on the 2005 Grade 12 exam. Thus, use of the Mapmark method to set standards on this test was defensible from both theoretical and empirical perspectives.

Research related to justification of the method

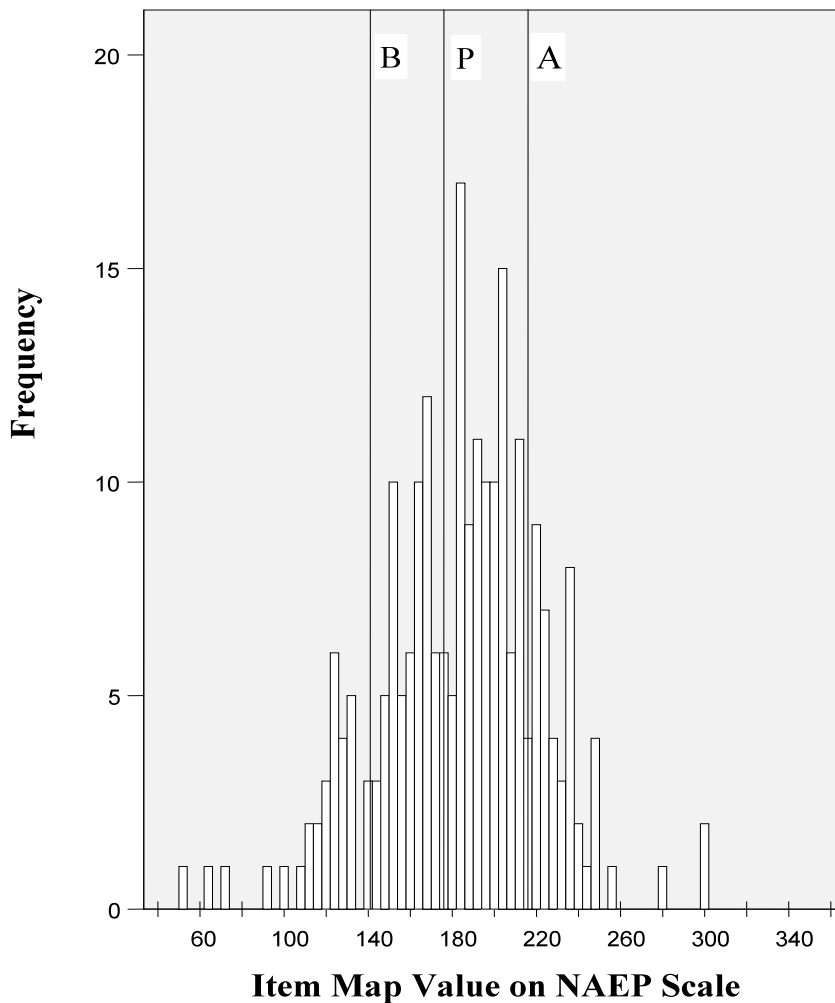
Recently, Reckase conducted two simulation studies to evaluate the degree to which standard setting panelists’ “intended” cut scores were recovered from the standard-setting process (Reckase, 2006a, 2006b). The simulations mimicked panelists’ ratings under two standard-setting method conditions—Angoff and Bookmark. The studies also involved simulated data under error-free and error-perturbed conditions. For the Bookmark method, the RP67 criterion (i.e., the probability that a borderline student will answer the item correctly is .67) was used, which was the same criterion used on this NAEP exam. Reckase (2006a) found that under the error-free condition, the Bookmark cut scores were statistically negatively biased (i.e., lower than the simulated cut scores) and the magnitude of the bias increased when error in panelists’ judgments was introduced. Systematic bias was not present under the Angoff condition. He attributed this finding to the distance between the cut score for each panelist based on the bookmark placement (which is based on the difficulty of the item just prior to the bookmark) and the panelists’ intended cut score. Although these simulation results could only describe the first round of a Bookmark study (before panelists discuss ratings), the systematic bias was troubling.

Schulz (2006) criticized the simulation methods used by Reckase (2006a) and claimed that the simulations were not representative of actual bookmark applications. Specifically, he commented that panelists are explicitly encouraged to go beyond their initial bookmark placements to define a “range of uncertainty” (p. 7) before finalizing their bookmark placement, and so their judgments are much more complex than those simulated by Reckase (2006a). Reckase (2006b) extended his simulation conditions to approximate panelists searching for a range of uncertainty. Under this condition, he found very little negative bias for the majority of the score scale, but still substantial bias at the extremes where there were relatively few items. He concluded that the practice of encouraging bookmark panelists to consider a range of items “shows considerable promise over the method based on selecting the first item judged to have a correct response below the mapping probability” (p. 17).

Our evaluation of the Reckase and Schulz studies with respect to their pertinence to the Mapmark standard setting on the 2005 Grade 12 NAEP Math test is that the multiple rounds involved in this Mapmark application probably nullified any negative bias that may have been present in the panelists' initial bookmark placements from Round 1. ACT used the range of uncertainty approach in instructing the panelists, which would have mitigated any potential negative bias.

The issue of item density, however, remains. That is, a potential problem could occur if the median cut scores for panelists were placed where there were relatively few items on the score scale and where there were large gaps between these few items. Therefore, in considering the justification for using the Mapmark method on this NAEP exam, we evaluated the density of these NAEP items with respect to where the cut scores were set. A histogram of the item locations along the 2005 Grade 12 NAEP Mathematics score scale is presented in Figure 4. These items are located according to their NAEP item map value, which is close to their RP67 values that were used to order the items in the OIB. For the polytomous items, each score point is mapped at its RP65. For the multiple-choice items, the items are located at RP74 or RP72, for four-choice, and five-choice items respectively. Although these are not the precise RP values used to create the OIB, they are close enough for evaluating the density of the items along the score scale in locations near where the cut scores were set. The cut scores resulting from the standard setting study are illustrated with vertical lines in Figure 4. As is evident from the figure, these cut scores occur where there is sufficient item density. This finding suggests that negative statistical bias was not likely to be a problem in this study.

Figure 4. Item Density for 2005 Grade 12 NAEP Mathematics Items



Note: Vertical lines preceding B, P, and A indicate the cut scores for the Basic, Proficient, and Advanced achievement levels, respectively

Another issue related to the justification of the Bookmark method that some researchers have challenged is the choice of RP criterion. ACT based the choice of the RP67 criterion on two field trials where they compared RP50 and RP67 criteria. In these field trials, they found that panelists strongly preferred the RP67 criterion and that use of that criterion led to cut scores that were more similar to the operational cut scores established using the item-rating method (ACT 2005b; Williams and Schulz, 2005). The RP67 criterion was then used in a special study on the 2003 Grade 8 NAEP Math test and was reviewed favorably. Based on this information, ACT's Technical Advisory Committee for Standard Setting recommended use of the RP67 criterion. As described in Karantonis and Sireci (2006), selection of the RP criterion is a controversial aspect of the Bookmark method, but the RP67 criterion seems to be the most common choice and has theoretical justification in that it represents the location where the information for the correct response (of dichotomously scored items) is maximized (Huynh, 2006).

In considering the field trials, pilot studies, research on the Bookmark method, and the ways in which the Mapmark method improves upon the Bookmark process, we conclude the Mapmark method is well justified as a *reasonable* method for setting achievement level cut scores on this exam. We turn now to a discussion of how well this reasonable method was implemented on the 2005 Grade 12 NAEP Mathematics assessment.

Implementation of the Method

Before the standard setting was conducted, all panelists received a briefing booklet describing the process and the activities planned during the achievement level study. The booklet included a description of the activities for each round, the study schedule, and a glossary of terms.

The standard setting took place in one room, which had six tables to which the panelists were assigned. Five tables had five panelists and one table had six panelists. Each table had three men and two women, and an ethnic distribution of three Caucasians and two members of a different ethnicity (e.g., one black and one Hispanic). The table with six panelists had three men, three women, and four Caucasians.

The six tables were evenly divided into two groups, group A and group B. Two groups were created for reasons of efficiency. It was not expected that all panelists could review all 180 items and so two sets of items were created (one for group A, one for group B). This strategy allowed all items to be rated by at least half the panelists. The items sets also had overlap consisting of 34 items common to both sets. The common items allowed for an evaluation of the consistency of panelists' judgments across the two initially independent groups. The two sets of items were balanced for difficulty, item type, and representation of teacher domains.

The difficulty and response probabilities for the items were estimated from a 2004 pilot test of the exam using a similarly representative sample of grade 12 students. About 10,000 students participated in this pilot test.

Panelist training

Following typical orientation procedures (e.g., welcome, introductions, descriptions of NAEP, etc.), an explanation of the achievement level standards and their importance was presented. Next, one form of the exam was given to each group under the same testing conditions as the students. The purpose of this exercise was to give the panelists an idea of the difficulty of items under test-like conditions and to familiarize them with the items. After taking the test form, panelists were given the answer keys and scored their own exams.

Next, the panelists were trained on the Mapmark method. This training included explanations of item maps, item characteristic curves and RP values, the OIB, and how to understand and identify the KSAs measured by the items. The next phase of training described their primary Round 1 task—placing their bookmarks, as well as a thorough discussion of the achievement level descriptors and the concept of “borderline” students. The panelists were also informed of the types of feedback they would experience in subsequent rounds.

Data collection

The data collection was summarized earlier in describing the Mapmark method and so only a brief overview with a few more details is presented in this section. Complete details are provided in ACT (2005b). The procedure involved four rounds of data collection. In Round 1 panelists were asked to determine the last item a borderline Proficient student would have a probability of 0.67 of answering correctly. (The cut score for Proficient was set first, followed by Basic, and then Advanced.) The second round began with the panelists reviewing feedback data

including the median and range of Round 1 cut scores, domain scores, and item maps with the provisional (Round 1) cut scores highlighted (see Figure 2). Panelists were asked to circle their cut score within the Domain Score Chart, an example of which is presented in Figure 5. The domain score chart presents the expected percent correct score within each domain for every scale score within 10 points of the lowest and highest panelist cut score. According to ACT (2005b), this activity enabled the panelists to “see how much difference there was between their cut score and the median both numerically and in criterion-referenced terms” (p. 50).

During Round 2, panelists reviewed the Teacher Domain definitions, the items in each domain, and the other feedback information described earlier such as the expected percentage correct scores for borderline students. These expected scores were used to give the panelists an idea of how the items’ expected percentage scores increased within and between achievement levels, how different items’ expected percentage scores varied within domains, and how the items’ expected percentage scores function over all the achievement levels (see Figure 3 for an example).

Panelists were given specific tasks to ensure they understood the domain scores and then were asked to judge whether the domain score data for borderline students’ performance was consistent with their expectations (i.e., was the percent correct for a borderline student in each domain based on their cut score placement too high, too low, or correct?). The panelists placed their Round 2 ratings by choosing scale values on their Domain Score Chart (Figure 5). They were asked to make these ratings by considering all of the information reviewed during Round 2.

Round 3 began with a review of feedback information from Round 2. This information included revised item maps, percent correct tables, and domain score charts based on the revised cut scores. The OIB page numbers that corresponded to the lowest and highest cut scores across panelists (for each achievement level) were also given to the panelists. The majority of Round 3 consisted of facilitated table and whole group discussion about the information presented thus far and descriptions of the rationales panelists used in making their cut score recommendations. After the group discussion, the panelists worked independently and made a third set of cut score recommendations based on all the information provided.

The fourth round began by distributing revised feedback data (i.e., item map, domain score chart, percent correct table) as well as information about the variability and central tendency of the cut scores over Rounds 1 through 3. In addition, “consequences data” were presented, which were the percentages of students within each achievement level based on the median cut scores from Round 3. These percentages were based on the 2004 pilot assessment. The percentages of students within each achievement level and at or above each achievement level were presented. The percentage of students “below Basic” was also presented. The panelists were allowed to adjust their cut scores based on the feedback information from Round 3, the consequences data, and the discussions of these data, again by choosing scale values on the domain score chart.

Following the fourth round of cut scores, new feedback data were provided to the panelists and they were informed that the Round 4 cut scores would be presented to NAGB. As a final task, panelists were asked to pick exemplar items for each achievement level category using the probability of 0.67 of a student in the middle of the achievement level answering the item correctly.

of the 30 panelists¹⁴ selected a response of 2, which was near the “totally inadequate” endpoint of the scale. Two other questions related to clarity of tasks were answered similarly, with median responses of “adequate” and only one or two panelists responding below the midpoint. The survey data related to clarity of tasks in the final round of the standard setting study is summarized in Table 10. The data indicate that the vast majority of panelists felt they had a good understanding of the final tasks involved in round 4.

Panelists were also asked whether they had sufficient time to complete their tasks. The data for two particularly relevant questions are presented in Table 11. Most of the panelists felt the amount of time given for the tasks was “about right.” The response option suggesting “too much time was given,” was the next most frequent response.

Table 10. Summary of Panelists’ Round 4 Survey Data: Understanding the Standard Setting Process

Survey Statement	Rating Scale Point Frequency					Median Response
	1 (Totally Disagree)	2	3	4	5 (Totally Agree)	
I understood the purpose of this meeting.	0	0	0	6	24	Totally Agree
I understood the Round 4 median cut scores.	0	0	1	6	23	Totally Agree
I understood the domain score feedback.	0	0	1	7	22	Totally Agree
I understood what students at the Round 4 median cut scores can do.	0	0	1	8	21	Totally Agree
I understood how to complete the Consequences Questionnaire.	0	1	4	4	21	Totally Agree
I understood the Round 4 consequences data.	0	1	3	5	19	Totally Agree

The exit survey also asked panelists several questions about their confidence in the process and in their ratings. A particularly interesting statement presented to panelists was, “I would be willing to sign a statement (after reading it, of course) recommending the use of the cut scores resulting from the [this] process.” Of the 29 panelists who responded to this question, 19 responded “yes, definitely,” 9 responded “yes, probably,” and one responded “no, probably.” Panelists were also asked to indicate their level of confidence in their cut score recommendations. Using a five-point scale from 1 (not at all confident) to 5 (totally confident), 28 of the 30 panelists responding to this question selected 4 or 5. The other responses to this statement were 2 and 3.

¹⁴ One panelist left the meeting before completing all the survey data.

Table 11. Summary of Panelists' Exit Survey Data: Sufficient Time

Survey Statement	Rating Scale Point Frequency					Median Response
	1 (Far Too Long)	2	3	4	5 (Far Too Short)	
The amount of time allocated for the Consequences Questionnaire was:	3	1	25	1	0	3
The amount of time I had to complete the tasks I was to accomplish during each round was:	2	2	21	5	0	3

Panelists were also presented with a statement that read, “I believe that the achievement levels capture meaningful distinctions in mathematics performance as described in the [achievement level descriptions].” Twenty-four of the 30 panelists responding to this question selected one of the top two agreement options (the highest option of 5 was labeled “totally agree”); however, two panelists selected a response below the midpoint of the scale, indicating they disagreed with the statement. Panelists were also asked about their confidence in the standard setting *process*. A summary of their responses to these items is presented in Table 12.

In general, these survey data indicated panelists understood their tasks and had confidence in their ratings (although we noted one or two panelists had relatively low confidence ratings on some survey items). Other data, not reported here, indicate the panelists felt the facilitators treated them appropriately. Therefore, our conclusion from the survey data is that the method was implemented well and the vast majority of the panelists believed the process was defensible for setting achievement level standards on this exam. One potentially disconcerting finding is that a few panelists did not agree that the achievement levels captured “meaningful distinctions in mathematics performance as described by the ALDs.”

Table 12. Summary of Panelists' Exit Survey Data: Confidence in Process

Survey Statement	Rating Scale Point Frequency					Median Response
	1 (Not At All)	2	3	4	5 (To a Great Extent)	
I feel that the panelists in this meeting are appropriately qualified for setting NAEP achievement levels.	0	1	1	5	22	5
I feel that this ALS process provided me an opportunity to use my best judgment to recommend cut scores for the NAEP mathematics assessment.	0	0	2	9	19	5
I feel that the panel in this meeting is widely inclusive of groups that should have a say in setting NAEP achievement levels.	0	1	3	5	19	5
I feel that this ALS process has produced achievement levels that are defensible.	0	0	2	14	14	4
I feel that this ALS process has produced achievement levels that will generally be considered reasonable.	0	0	5	11	14	4

Documentation

The standard-setting study was thoroughly documented (see Table 5). A particularly important document is the *Process Report* (ACT, 2005b) because it describes all standard-setting steps in detail, includes examples of meeting materials, summarizes the field trials and pilot studies, presents the results of all analyses conducted, and provides references to other relevant reports. Minutes from all committee meetings (i.e., COSDAM, TACSS) were comprehensive and included data tables, PowerPoint slides, and comprehensive descriptions of completed tasks. When all documentation related to the study is considered as a whole, there are over 500 pages.

In addition to the extensive documentation, there were several evaluators representing different organizations who independently monitored the standard-setting process. At the operational standard-setting study, in addition to the first two authors of this report, NAGB had two staff observers, the ACT TACSS had two representatives, and ETS had one representative.

Summary of Procedural Evidence

In considering all procedural evidence evaluated for this standard-setting study, we conclude the method was adequately supported by evidence from a procedural perspective. Well-designed selection criteria were used to recruit a large number of qualified panelists; there was strong justification for the standard-setting method employed; the implementation of the method involved comprehensive training, several rounds of data collection, facilitated discussion, and several rounds of panelist surveys; and the process was well documented. It is clear that intense thought and care went into the design of the study, the preparation of materials for the meetings, and gathering the data.

Internal Evidence

The use of parallel groups and tables allowed for several types of consistency checks across rounds of the study. In addition, we were able to evaluate the consistency of panelists' recommended cut scores across different panelist groups. Results relating to internal validity evidence reported here include statistical tests of cut scores derived from subgroups of panelists, estimates of the standard errors of the cut scores, and analyses of the changes in panelists' variability across rounds of data collection. ACT conducted many of the same analyses reported in this section (e.g., ACT 2005b, 2005d) and so some of the results here represent reanalysis of the standard-setting data. We also conducted other analyses we felt were appropriate.

Estimates of standard errors of the cut scores

Estimates of the standard errors of the cut scores provide information regarding the stability of the cut score recommendations from a standard-setting panel. Ideally, these standard errors should help evaluate the likely fluctuation of the cut scores, if the study were replicated. It is difficult to properly estimate the standard errors for cut scores due to the interactive discussions that occur among panelists throughout the study. By the time panelists provide their final (e.g., Round 4) cut score recommendations, they have influenced one another, which makes their cut scores somewhat dependent. Thus, using the standard error of the mean cut score does not provide the correct index of the likely variability of this cut score over replications because that statistic is based on an assumption of independence. It should also be noted that the final cut scores recommended by ACT and accepted by NAGB were based on the *median* panelist cut score, not the mean. Nevertheless, since ACT's supporting documentation for the standard setting study included standard errors based on the mean cut score, we also evaluated those data. The mean and median cut scores across rounds are presented in Table 13. Our estimates of the standard errors of the cut scores at each round follow. From Table 13 it can be seen that the mean and median cut scores were consistently similar and the average cut scores did not change very much from one round to the next.

Table 13. Average Cut Scores Across Rounds

Cut Point	Round							
	1		2		3		4	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Below Basic/Basic	240.26	239	240.23	240	239.71	241	239.48	241
Basic/Proficient	273.71	274	275.35	276	274.58	275	274.48	275
Proficient/Advanced	315.00	314	313.98	314	311.35	314	311.39	314

Given that there is no single, commonly accepted procedure for estimating standard errors for cut scores, we (a) reviewed the standard errors computed by ACT in the technical documentation associated with the 2005 Grade 12 NAEP Math assessment (ACT, 2005b, 2005d) and (b) computed several estimates of these standard errors based on our own, independent analysis of the data.

We computed several estimates of the standard errors of the cut scores. First, we computed the standard error of the mean cut score for Round 1, which represented independent judgments. These cut scores are based on the panelists' original bookmark placements that occurred prior to group discussion. Although they are independent, they are not representative of their final cut scores. We also computed the standard errors of the mean cut scores for Rounds 2

through 4, although we realize doing so deviates from standard statistical practice due to the violation of the assumption of independence of observations. A third method for estimating the standard errors was based on an analysis of the cut score data from panelists who participated in the operational study (November 2004) and those participating in the pilot study (July 2004). Our computation of these estimates is described next.

Estimates based on the standard error of the mean

The estimates of the standard errors of the cut scores based on computing the standard error of the mean are presented in Table 14. Two observations in Table 14 are particularly notable. First, the standard error for the Proficient/Advanced cut score at Round 1 is very large, relative to the other two cut scores. This finding may be due to one panelist who placed the bookmark after the last page of the OIB. Second, the standard errors decrease for all cut scores across rounds. The latter finding could reflect dependencies among the panelists due to discussion and so the Round 4 standard errors are probably lower bounds for the estimates if they were based on independent panelists' judgments (or at least they have a negative bias). ACT also computed estimates of the standard errors of the cut scores based on the standard error of the mean. However, they used the square root of $n-1$ in the denominator of the standard error, and we are not sure why. To calculate our standard errors reported in Table 14, we used the formula for the standard error of the mean, which uses the square root of n . Hence, the standard errors reported in Table 14 are slightly *lower* than those reported in ACT (2005b).

Table 14. Standard Errors for Mean Cut Scores ($n=31$)

Achievement Level Cut Score	Round 1	Round 2	Round 3	Round 4
Below Basic/Basic	2.86	1.67	1.60	1.52
Basic/Proficient	2.56	1.49	1.32	1.46
Proficient/Advanced	4.25	2.00	1.75	1.65

Notes: Shaded cells indicate statistic calculated from independent cut scores. All standard errors are slightly lower than those reported in ACT (2005b) due to differences in the denominator of the standard error (n versus $n-1$).

Estimates based on bootstrapping the median

As mentioned earlier, the final recommended cut scores were based on the median panelist rating for each achievement level instead of the mean. Thus, the standard errors reported in Table 14 (and in ACT, 2005b) are even less likely to represent proper estimates of the standard errors of the cut scores. There is no computational formula for the standard error of the median, and so estimating the standard errors of a cut score based on the median is problematic. Probably for this reason, the technical reports we received for this study (ACT 2005b, 2005c) did not include standard errors for the median cut scores.

The median cut score for each achievement level boundary was very close to the mean. However, we wanted to estimate the expected variability of the median cut score, given the final cut scores provided by the 31 standard-setting panelists. To do this, we used a nonparametric bootstrapping procedure. Essentially, bootstrapping constructs an empirical distribution for the statistic of interest and then uses the standard deviation of the empirical distribution as a measure of the standard error. To complete the bootstrapping, we used the following steps:

1. Sample, with replacement, M datasets of size N from the 31 panelists' cut scores. In this case, M equaled 10,000 and N equaled 30.
2. Compute and record the median for each dataset.

3. Compute the standard deviation of the recorded medians. The standard deviation represents the standard error for the median.

Following these steps, we estimated the standard errors to be 2.37, 1.33, and 0.96 for the Basic, Proficient, and Advanced cut scores, respectively. The estimate for Basic is almost a full-point larger than that based on the mean (2.37 vs. 1.52), the estimate for Proficient is just slightly smaller than the one based on the mean (1.33 vs. 1.46), and the estimate for Advanced is about seven-tenths smaller (0.96 vs. 1.65). Clearly, more research is needed on estimates of standard errors for the median, but these estimates based on nonparametric bootstrapping provide another set of data for evaluating the variability of the cut scores across panelists. With the exception of the Round 1 cut score for Advanced, regardless of the method used to estimate the standard error, all estimates were less than 3 points, and most were around 1.5 points.

Estimates based on independent panels

When independent panels are used to set standards on the same assessment, the difference in the standards across panels can be used to estimate the standard errors of the cut scores. The operational standard-setting panel (November 2004) and the pilot study standard-setting panel (July 2004), both set standards on the same item pool. A comparison of the cut scores derived from these independent groups represents a more accurate estimate of the degree to which the cut scores replicate across independent panels.

Brennan (2002) derived a formula for computing an estimate of the standard error of the mean when $n=2$, such as is the case when comparing cut scores derived from two independent panels (i.e., each cut score represents a sample size of one). This formula is:

$$\frac{|X_1 - X_2|}{2}$$

where X_1 and X_2 represent cut scores from two independent groups. Applying this formula to the Round 4 Mapmark cut scores from the pilot and operational standard-setting studies yields standard error estimates of 0.51, 1.79, and 1.71, for the Below Basic/Basic, Basic/Proficient, and Proficient/Advanced cut scores, respectively. In comparing these estimates to those derived from the standard error of the mean or median based on the operational standard setting panel, the estimate is lower for the Below Basic/Basic cut score and higher for the other two cut scores.

We evaluated the statistical difference between the cut scores derived from the pilot and operational standard setting studies by conducting independent t-tests and computing the standard error of the mean difference of the round 4 cut scores for each achievement level cut score. These results, including the standard errors of difference (provided here as another way of viewing standard errors of the cut scores), are summarized in Table 15. All tests failed to reach statistical significance at $p < .05$. The difference for the Basic cut score was only about one point and the differences for the other two cut scores were about 3.5 points. The standard errors of difference were between 2.0 and 2.5 points for each achievement level.

Table 15. Statistical Comparison of Operational and Pilot Study Final Cut Scores

Achievement Level Cut Score	Mean difference	t	df	p	SE of Difference	95 percent CI
Below Basic/Basic	1.01	0.43	50	.66	2.28	-3.57, 5.58
Basic/Proficient	3.58	1.72	50	.09	2.08	-0.67, 7.77
Proficient/Advanced	-3.42	-1.44	50	.16	2.38	-8.20, 1.34

A summary of the different methods used to estimate the standard errors of the cut scores is presented in Table 16. Although there are differences in the estimates across methods, the largest estimate across methods is always less than 2.5 points.

Table 16. Summary of Estimates of Standard Errors of (Round 4) Cut Scores

Achievement Level Cut Score	Method			
	SE Mean	SE Median	Brennan (2002)	SE Difference
Below Basic/Basic	1.52	2.37	0.51	2.28
Basic/Proficient	1.46	1.33	1.79	2.08
Proficient/Advanced	1.65	0.96	1.71	2.38

Consistency across subgroups of panelists

Consistency of the cut scores for each achievement level was evaluated across selected subgroups of panelists. In selecting these subgroups we focused on demographic variables that were thought to most likely reflect differences across panelists. Specifically, we compared cut scores across panelist groups defined by type (teacher, non-teacher educator, general public), group (group A or group B, with groups defined by two different subsets of items), table (six tables, three nested within each group), sex, and geographic region (Midwest, Northeast, South, West), and ethnicity. Table 17 gives the sample sizes for each of these groups. The “other” ethnic group represents an aggregation of African American (n=4), Hispanic/Latino American (n=3), and Asian American (n=2), since the sample sizes within each of these groups were too small for separate analysis.

Table 17. Sample Sizes for Panelist Subgroups

Grouping Variable	Subgroups	n	Percent
Type	Teacher	17	54.8
	Non-teacher educator	5	16.1
	General public	9	29.0
Group	A	15	48.4
	B	16	51.6
Table	1–6	(5 at 5 tables, 6 at 1 table)	19.4
Sex	Female	13	41.9
	Male	18	58.1
Region	Midwest	8	25.8
	Northeast	10	32.3
	South	6	19.4
	West	7	22.6
Ethnicity	Euro-American	22	71.0
	Other	9	29.0

Three-way mixed model ANOVAs were conducted for each subgroup analysis depicted in Table 17. The two repeated measures factors were round (1-4) and achievement level (Basic, Proficient, Advanced), and the between groups factor was the grouping variable (type, group, region, etc.). Statistically significant ($p < .001$) main effects for rounds and level, and statistically significant round-by-level interactions, were consistently found. However, the only statistically significant differences across subgroups were for the Group and Table analyses, which had statistically significant three-way interactions. The eta-squared effect size measure associated with the three-way interaction for Group was .09 and for Table it was .03; thus these effects, though statistically significant, are small. In fact, when Bonferroni corrections were applied to the Table comparisons, no differences were statistically significant at $p < .05$. These findings support the consistency of the cut scores across subgroups of panelists.

Figures summarizing the consistency of cut scores across rounds, broken down by Group membership, are presented in Figures 6 through 11. Figure 6 presents the results for Group (A or B), Figure 7 presents the results for Tables, the results for ethnic groups are presented in Figure 8, panelist type (teacher, non-teacher educator, general public) is presented in Figure 9, sex is presented in Figure 10, and geographic region is presented in Figure 11. In general, these figures illustrate a decrease in variability within panelists groups across rounds, and fair consistency of cut scores across groups.

The greatest variability across groups was seen for the smallest groups (e.g., tables, ethnicity). It is also interesting to note that, in general, the cut scores for *Advanced* changed the least across rounds. This could be due to a ceiling effect caused by the items, or by the panelists' earlier bookmark placements for the Basic and Proficient cut scores. It is also interesting that the Basic cut scores exhibited the largest variability across groups, perhaps due to perceived higher consequences associated with that cut score. With respect to panelist type, it is interesting to note that the "general public" panelists had higher cut scores at Round 1, but they became congruent with the cut scores for the other two types of panelists in subsequent rounds.

Figure 6. Variability in Panelists' Median Cut Scores Across Rounds and Groups

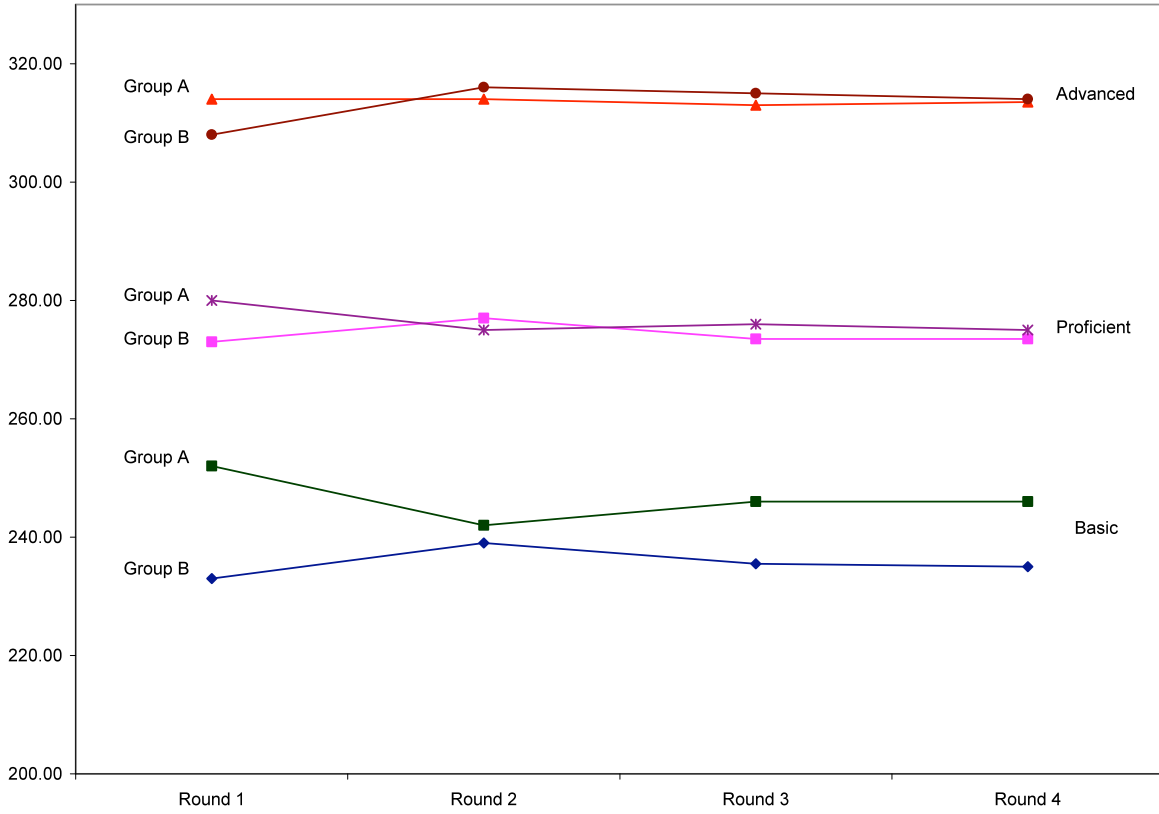


Figure 7. Variability in Panelists' Median Cut Scores Across Rounds and Tables

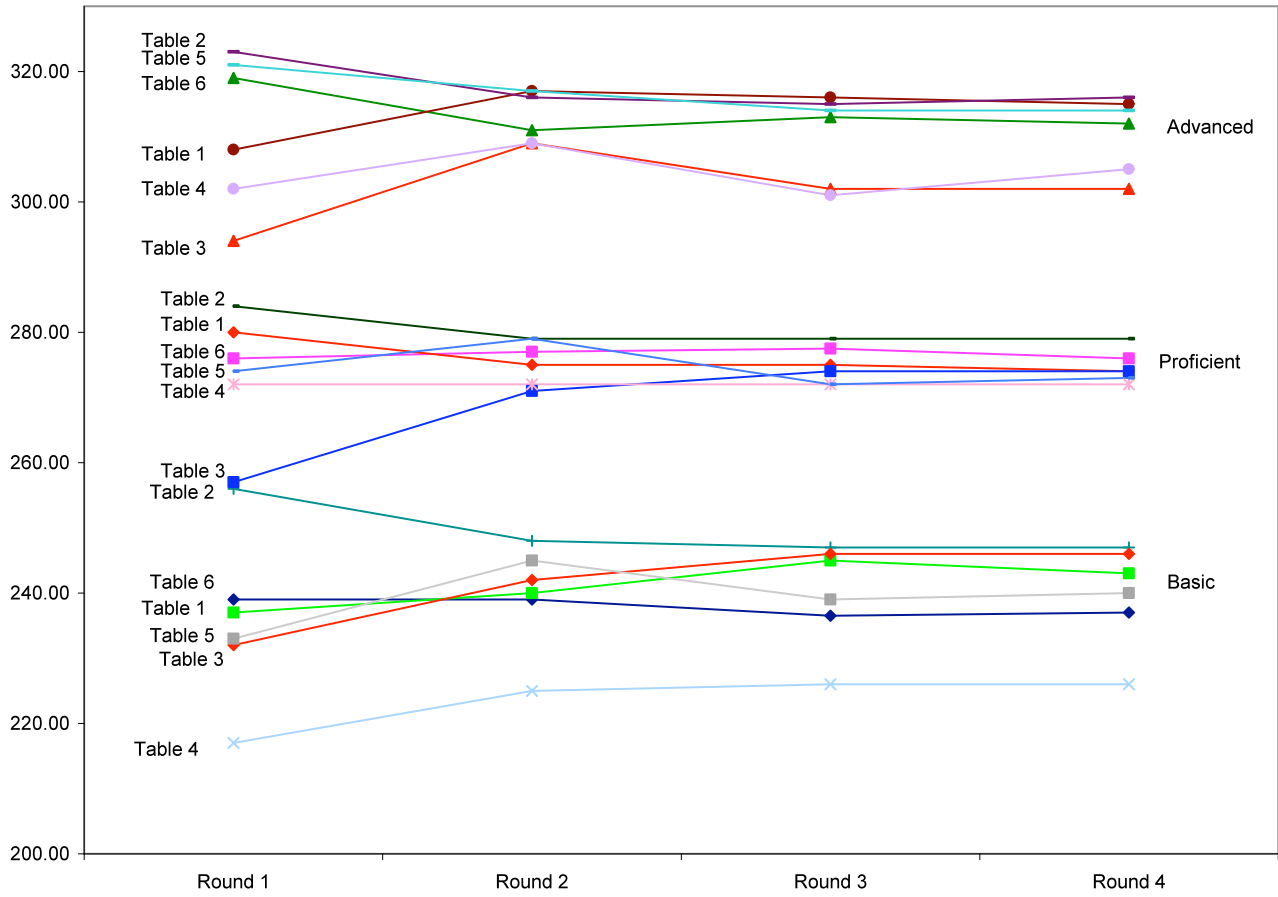


Figure 8. Variability in Panelists' Median Cut Scores Across Rounds and Ethnic Groups

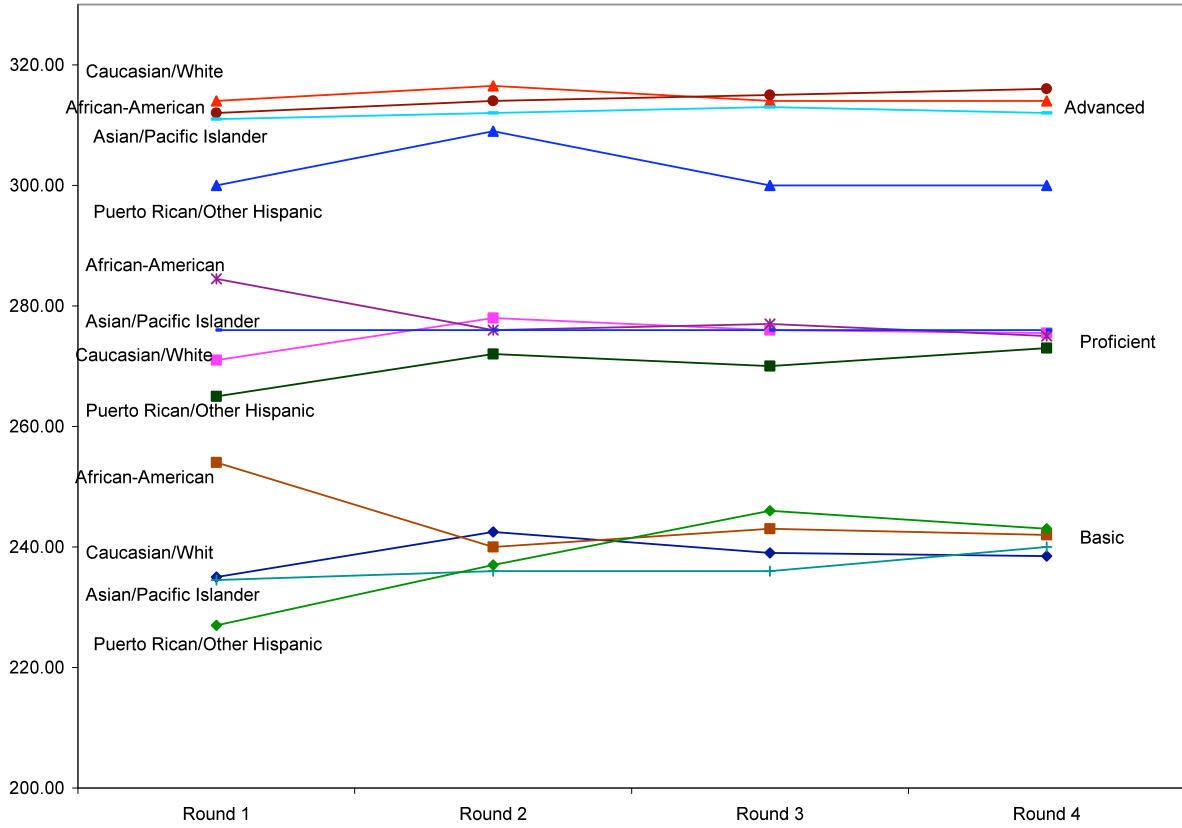


Figure 9. Variability in Panelists' Median Cut Scores Across Rounds by Panelist Type

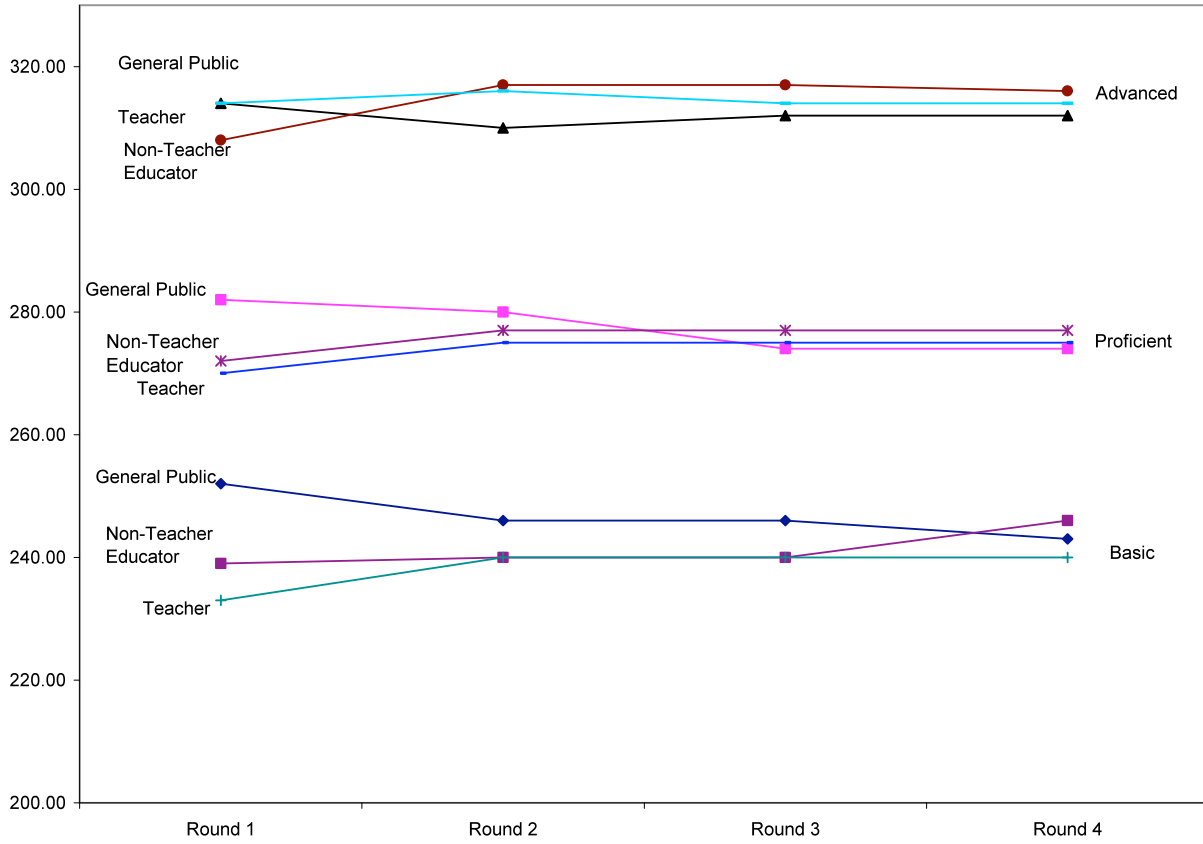


Figure 10. Variability in Panelists' Median Cut Scores Across Rounds by Sex of Panelist

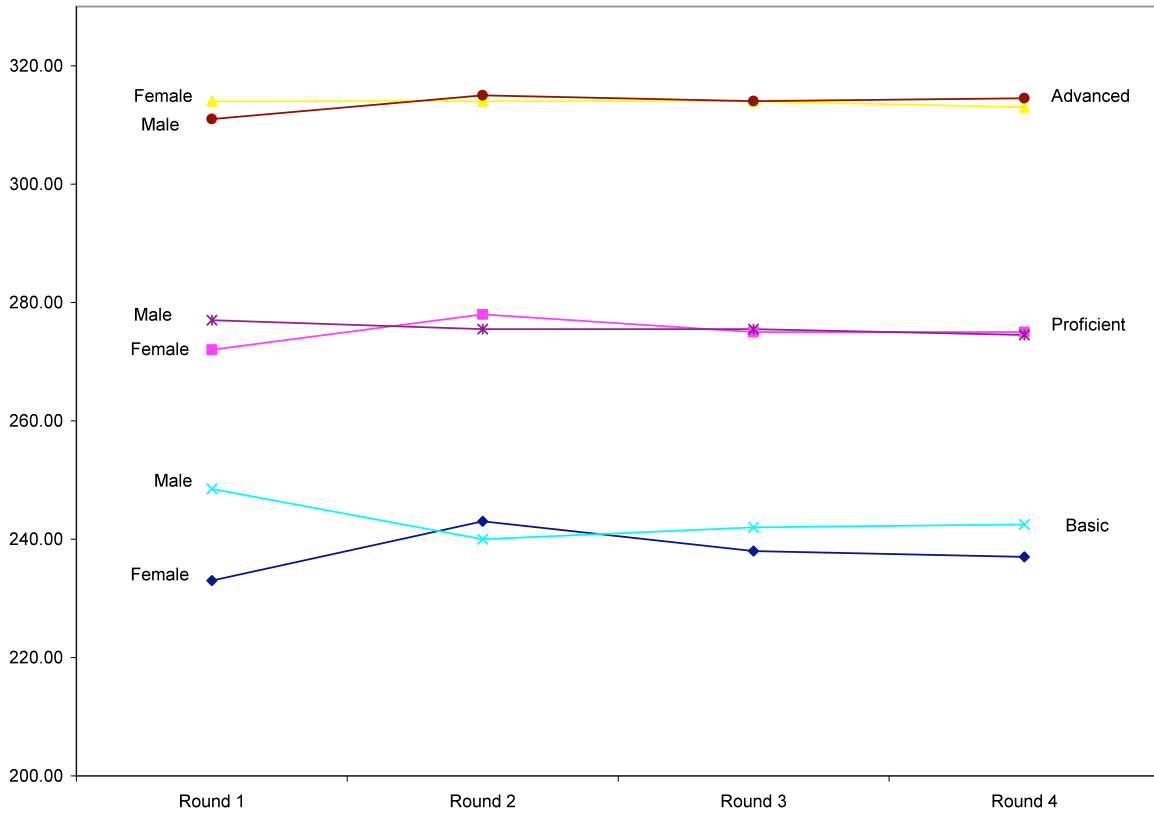
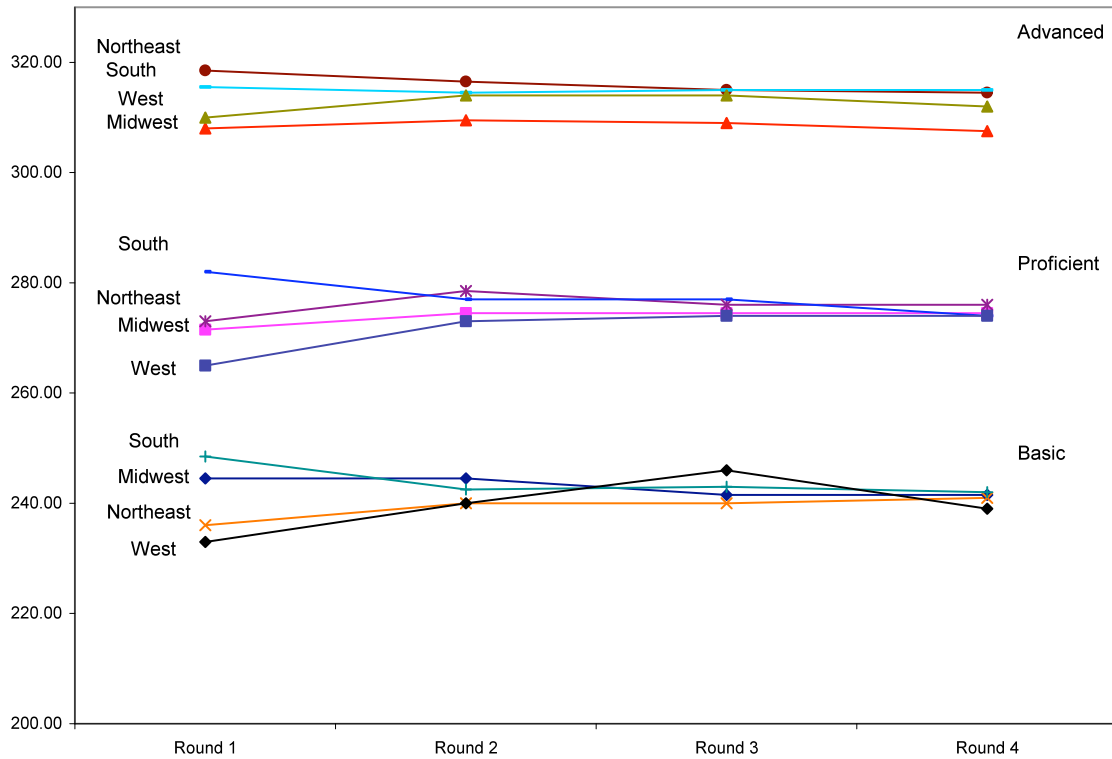


Figure 11. Variability in Panelists' Cut Scores Across Rounds by Geographic Region



Summary of internal evidence

Our estimates of the standard errors associated with the cut scores resulting from the standard setting study found they tended to be under 2.5 points for each achievement level. Given that the standard deviation for the scores on this assessment is about 34 points,¹⁵ the magnitude of error seems small. Furthermore, our analysis of the consistency of the cut scores across various subsets of panelists revealed adequate consistency. Therefore, we conclude the internal evidence for validity of this standard setting is strong.

¹⁵ Information downloaded from <http://nces.ed.gov/nationsreportcard/nde/viewresults.asp> on Feb. 25, 2007.

External Evidence

As described earlier, external validity evidence for standard setting can come from at least two sources. One source is the degree to which different standard-setting methods lead to similar cut scores and examinee classifications when applied to the same test. Another source is when separate tests or other measures of examinee performance are used to classify examinees with respect to their standing on the same or similar construct measured. In this section, we report data relevant to both types of external validity evidence.

Similarity of cut scores across different standard-setting methods

The consistency between the cut scores from the operational Mapmark standard setting (November 2004) and the pilot study Mapmark standard setting (July 2004) was discussed in the previous section with respect to estimating the standard errors of the cut scores (i.e., standard errors of the difference), as part of our evaluation of internal validity evidence. However, this pilot study is also relevant to external validity evidence for these standards because it also involved an application of the item-rating standard-setting method to this assessment. Thus, a comparison of the Mapmark cut scores from the operational study to the item-rating cut scores results from the pilot study provides external validity evidence. The results of this comparison are summarized in Table 18. The *Basic* and *Proficient* cut scores were within one-point of one another. However, the cut score for *Advanced* was about nine points higher for the Mapmark method, which resulted in almost 2 percent fewer students falling into that achievement level.

Table 18. Summary of Consistency of Cut Scores Across Item Rating and Mapmark Methods

Achievement Level	Pilot Study Item Rating Method (July 2004)		Operational Mapmark Method (November 2004)	
	Cut Score	percent At or Above	Cut Score	percent At or Above
Basic	142	61.5	141	62.6
Proficient	177	22.4	176	24.5
Advanced	207	4.0	216	2.3

Grade 12 math performance trends on other assessments

An important source of external validity evidence for standards set on educational tests can be obtained by comparing the achievement level classifications of students with comparable classifications on similar, but independent, assessments of the construct measured. If students were classified into the same or similar achievement levels on two or more assessments measuring the same construct, the validity of the standards on each assessment is supported. Consistency of student classifications across multiple measures is a form of convergent validity, meaning that the standards set on the different assessments converge at the same levels of student performance.

In an ideal external validation of achievement level results, two tests that are designed to measure the same construct would be available, comparable achievement level standards would be set on both tests, and a common group of students would take both tests. In such a situation, the degree to which students were classified into the same achievement levels across tests would provide external evidence that the classifications resulting from the standard setting were reasonable.

Unfortunately, these ideal conditions generally do not exist, and they certainly do not exist for the 2005 Grade 12 NAEP Mathematics assessment. First of all, individual students are not classified into achievement levels on NAEP. Rather, percentages of students scoring within each achievement level are estimated. Second, there is no parallel, national test that measures grade 12 students' mathematics proficiency and classifies students into achievement level categories that are directly comparable to those used by NAEP. Third, different tests usually measure somewhat different constructs, even if they have similar names. Fourth, the performance standards developed by different groups in different contexts are not likely to yield the same results, even when they use the same level (e.g., "proficient"). Therefore, other strategies for the external validation of the standards set on the 2005 Grade 12 NAEP Mathematics assessment were needed.

The achievement level results for the 2005 Grade 12 NAEP Mathematics assessment are presented in Table 19, along with brief descriptions of the achievement levels. Following the recommendations of our Technical Work Group, we sought national test data that could characterize the math proficiencies of 2005 Grade 12 students in a manner that could be related to these achievement categories. Data were available for three national assessment programs that reported results for 2005 high school seniors. These programs were the Advanced Placement (AP) Calculus tests, the ACT Assessment, and the SAT (including the SAT-II subject tests in math). Data from these assessment programs were used because cut scores or benchmarks had been set on these exams, and these benchmarks were considered relevant to at least one of the achievement levels set on the Grade 12 NAEP Mathematics assessment. However, it should be noted that although these testing programs are "national," they are not nationally representative, as are NAEP assessments. Thus, while NAEP assessments represent students of various proficiency levels, these other assessments are more likely to reflect the college-bound senior population.

Table 19. 2005 Grade 12 NAEP Math Achievement Level Descriptions and Results

Achievement Level	Description*	Percent At or Above
Basic	Should be able to solve math problems that require the direct application of concepts and procedures in familiar situations (e.g., perform computations with real numbers and estimate the results of numerical calculations). Should be able to estimate, calculate, and compare measures and identify and compare properties of two- and three-dimensional figures, and solve simple problems using two-dimensional coordinate geometry. Should be able to identify the source of bias in a sample and make inferences from sample results, calculate, interpret, and use measures of central tendency to compute simple probabilities. Understand the use of variables, expressions, and equations to represent unknown quantities and relationships among unknown quantities. Solve problems involving linear relations using tables, graphs, or symbols; and solve linear equations involving one variable.	60.6%
Proficient	Should be able to select strategies to solve problems and integrate concepts and procedures. Should be able to interpret an argument, justify a mathematical process, and make comparisons dealing with a wide variety of math tasks. Perform calculations involving similar figures including right triangle trigonometry. Understand and apply properties of geometric figures and relationships between figures in two and three dimensions. Select and use appropriate units of measure as they apply formulas to solve problems. Use measures of central tendency and variability of distributions to make decisions and predictions; calculate combinations and permutations to solve problems, and understand the use of the normal distribution to describe real-world situations. Identify, manipulate, graph, and apply linear, quadratic, exponential, and inverse functions $y = k/x$; solve routine and non-routine problems involving functions expressed in algebraic, verbal, tabular, and graphical forms; solve quadratic and rational equations in one variable and solve systems of linear equations.	23.0%
Advanced	Should demonstrate in-depth knowledge of the mathematical concepts and procedures represented in the framework. Integrate knowledge to solve complex problems and justify and explain their thinking. Analyze, make and justify mathematical arguments, and communicate their ideas clearly. Describe the intersections of geometric figures in two and three dimensions, and use vectors to represent velocity and direction. Describe impact of linear transformations and outliers on measures of central tendency and variability; analyze predictions based on multiple data sets; and apply probability and statistical reasoning in more complex problems. Solve or interpret systems of inequalities; and formulate a model for a complex situation (e.g., exponential growth and decay) and make inferences for predictions using the mathematical model.	2.2%

*Adapted from Appendix A of ACT (2005b).

Because we did not have data for a common group of students who took these tests and the 2005 Grade 12 NAEP Mathematics assessment, our analyses were limited to comparing the percentages of students reported in the NAEP achievement level results (i.e., the last column of Table 19) with the percentages of students in specific categories derived from the external testing programs. For each testing program, we identified a benchmark performance criterion that could be considered more-or-less congruent with a specific NAEP achievement level. A description of these criteria and our rationales for selecting them follow. Given that these external tests differed from the NAEP assessment in several important ways (e.g., test content, student motivation conditions or stakes, substantially different standard-setting methods), they must not be taken as unequivocal estimates of what the NAEP achievement results should be. However, they may be useful for gauging whether the NAEP achievement level results are within reason, given the performance of 2005 grade 12 seniors on other mathematics assessments.

As described below, the data from the external testing programs were used to estimate percentages of 2005 grade 12 students in the U.S. at or above specific achievement level categories. To arrive at those estimates, the total number of 2005 grade 12 students was needed. Enrollment data are gathered by the U.S. Department of Education, and are broken down by jurisdiction and public versus private school. Using the public school enrollment for the 50 states and Washington D.C., and an estimate of the national private school enrollment, the total number of 2005 grade 12 students was estimated to be 3,402,883.¹⁶

AP Calculus exams

The AP testing program has two math exams: Calculus AB and Calculus BC. The AB exam assesses course content typically covered in advanced precalculus and introductory calculus courses. The BC exam covers material consistent with calculus taught at the postsecondary level, such as that required for engineering majors.¹⁷ Standards are set on the AP exams so that students can earn credit for math courses offered in college or be placed out of specific college courses. AP exam results for students are reported on a five-point scale where 5=extremely well qualified, 4=well qualified, 3=qualified, 2=possibly qualified, and 1=no recommendation. Most colleges award course credit or placement to students obtaining a score of 3 or higher (College Board, 2006a). Therefore, we used a criterion of a score of 3 or better on either of these two AP exams to define students as “advanced.”¹⁸ Although we realize the NAEP and AP Calculus tests are measuring very different constructs, our rationale for this criterion was that if students were proficient in calculus and would receive college credit for calculus courses, they would be advanced in mathematics in general, and should be able to perform at the “advanced” level with respect to the math domain measured on the Grade 12 NAEP exam.

Data were available from the College Board on the numbers of 2005 high school seniors who earned scores of 3 or above on each of these exams. These numbers were used to estimate the percentage of 2005 high school seniors who earned scores of 3 or higher on these tests. These proportions were compared to the percentage of grade 12 students classified as *Advanced* on the 2005 Grade 12 NAEP Math assessment (i.e., 2.2 percent).

The numbers and percentages of 2005 grade 12 students earning scores of 3 or higher on the AP calculus exams are presented in Table 20. Using the national enrollment figure of 3,402,883, just over 2 percent of the national group surpassed this criterion on the AB exam and just under 1 percent surpassed this criterion on the BC exam. About 19 percent of the students

¹⁶ Other jurisdictions not participating in the 2005 grade 12 NAEP data are excluded from this total (i.e., Puerto Rico, American Samoa, Guam, Northern Marianas, Department of Defense schools, and the Virgin Islands). Private school enrollment was based on data from the 2003–2004 school year since data for 2004–2005 were not yet available for private schools. Source: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (<http://nces.ed.gov/pubs2007/2007309.pdf>).

¹⁷ John Dossey, personal communication, Aug. 15, 2006.

¹⁸ This is the same criterion used by Shepard, Glaser, and Linn (1993) in their external analysis of the 1992 Grade 12 NAEP Math achievement levels.

who took the BC exam previously took the AB exam. Therefore, we multiplied the 31,453 students who earned a score of 3 or higher on the BC exam by .81 to arrive at the projected number of BC exam students who surpassed the criterion, but did not take the AB exam. We then added that number to the 78,171 who surpassed the criterion on the AB exam. Using these data, approximately 3% of the 2005 grade 12 student population earned a score of 3 or higher on either test. This percentage is slightly higher than the 2.2% of students classified as Advanced on NAEP.

Table 20. AP Calculus Exam Results for 2005 High School Seniors

	AB	BC
# Taken	142,091	39,910
# scoring ≥ 3	78,171	31,453
% of test takers scoring ≥ 3	55%	79%
% 2005 HS Seniors scoring ≥ 3	2.3%	0.9%

ACT assessment

The ACT is a comprehensive assessment used to help college admissions officers make admissions decisions. According to the ACT Web site, it measures both high school achievement and ability to do college-level work. There are four tests within the ACT assessment: English, Mathematics, Reading, and Science. A composite score for these four tests is also reported, which is the average of the four subtest scores rounded to the nearest integer. The score scale for each subject test and the composite ranges from 1 to 36.

Based on research regarding what college students need to know and be able to do to succeed in college, ACT established “college readiness benchmarks.” These benchmarks are cut scores on the ACT subtest score scales that indicate readiness for college-level work. The cut score for the ACT Math test is 22, and is referenced to a grade of B or higher on a typical (non-remedial) first-year course in college algebra that is taken by a large proportion of first-year students (Allen and Sconing, 2005).¹⁹ We used this benchmark as an external criterion for evaluating the “proficient” achievement level on the NAEP exam. To use this criterion, we identified the number of ACT test-takers from the 2005 grade 12 class who met or surpassed the benchmark. However, given that the ACT is one of two widely accepted college entrance exams (the SAT being the other), we could not simply divide the number of students meeting this benchmark by the number of 2005 grade 12 students in the U.S. Therefore, it was necessary to estimate the number of students taking both the ACT and SAT who would meet this benchmark.

We used two approaches to estimate the number and percentage of students across the two tests who would meet the ACT college readiness criterion in math. For the first estimate, we took the percentage of seniors with ACT Math test scores greater than or equal to 22, and multiplied it by the number of students who took the SAT math test. This gave us the number of students who had an SAT percentile rank greater than or equal to the percentile rank associated with an ACT math score of 22. We then added the numbers of students on each test who met the criterion.²⁰ For the second estimate, we used an ACT-SAT (composite) score concordance table to find the SAT score that corresponded to an ACT score of 22 (College Board, 2006b). We then calculated the number of SAT students at or above this converted score. It should be noted that neither approach resolves the problem that many students took both tests, and so the estimates are likely to overestimate the percentage of “proficient” 2005 grade 12 students.

¹⁹ Allen and Sconing derived this benchmark 22 by finding the ACT score in each course in each college studied that was associated with a .50 probability of earning a grade of “B” or higher in the course, and then taking the median of these ACT scores across colleges.

²⁰ By adding these numbers together, we get an estimate of the total number of students who met the standard as if they had all taken the ACT (rather than either the ACT or SAT).

The results from these estimates are summarized in Table 21. The two different estimates for the national population were within 5 percentage points of each other and suggest that about 34 percent of the 2005 grade 12 students were proficient in math according to our ACT criterion; however, given that the unknown overlap between the two groups of test takers would inflate the percentage to some degree, and the fact that neither test is a representative sample of 2005 seniors, this result is hard to interpret with respect to the 23 percent found to be *Proficient* or above on NAEP.

Table 21. 2005 ACT Results: High School Seniors

# Test Takers	1,186,251
# ≥ 22	481,273
% ≥ 22	40.57
% of 2005 HS Seniors	14.14%
% 2005 ACT/SAT seniors: estimate 1	31.7%
% 2005 ACT/SAT seniors: estimate 2	36.4%

Source: *2005 National Score Report: Data Tables* (ACT, 2007). Downloaded on Jan. 8, 2007 from <http://www.act.org/news/data/05/data.html>.

SAT-II Math subject tests

The SAT testing program consists of the SAT Reasoning Test and SAT Subject Tests. The Reasoning Test consists of three sections: Critical Reading, Math, and Writing. There are two subject tests related to math—Mathematics Level 1 and Mathematics Level 2 (formerly called Mathematics IC and IIC respectively). The Mathematics Level 1 subject test is a “broad survey test intended for students who have taken three years of college-preparatory mathematics, including two years of algebra and one year of geometry.”²¹ The Mathematics Level 2 subject test is “intended for students who have taken college-preparatory mathematics for more than three years, including two years of algebra, one year of geometry, and elementary functions (precalculus) and/or trigonometry.”²² Like the ACT, AP, and SAT tests, not all high school seniors take these exams, but many colleges and universities use them for admission, course credit, or course waivers. The California State University system, for example, allows students to meet its “entry level mathematics requirement” with an SAT math reasoning score of 550 or above, or scores on the Mathematics Level 1 or Level 2 tests of 550 or above. As another example, the University of Texas at Austin gives students credit in a lower-level course (Elementary Functions and Coordinate Geometry) to all students who earn a score of 560 or above on the Math Level 1 exam or 530 or above on the Math Level 2 exam. Given that data were available in only 50-point intervals on these subject tests for 2005 high school seniors, we calculated the number of seniors who scored at or above 550 on each test. That total was 190,298, which represents only 5.6 percent of the 2005 high school senior population. That percentage is far below the 23 percent of students classified as *Proficient* or above on the NAEP exam. The large difference could be due to the fact that the students who take these tests are a nonrepresentative subset of the high school senior population and so they do not represent the population with respect to size or proficiency.

²¹ Downloaded from http://www.collegeboard.com/student/testing/sat/lc_two/math1c/math1c.html?math1c, Jan. 26, 2007.

²² Downloaded from http://www.collegeboard.com/student/testing/sat/lc_two/math2c/math2c.html?math2c, Jan. 26, 2007.

NAEP high school transcript study

Recently, the National Center for Education Statistics (NCES) completed a high school transcript study for subsets of 2005 high school students who participated in the NAEP assessment (NCES, 2007). In this study, they related course-taking patterns and high school grades to performance on the 2005 NAEP Mathematics and Science assessments. In general, the results showed that more rigorous course-taking patterns and higher grades in high school were associated with higher NAEP scores. Although the focus of the study was not a validation of the NAEP achievement levels, there are several findings that are relevant to the present evaluation from the perspective of external validity evidence.

One relevant finding was that graduates who had not taken a very rigorous set of math courses (i.e., geometry or below) performed, on average, below the NAEP *Basic* achievement level. At the other extreme, graduates who took and did well in calculus, on average, performed at the *Proficient* level. Unfortunately, the report did not indicate the percentages of students who passed their calculus course who fell into the various achievement levels. However, the report did provide profiles of students who were classified as *Advanced* and *Below Basic* on the NAEP assessment. Most students who were classified as *Advanced* (85 percent) had taken calculus, and most (85 percent) had math grade point averages in the top quartile. The students classified as *Below Basic* had a much less rigorous mathematics course profile, with only 1 percent taking calculus and only 7 percent falling into the upper quartile of math grade point average. Interestingly, the vast majority of students classified as *Advanced* (86 percent) took an AP or International Baccalaureate math course, as compared to only 1 percent of the *Below Basic* students.

Summary of external data

Several caveats were described with respect to the utility of external data for evaluating the 2005 Grade 12 NAEP Math achievement level results. For the comparisons to external assessment programs, all of the tests studied involved self-selected samples of college-bound seniors, rather than a nationally representative sample of students who are tested by NAEP. There are obvious motivational differences across students taking an exam for college admission or course credit and students taking an exam that has no direct benefit to them. Such differences in motivation are certain to affect students' performance on a test. In addition, the benchmarks set for college readiness on the ACT, or to meet college entry requirements in math, are not designed to be congruent with the NAEP achievement levels and these benchmarks were set in ways very different from how standards were set on NAEP. Another important caveat associated with these comparisons is that even though all the tests we compared were measuring mathematics proficiency, they each assess different aspects of mathematics and so the content tested differs across these exams. Thus, these external data are far from perfect evaluation criteria.

The utility of the ACT and SAT data is particularly debatable due to the fact that we could not estimate the overlap between 2005 seniors who took both the ACT and SAT. However, the fact that the percentage of students who were at or above *Proficient* on the NAEP exam (23 percent) fell between the 5.6 percent of seniors who surpassed our SAT-II math criterion and the roughly 34 percent who surpassed our SAT-I/ACT criterion, suggests the NAEP achievement level results for *Proficient* are not outside the range of possibility.

The AP calculus data are more defensible for the purpose of evaluating the *Advanced* NAEP achievement level results because there is no other national test students take to earn college credit and students who can handle calculus can obviously handle challenging mathematics content. If we assume students who scored 3 or better on these AP exams truly should be considered *Advanced* as defined by NAGB (see Table 19), we can conclude that the

percentage of 2005 seniors who earned scores of 3 or better on AP calculus tests (about 3 percent after considering the overlap of the AB and BC exams) was slightly higher than the percentage of students estimated to be *Advanced* on the NAEP exam (2.2 percent).

This finding implies that the NAEP results for the *Advanced* achievement level are not obviously too high or too low, although we note that all students who might meet the NAEP definition of *Advanced* may not have taken an AP Calculus test. The larger proportion of students who achieved AP Calculus scores of 3 may suggest that about 1 percent more students should have been classified as *Advanced* on this NAEP assessment. That conclusion is subject to the caveats mentioned earlier. In any case, the difference between the NAEP and AP percentages appears too small to conclude that the NAEP standard is entirely too rigorous.

Consistency of Cut Score Results Across Field Test and 2005 Assessment

In addition to external information regarding students' mathematics proficiency, we also looked at the consistency of information between the field test conducted for the 2005 Grade 12 Math test and the actual assessment. The standard setting study for the 2005 Grade 12 NAEP Math test relied heavily on the 2004 Grade 12 NAEP Math field test in several ways. The item parameters that were used to compute the response probabilities for each item were based on this field test and so the ordering of the items within the OIB was dependent on these data. The domain characteristic curves, expected percentages correct, and consequences data were also computed from these data. Therefore, an important evaluation area is the degree to which these data were representative of the results of the 2005 Grade 12 Math assessment. The achievement level results from the 2004 pilot study and the 2005 assessment are reported in Table 22. It should be noted that these results pertain to two different cohorts of students (seniors in 2004 and 2005) and so they should not be expected to be identical. Given that they are close, it is reasonable to conclude the 2004 data on which the standard-setting study was based were sufficiently representative for the purposes of creating the OIB and providing expected performance and consequences data to the panelists.

Table 22. Comparison of Achievement Level Results: 2004 Pilot and 2005 Assessment

Achievement Level	% At or Above	
	2004 Pilot	2005 Assessment
Basic	62.6	60.6
Proficient	24.5	23.0
Advanced	2.3	2.2

Summary and Conclusions

Our evaluation of the establishment of the achievement level standards on the 2005 Grade 12 NAEP Mathematics assessment began with a review of criteria for evaluating standard-setting procedures on educational tests. The criteria we used borrowed heavily from the *Standards for Educational and Psychological Testing* (AERA et al., 1999) and Kane (1994, 2001) who proposed three sources of validity evidence for standard setting—procedural, internal, and external.

Our evaluation used several data sources. These sources included published literature, documentation associated with the standard-setting activities, observation of the standard-setting process, and analysis of the standard-setting and item data.

With respect to procedural evidence, we found that the standard-setting process put in place by ACT had adequate procedural validity. The choice of standard-setting method was based on empirical research and was designed to address criticisms of prior NAEP standard-setting processes. The training of the panelists was comprehensive and the panelists generally reviewed the process favorably. The study was extremely well documented.

With respect to internal evidence, we could not find systematic evidence that the standards varied widely across different types of panelists, methods, or replications. Using three different approaches to estimating the standard errors of the cut scores, we found them to be very small (less than 2.5 points), relative to the standard deviation of the test score scale (34 points).

Our evaluation of external evidence was limited. We were unable to locate external criteria that would be commensurate with this NAEP Mathematics assessment, let alone samples of students who would be similarly representative of 2005 grade 12 students. We were able to conduct some analyses of 2005 grade 12 students' performance on independent measures of mathematics proficiency and found no evidence that the achievement level results for the 2005 Grade 12 NAEP Mathematics assessment were beyond reason. Our analysis of AP data suggests that the percentage of *Advanced* students might be slightly higher (about 1 percent) than the percentage reported on this NAEP assessment. The findings from the NAEP High School Transcript Study (NCES, 2007) illustrated that students taking calculus courses, on average, were classified as *Proficient*, but those data should be mined further to determine the percentage of students who did well in calculus who were classified as *Advanced*. In general, the external data do not explicitly support the validity of these NAEP standards, nor did they refute the standards.

In the introductory section of this report, we described several specific criteria that could be used to evaluate a standard-setting study. These criteria are repeated in Table 23, along with our conclusion regarding whether the 2005 Grade 12 NAEP math standard setting satisfactorily met each criterion. As is evident from the table, the study fared very well with respect to our evaluation criteria. Two criteria could not be evaluated: (a) the consistency of standards across item formats (i.e., multiple-choice and free-response items), and (b) comparing students from different achievement levels (identified using external data) on their NAEP performance.

Under the Mapmark method, panelists do not rate all items and so it was not possible to derive separate cut scores for each panelist by item format. However, the idea that standards should be consistent across sub-scores based on item formats is controversial. Hambleton et al. (2000) criticized this criterion when commenting on the prior evaluation of NAEP standard setting (Pellegrino et al., 1999). They argued that performance standard differences across item types could merely reflect the fact that the different item formats are measuring different aspects of student achievement. In making this argument, they cited Kane (1995) who stated:

One could assume that the apparent difference in the quality of student performance between dichotomous items and extended response items. . . [is] real and that students are meeting judges' expectations, on recognition tasks, but not doing as well, relative to judges expectations, on tasks that require an extended response. . . . The fact is that many scholars believe extended response items tap aspects of student achievement not directly assessed by multiple-choice items. (Kane, 1995, p. 125, cited in Hambleton et al., 2000)

They also hypothesized that such differences could be due to issues in scaling dimensionally distinct item types onto a common scale (Brennan, 1998, cited in Hambleton et al., 2000).

Table 23. Summary of Results Regarding Evaluation Criteria

Criterion	Evidence	Criterion met?
Care in selecting participants	ACT (2005b)	Yes
Justification of standard setting method(s)	ACT (2005a, 2005b, 2005c), Schulz et al. (2005)	Yes
Panelist training	Observations, Survey data	Yes
Clarity of goals/tasks	Observations, Survey data	Yes
Appropriate data collection	ACT (2005b), Observations	Yes
Proper implementation	ACT (2005b), Observations	Yes
Panelist confidence	ACT (2005b), Survey data	Yes
Sufficient documentation	ACT (2005a, 2005b, 2005c)	Yes
Sufficient inter-panelist consistency	Analysis of panelist and pilot study data	Yes
Decreasing variability across rounds	Analysis of panelist data	Yes
Small standard error of cut score (consistency within method)	Analysis of panelist data	Yes
Consistency across independent panels	Analysis of panelist and pilot study data	Yes
Consistency across panelist subgroups	Analysis of panelist data	Yes
Analysis of borderline students performance on specific items	ACT (2005b)	Yes
Consistency across standard setting methods	Analysis of panelist and pilot study data	Yes
Mean differences across proficiency groups on external criteria	NAEP High School Transcript Study (NCES, 2007).	Yes
Consistency across other student classification data	Comparison of achievement level results with ACT, AP, and SAT data	To some extent
Consistency across item formats		Not applicable
Reasonableness	All of the above	Yes

The criterion of comparing test score performance across groups which are known to be from different achievement levels is a particular type of external validity evidence, but, again, data on such groups were not available for analysis. Such analyses would be possible only if groups of

students from different achievement levels could be identified independently of the NAEP assessment (perhaps based on success in rigorous courses defined in the NAEP transcript study), and then differences in mean NAEP scores across these groups were calculated to determine how well NAEP performance differentiated these student groups. Perhaps future research could use some consensus process to identify such students and administer NAEP test items to them for external validation purposes. Without that type of special study, it is not possible to evaluate that criterion.

Criticism of Earlier NAEP Standard Setting and Its Relevance to the Current Study

As described in the introduction to this report, there have been at least four major criticisms of prior standard-setting activities for NAEP. These criticisms are: (a) the item-rating (Angoff) tasks are too cognitively complex for standard-setting panelists to successfully complete (Pellegrino et al., 1999; Shepard et al., 1993; U.S. General Accounting Office, 1993); (b) there is no evidence that students within an achievement level possess the knowledge and skills implied in the achievement level description (Linn, 1998; Pellegrino et al., 1999; Shepard et al., 1993); (c) there is no external evidence to corroborate student classifications (Linn, 1998; Pellegrino et al., 1999; Shepard et al., 1993); and (d) there are inconsistencies in cut scores when derived separately from multiple-choice and free-response items (Linn, 1998; Pellegrino et al., 1999; Shepard et al., 1993). These criticisms were addressed previously, but we revisit them here in light of the evaluation results.

Cognitive complexity of standard setting tasks

As mentioned earlier, the criticism of cognitive complexity with respect to setting standards on NAEP tests claims that asking panelists to judge the probability of success on an item for “borderline” students is asking for more than they can successfully process from a cognitive perspective, particularly when there are multiple standards to set. That is, the standard-setting task involved in Angoff and other item rating methods is too complex for panelists. Although several psychometricians refuted this criticism, including the 11-member panel that advised ACT on the 1996–2000 NAEP standard-setting activities (i.e., Hambleton et al., 2000), a natural question is whether the Mapmark method, as implemented here, requires more or less cognitive complexity than item-rating methods.

The tasks involved in the Mapmark method appear to be cognitively complex even though the procedure was designed to present tasks to panelists that are more familiar to them, relative to tasks associated with item-rating methods. The cognitive complexity criticism of item-rating methods stems from the fact that panelists are required to make conditional probability judgments for each item. Mapmark, and its predecessor Bookmark, requires panelists to focus on the content of the test items and consider the knowledge and skills measured by each item prior to making a bookmark placement. The panelists do not make conditional probability judgments for each item, and so it is considered less complex. As described by ACT (2005b), “By ordering items according [to] the RP criterion and student performance data in the KSA review, the role of probability judgment in the panelists’ task is minimized and panelists are free to concentrate more on test content, on what higher levels of performance on the test mean, and on mapping the achievement level descriptions to actual levels of student performance” (p. 12).

We do not necessarily concur with the claim that item-rating standard setting methods, such as the Angoff method, are overly cognitively complex. We also believe the tasks required of panelists participating in this Mapmark study were very complex. However, we acknowledge that the theory underlying the Mapmark method proposes reduced complexity, and we were

impressed by the comprehensive training of panelists throughout the standard-setting study and the degree to which the facilitators attended to them throughout the process. Given our observations of the standard-setting process, and the results of the panelist surveys (from the current study, the field trials, and the pilot study) we conclude that the tasks presented to the panelists in this study were not too complex for them to successfully complete. Thus, we conclude that the prior criticism of NAEP standard-setting tasks as too cognitively complex does not apply to the 2005 Grade 12 NAEP Mathematics assessment. However, we also noted that almost all of the panelists participating in this study had strong math backgrounds, which probably helped them understand and use the substantial statistical information that was presented throughout the process (e.g., domain scores, percent correct tables, etc.). The degree to which these tasks could be successfully comprehended and implemented by panelists from other subject areas remains an open question.

Confirmation of achievement level descriptions

Reporting test results by using achievement levels involves careful development of the achievement level descriptions that describe what students classified into each level know and are able to do. Standards may be invalid if the students within a particular achievement level have low probabilities of success on items that the achievement level descriptions imply they can master. Linn (1998), among others, claimed that some standards set on NAEP exams suffered from this problem, which he described as “discrepancies between descriptions of achievement levels with their associated exemplar items and the locations of cut scores on the scale” (p. 25).

The Mapmark method used on the 2005 Grade 12 NAEP math exam addresses this criticism by having panelists focus on the performance of students within each achievement level, and at the border between achievement levels, on groups of items organized by specific skills—skills typically articulated in the performance level descriptions. In fact, the final phase of the Mapmark standard setting was the selection of exemplar items that best represent what students within an achievement level know and can do.²³ The second, third, and fourth rounds of the Mapmark presented data to panelists regarding the performance of students within each achievement level (e.g., domain score charts, see Figure 5) and the performance of borderline students (e.g., percent correct tables, see Figure 3). Thus, one clear benefit of switching from an item-rating method to the Mapmark method is that the Mapmark method, if implemented properly, specifically addresses the criticism of a disconnect between the achievement level descriptions and the performance of students within each achievement level. We noted that the current implementation of the Mapmark revisited the achievement level descriptions at several points during the process.

Corroborating external evidence

Another criticism of NAEP achievement levels is that there are no external data to confirm the percentages of students falling into the different achievement levels. The present evaluation found no data to refute this criticism. Rather, our efforts to gather such data confirm the difficulty in obtaining them. To properly gather external validity data, special studies are probably needed, and are recommended for future NAEP standard setting activities. For example, teachers could be trained to understand NAEP achievement level descriptions and then identify students who are clearly within certain achievement levels. The students could then be given a NAEP assessment (perhaps using all or most of the blocks rather than the one-hour’s

²³ Although the identification of potential exemplar items for each achievement level was part of the standard-setting meeting, it is essentially subsequent to the standard-setting process and so it was not a focus of our evaluation. In our opinion, that process also had high procedural validity.

worth of items used in the typical NAEP assessment). This type of study would allow for; (a) evaluating mean score differences on NAEP across these student groups, and (b) calculating of classification consistency across the NAEP achievement level results for each student and his or her teacher classifications. Without such control over student proficiency and test content, external evaluations based on extant data, such as ours in the present study and those conducted by Shepard et al. (1993) will not be convincing with respect to confirming or refuting NAEP results, except in the most extreme circumstances.

One other way in which external data could be used to evaluate NAEP achievement levels is to gather data from grade 12 students who take NAEP assessments on whether they took advanced placement tests (or other standardized, criterion-referenced tests), and if so, what their scores were. Using NAEP's plausible values methodology, achievement level results could be reported for specific student groups (e.g., students with AP scores ≥ 3) as a means for checking these results against expectations for specific groups of students. For example, if reliable data could be gathered on whether grade 12 students had AP scores of 3 or greater in a NAEP subject area, the achievement level results for that group could be used to check the reasonableness of the standards. It could be hypothesized that the vast majority of that group would be classified as "Proficient" or "Advanced," assuming it could be argued that similar constructs were measured. Similar analyses could be applied to the groups of students with different course-taking patterns identified in the NAEP High School Transcript Study. The High School Transcript Study illustrates the types of data that can be gathered from both operational and pilot NAEP assessments. Such data could be considered when setting the final cut scores for the NAEP achievement levels.

Consistency of cut scores over item formats

The idea that cut scores should be consistent across subsets of items defined by item format was discussed earlier and rejected by some scholars as a valid criterion for evaluating the NAEP achievement levels. We concur with the rejection of the criterion. Different item formats are used on NAEP to measure different skills, and they appear in different proportions. Thus, sub-scores based on item format are likely to have different levels of measurement precision and measure distinct aspects of the construct measured. Such comparisons may, however, be illuminating for describing what students at different achievement levels know and are able to do when presented with different tasks related to a common subject domain.

There are some subsets of NAEP items that are relevant for separate analysis, because NAEP creates (and equates) sub-scales as a preliminary scaling step before forming the composite score scale. However, NAEP sub-scales are highly correlated and are not defined by item format. In fact, sub-scale scores, are not even reported for many NAEP assessments, such as the 2005 Grade 12 Mathematics assessment.

Item-rating standard-setting methods, such as the Angoff method, allow for dissecting panelists' data to derive separate standards by item format. Such analyses may be informative in some circumstances, but when differences are seen, we do not believe it should automatically be taken as evidence of a problem with the standard setting. With the Mapmark method, separate standards cannot be derived for each item format. Thus, evaluation of the consistency of standards across item formats is not possible when this method is used.

Limitations of the Mapmark Method

As is evident from our conclusions (summarized in Table 23), we believe the process for setting standards on the 2005 NAEP Math assessment was appropriate for the purpose of reporting achievement level results. However, there are some limitations of the Mapmark method that warrant discussion, especially given that this method may be used on future NAEP exams.

One potential limitation of the method is that successful implementation may not generalize outside of subject areas in which panelists have strong quantitative skills. Panelists in the current study reviewed complex statistical information such as item maps, subscale item maps, domain item maps, item characteristic curves, domain characteristic curves, response probabilities, and consequences data. Although the consequences data are likely to be easily interpreted by nonmathematical panelists, the other data require a fair degree of mathematics proficiency. Application of Mapmark to other subject areas may warrant explicit selection of panelists with sufficient mathematics proficiency to absorb the complex data displays, in addition to the other criteria used for selection. The degree to which nonmath panelists understand the information distributed in a NAEP Mapmark study should also be evaluated through pilot studies.

Another limitation of the Mapmark method is that because it is so new, there are probably only a handful of people who know it well and could successfully implement it. Furthermore, the method is incredibly resource intensive. Thousands of pages of material were required and the preparation time must have been substantial. For example, there were 469 PowerPoint slides used throughout the three-and-a-half-day meeting. Thus, although we applaud the Mapmark implementation on the 2005 Grade 12 NAEP Mathematics assessment, we realize future successful implementations of the method need to be similarly intense from a resource perspective.

Limitations of the Evaluation

In addition to the limitations associated with the Mapmark standard-setting method, it is also important for us to point out the limitations of our evaluation. Although our evaluation was comprehensive and used several data sources, it certainly was not exhaustive. Our observations confirmed much of what was contained in the ACT standard-setting reports, but we did not observe the field test, pilot studies, or the meetings at which it was decided to switch from the item-rating to the Mapmark method. There are other evaluation activities that could have been conducted such as testing students who differ with respect to mathematics achievement to discover the achievement levels into which they would be classified by this NAEP assessment. We could have also replicated the Mapmark procedure, or implemented a different method altogether, to look at the consistency of cut scores across an independent replication or method. However, given the work done by ACT and the resources for this evaluation, these activities were seen as outside the scope of this study.

Based on our analysis of the procedural, internal, and external evidence pertaining to the validity of the process of setting achievement level standards on the 2005 Grade 12 NAEP Mathematics assessment, we conclude that the procedure was sound, followed recommendations for best practices in the area of standard setting, and involved multiple quality control checks to support the defensibility of the process. The validity of any type of test score interpretation is not something that can be unequivocally established. However, the multiple sources of validity evidence we analyzed for this exam lead us to conclude the standards set on the 2005 Grade 12

NAEP Mathematics assessment are valid for the purposes of describing the performance of 2005 Grade 12 students with respect to the NAEP achievement level descriptors.

Recommendations

Based on our evaluation, we have six recommendations for future standard-setting activities on NAEP assessments.

1. Pilot studies should continue to be conducted whenever changes to NAEP standard-setting processes are proposed. These studies should gather data on panelists' comprehension of their tasks and of the information presented to them (as was done in the 2005 study reviewed here).
2. If Mapmark is proposed to be used to set standards on other NAEP assessments, the degree to which panelists in these other domains can comprehend and use the complex statistical information presented should be confirmed. One group of particular interest here would be the panelists who represent the public (about 30 percent of the panelists).
3. When standard-setting procedures involve bookmark-like judgments, the distribution of items along the score scale should be investigated to ensure there are sufficient numbers of items within the cut score regions of the score scale.
4. Independent data should be used to confirm that students of different achievement levels fall into the expected NAEP achievement level. Focused studies would be needed to gather independent measures of the expected NAEP achievement levels into which these students would fall, and systematic sampling procedures should be used to select sufficient numbers of students who would fall into each achievement level.
5. If Mapmark were used to set standards on a future NAEP assessment, it should be implemented in a fashion consistent with the process used to set standards on the 2005 Grade 12 Math assessment. This process included comprehensive panelist training, facilitation throughout the process, and several quality control checks to confirm that panelists understood their tasks and the decisions they were making.
6. Standard-setting activities on NAEP assessments should continue to be thoroughly documented so that others can understand and evaluate the process. The documentation should include the logic regarding the choice of standard setting method, and procedural, internal, and external validity evidence.

This page left intentionally blank

References

- ACT. (1995). *Research studies on the achievement levels set for the 1994 NAEP in geography and U.S. history*. Iowa City, Iowa: Author.
- ACT. (April, 2005a). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Executive summary*. Iowa City, Iowa: Author.
- ACT. (April, 2005b). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Process report*. Iowa City, Iowa: Author.
- ACT. (May, 2005c). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Special studies report*. Iowa City, Iowa: Author.
- ACT. (May, 2005d). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Technical report*. Iowa City, Iowa: Author.
- Allen, J., and Sconing, J. (August, 2005). Using ACT Assessment scores to set benchmarks for college readiness. *ACT research report series 2005-3*. Iowa City, ACT.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Brennan, R. L. (1995). Standard setting from the perspective of generalizability theory. *Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board and the National Center for Educational Statistics* (Vol. 2, pp. 269–287). Washington, D.C.: National Assessment Governing Board and National Center for Educational Statistics.
- Brennan, R. L. (October, 2002). Estimated standard error of a mean when there are only two observations. *Center for Advanced Studies in Measurement and Assessment technical note number 1*. Iowa City: College of Education, University of Iowa.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30, 93–106.
- Cizek, G. J. (1996a). Setting passing scores. *Educational Measurement: Issues and Practice*, 15(1), 12–21.
- Cizek, G. J. (1996b). Standard setting guidelines. *Educational Measurement: Issues and Practice*, 15(2), 20–31.
- Cizek, G. J. (2001a). Conjectures on the rise and fall of standard setting: An introduction to context and practice. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Hillsdale, NJ.: Erlbaum.
- Cizek, G. J. (2001b (Ed.)). *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Hillsdale, N.J.: Erlbaum.
- Cizek, G. J., Bunch, B. B., and Koons, H. (2004). Setting performance standards: Contemporary methods [An NCME instructional module]. *Educational Measurement: Issues and Practice*, 23(4), 31–50.
- College Board (2006a). *Advanced placement: Report to the nation 2006*. New York, NY: Author.
- College Board (2006b). *SAT program handbook: A comprehensive guide to the SAT program for school counselors and admissions officers*. New York, NY: Author.

- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427–439.
- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Hillsdale, N.J.: Erlbaum.
- Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsythe, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., and Zwick, R. (2000). A response to “setting reasonable and useful performance standards” in the National Academy of Sciences “Grading the Nation’s Report Card.” *Educational Measurement: Issues and Practice*, 19(2), 5–14.
- Hambleton, R. K., and Powell, S. (1990). A framework for viewing the process of standard setting. *Evaluation and the Health Professions*, 6, 3–24.
- Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice*, 25 (2), 19–20.
- Jaeger, R. M. (1990). Establishing standards for teacher certification tests. *Educational Measurement: Issues and Practice*, 9(4), 15–20.
- Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice*, 10(2), 3–6, 10, 14.
- Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working Paper 2003-11). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.
- Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, 10, 93–98.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425–461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–88). Mahwah, N.J.: Erlbaum.
- Karantonis, A., and Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice*, 25 (1), 4–12.
- Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., and Angeles, J. (May, 1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, D.C.: American Institutes for Research.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., and Patz, R. J. (April, 1998). *The Bookmark procedure: Methodology and recent implementations*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, Calif.
- Lewis, D. M., Mitzel, H. C., and Green, D. R. (June, 1996). Standard setting: A Bookmark approach. In D. R. Green (Chair), *IRT-based standard setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, Ariz.
- Lewis, D. M., Mitzel, H. C., Green, D. R., and Patz, R. J. (1999). *The Bookmark standard setting procedure*. Monterey, Calif.: McGraw-Hill.
- Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, 11, 23–47.
- Linn, R. L., Koretz, D. M., Baker, E. L., and Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in*

- mathematics*. (Technical Report). Los Angeles: University of California, Center for the Study of Evaluation.
- Livingston, S. A., and Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, N.J.: Educational Testing Service.
- Loomis, S. C., and Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, N.J.: Erlbaum.
- McGinty, D. (2005). Illuminating the “black box” of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, 18, 269–288.
- Meara, K. P., Hambleton, R. K., and Sireci, S. G. (2001). Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review*, 12 (2), 17–23.
- Mehrens, W. A., and Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, 5, 265–283.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 249–281) Mahwah, N.J.: Erlbaum.
- National Assessment Governing Board (2004). *Mathematics Framework for the 2005 National Assessment of Educational Progress*. [online] Available: http://www.nagb.org/pubs/m_framework_05/761607-Math%20Framework.pdf.
- National Center for Education Statistics (2007). *America’s high school graduates: Results from the 2005 NAEP high school transcript study* (publication number NCES 2007467) Washington, D.C.: Author.
- Pellegrino, J. W., Jones, L. R., and Mitchell, K. J. (1999). *Grading the nation’s report card*. Washington, D.C.: National Academy Press.
- Raymond, M. R., and Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 119–157). Mahwah, N.J.: Erlbaum.
- Reckase, M. D. (2006a). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice*, 25 (2), 4–18.
- Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice*, 25 (3), 14–17.
- Reid, J. B. (1991). Training judges to generate standard setting data. *Educational Measurement: Issues and Practice*, 10(2), 11–14.
- Schulz, E. M. (2006). Commentary: A response to Reckase’s conceptual framework and examples for evaluating standard setting methods. *Educational Measurement: Issues and Practice*, 25 (3), 4–13.
- Schulz, E. M., Lee, W., and Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement*, 42, 1–26.
- Shepard, L. A., Glaser, R., Linn, R. L., and Bohrnstedt, G. (1993). *Setting performance standards for student achievement (final report)*. Stanford: National Academy of Education.
- Stufflebeam, D., Jaeger, R. M., and Scriven, M. (1991, August). *Summative evaluation of the National Assessment Governing Board’s inaugural 1990–91 effort to set achievement*

- levels on the National Assessment of Educational Progress.* Washington, D.C.: National Assessment Governing Board.
- Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement*, 4, 271–297.
- U. S. General Accounting Office (1993). *Educational achievement standards: NAGB's approach yields misleading interpretations* (Rep. No. GAO-PEMD-93-12). Washington, D.C.: Author.
- Vinovskis, M. A. (1998). *Overseeing the nation's report card: The creation and evolution of the National Assessment Governing Board.* Washington, D.C.: National Assessment Governing Board.
- Williams, N. J., and Schulz, E. M. (April, 2005). *An investigation of response probability (RP) values used in standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Yin, P., and Schulz, E. M. (April, 2005). *A comparison of cut scores and cut score variability from Angoff-based and Bookmark-based procedures in standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Zenisky, A. L., Hambleton, R. K., and Sireci, S. G. (March, 2007). Comprehensive evaluation of NAEP: Utility study final report. *Center for Educational Assessment Research Report No. 624.* Amherst, Mass.: University of Massachusetts.

Chapter 3:
**How Do Other Countries Measure Up to the Mathematics Achievement Levels
on the National Assessment of Educational Progress?**

Ronald K. Hambleton, Stephen G. Sireci, and Zachary R. Smith
University of Massachusetts Amherst

This page intentionally left blank

Contents

List of Figures and Tables.....	3-v
Abstract.....	3-vii
Introduction.....	3-1
Purpose.....	3-2
Prior Research.....	3-2
Comparing NAEP, PISA, and TIMSS Content.....	3-2
Linking NAEP and TIMSS Results.....	3-4
Method.....	3-7
National and International Assessment Data.....	3-7
Linking of the NAEP Achievement Levels to the TIMSS and Pisa Score Reporting Scales.....	3-7
Analyses.....	3-8
Results.....	3-11
Discussion.....	3-19
Limitations of This Study.....	3-20
Relating Our Findings to Phillips (2007).....	3-21
References.....	3-23

This page intentionally left blank

Figures and Tables

Figures

Figure 1: Illustration of Placing NAEP Achievement Levels on an International Score Scale .. 3-9

Tables

Table 1: Achievement Levels on the International Reporting Scales	3-8
Table 2: Indiana NAEP-TIMSS Comparisons	3-10
Table 3: 2003 NAEP Mathematics vs. TIMSS Mathematics for the Advanced Level	3-12
Table 4: 2003 NAEP Mathematics vs. TIMSS Mathematics for At or Above Proficient..	3-13
Table 5: 2003 NAEP mathematics vs. TIMSS Mathematics for At or Above Basic	3-14
Table 6: 2003 NAEP Mathematics vs. PISA Mathematics for the Advanced Level	3-15
Table 7: 2003 NAEP Mathematics vs. PISA Mathematics for At or Above Proficient.....	3-16
Table 8: 2003 NAEP Mathematics vs. PISA Mathematics for At or Above Basic	3-17

This page intentionally left blank

Abstract

In this study, we mapped achievement levels from the National Assessment of Educational Progress (NAEP) onto the score scales for selected assessments from the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Achievement (PISA). The mapping was conducted on NAEP, TIMSS, and PISA Mathematics assessments in 2003. A focus of the study was on whether the NAEP achievement levels were set too high. The results indicated that students from many other countries had substantially larger percentages of students meeting NAEP mathematics achievement levels. In general, the findings suggest the NAEP standard for Advanced is high, but not too high when considered within an international context. With respect to the NAEP standard of Proficient, none of the top-performing countries approached 100 percent proficient, which seems to underscore the different conceptualizations of “Proficient” in NAEP and *No Child Left Behind*.

This page left intentionally blank

Introduction

Educational reform in the United States is very much about being “world class” and policymakers want world class curricula, teachers, achievement levels, and student performance. The National Assessment of Educational Progress (NAEP) may be the best mechanism we currently have in the United States for judging the quality of student achievement—the assessments are administered on a regular basis, involve nationally representative samples of students, and are of very high technical quality. Currently, NAEP is used to assess students at three grade levels (4, 8, and 12), in many more subjects than in any state assessments, and scores are interpreted against a set of performance standards or achievement levels that can be applied to all students across the country. NAEP in all respects is a national assessment.

Since 1992 one of the primary means by which NAEP results are reported is through “achievement levels.” There are three achievement levels reported on all NAEP assessments: Basic, Proficient, and Advanced. The National Assessment Governing Board (NAGB) provides general descriptions for these achievement levels as well as specific descriptions in each subject area. The general descriptions are:

Basic: This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Proficient: This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.

Advanced: This level signifies superior performance. (NAGB, 2007)

Many policymakers, the media, and other consumers of NAEP data find the achievement level results useful for determining how well U.S. students measure up to expected standards of performance in specific subject areas, and how their performance with respect to these standards changes over time (Zenisky, Hambleton, and Sireci, 2007, this volume). However, although the achievement level results are widely used, they have also been sharply criticized. Essentially, these critics argue that the science of educational standard setting is too arbitrary to promote meaningful and reliable interpretations of students’ performance (e.g., Pellegrino, Jones, and Mitchell, 1999), and in the case of NAEP, critics have also argued that the achievement levels were set too high (e.g., Rothstein, 2006).

There have been many counterarguments raised against these critics citing the volumes of research supporting the validity of NAEP achievement level standards (see Loomis and Bourque, 2001; Hambleton et al., 2000). Much of the research in this area is based on procedural validity evidence (e.g., Was the method used to set the standards rational, defensible, and appropriately implemented? Did the standard-setting participants understand the tasks and have confidence in their ratings?) and internal validity evidence (Was there general agreement in the standards across panelists in the process? Are the standards consistent across different groups of panelists?). There have even been studies that looked at the consistency of NAEP standards across different methods used to set them (e.g., ACT 2005a, 2005b). However, until recently, *external* data, that is, data from educational assessments other than NAEP, have not been used to address the validity of the NAEP achievement levels (see, for example, Phillips, 2007).

In this paper, we evaluated data from international assessments to estimate how well students from other countries would measure up to the mathematics achievement levels set on NAEP. Data from international assessments not only represent important, independent measures of U.S. students’ academic achievement; they also bring an international perspective to the

discussion of NAEP achievement levels. It is only through international studies such as the *Trends in International Math and Science Study* (TIMSS) and the *Program for International Student Achievement* (PISA) that we can really determine if NAEP's current achievement levels are "world class." The basic assumption in this study is that if other countries can meet our achievement levels in larger percentages than the U.S., then it would be less reasonable to argue that the NAEP achievement levels have been set too high. It is hard to argue that standards are too high if many students can meet them. Of course, ultimately, it will be a judgment about whether sufficiently large numbers of students exceeded the standards to determine that they were not set too high.

Purpose

The purpose of this study was to compare the performance of eighth-grade students in the U.S. and other countries on the TIMSS and PISA mathematics assessments using the NAEP achievement levels. These comparisons will provide data that can assist policymakers, educators, and the public in deciding whether the NAEP achievement levels were set too high as claimed by some policymakers and researchers in the U.S. To carry out the study, it was necessary to map the NAEP achievement levels to the TIMSS and PISA reporting scales. Before explaining our methodology, we briefly review research related to NAEP, TIMSS, and PISA that is relevant to our study. The research reviewed includes analyses of the content measured by these assessments and a recent study that performed a similar evaluation of how international students would fare on NAEP assessments.

Prior Research

There have been several recent studies that investigated the similarity of the content of NAEP assessments to the content of international assessments. These studies include Nohara and Goldstein (2001); Scott (2004); and Neidorf, Binkley, Gattis, and Nohara (2006) (see also Ginsburg, 2005). An additional study conducted by Dossey, McCrone, and O'Sullivan (2006) compared the problem solving questions on the 2003 Grade 8 TIMSS and PISA Mathematics and Science assessments.

Comparing NAEP, PISA, and TIMSS Content

The Nohara and Goldstein (2001) study is somewhat dated since it dealt with assessments in 2000, but it was a comprehensive study that compared the eighth-grade science and mathematics portions of NAEP 2000 with TIMSS-R²⁴ and the scientific literacy and mathematics literacy portions of PISA. Subject matter experts in science and mathematics education analyzed items from each of the three assessments in terms of content, response type, context, requirements for multistep reasoning, and other characteristics. The authors found greater similarity between NAEP and TIMSS, than between PISA and the other two assessments. They explained these similarities and differences with respect to differences in the purpose of each assessment. Both NAEP and TIMSS-R assessed students' mastery of basic knowledge, concepts, and subject-specific thinking skills tied to broad curriculum frameworks. As a result, both assessments had large numbers of items covering a broad range of topics, with items generally focused on a single, identifiable piece of knowledge, concept, or skill. In contrast, the purpose of

²⁴ The TIMSS acronym originally stood for the Third International Mathematics and Science Study. TIMSS-R was introduced in 2000 as the Third International Mathematics and Science Study-Repeat. Since that time, this assessment has been slightly renamed—*Trends in International Mathematics and Science Study*, with the appropriate year following the acronym (e.g., TIMSS 2003).

PISA is to assess students' abilities to handle everyday situations that require scientific and mathematical skills. PISA items were set in more real-life contexts and fit the general frameworks of curriculum topics less well.

Neidorf et al. (2006)²⁵ compared the content of the 2003 NAEP, TIMSS, and PISA Mathematics assessments. For NAEP and TIMSS, they compared the Grade 4 and Grade 8 assessments. For PISA, they included the Mathematics assessment for 15 year-olds (PISA assesses students by age rather than grade level). Subject matter experts classified the items from each assessment to the content specifications of the other assessments. They found strong agreement between the NAEP and TIMSS assessments. Almost all of the items from NAEP could be classified into the TIMSS framework and vice-versa. When differences were found across the assessments, they were primarily due to the cross-grade linking items included on the NAEP assessments. They also found that the mathematics content of most of the PISA items was at the eighth-grade level, but that PISA was easily differentiated from the other two assessments with respect to item formats (with PISA having a smaller proportion of multiple-choice items), cognitive skills being measured (with more higher-complexity items on PISA), and notable content differences (PISA placed greater emphasis on data analysis and less emphasis on algebra, in relation to NAEP and TIMSS).

Neidorf et al. (2006) concluded that the NAEP and TIMSS mathematics assessments were very similar at the broadest content level, and were also similar with respect to the cognitive complexities of the items. However, they were different with respect to the sub-content areas and objectives measured. They also concluded that PISA items matched the NAEP content topics very well, but as mentioned above, there were notable difference in cognitive complexity and the relative proportions of items from the different content areas. They summed up the findings by stating that the three assessments “address many similar topics and require students to use a range of cognitive skills and processes, [but] it cannot be assumed that they measure the same content in the same way” (p. iv). The results of this study suggest that overall achievement, such as the total scores on these assessments, may be roughly comparable for making general statements regarding students' performance, but more detailed statements about specific content areas or cognitive skills are probably not comparable.

In comparing the Mathematics assessments from the 2003 TIMSS (Grade 8) and PISA (15-year-olds), Dossey et al. (2006) found a higher proportion of problem solving items measured on the PISA assessment (48 percent) than on the TIMSS assessment (38 percent). However, they found no significant differences with respect to measurement of the mathematics content areas. They concluded, “Though not significant, the distribution of the problem-solving items among the content areas in the mathematics portions of the two assessments appear to mirror the overall differences in emphases found in comparison to the mathematics content in the National Assessment of Educational Progress (NAEP) (Neidorf et al. 2006)” (p. vi).

In summary, the comparisons of what these NAEP, TIMSS, and PISA mathematics assessments are measuring suggest strong content overlap in mathematics between NAEP and TIMSS, and modest overlap between NAEP and PISA.

²⁵ Scott (2004) summarized the results of other studies, including Neidorf et al. (2006), but many of the comparisons refer to different years and subject areas outside the relevance of the present study and so we did not summarize them here.

Linking NAEP and TIMSS Results

A recent study by Phillips (2007) explicitly examined how well students from other countries who participated in TIMSS would perform with respect to the NAEP achievement levels. To accomplish this comparison, Phillips used the results from an earlier study by Johnson, Cohen, Chen, Jiang, and Zhang (2005) that “linked” the 1999 TIMSS Mathematics and Science assessments to the 2000 NAEP assessments. The study’s goal was to estimate how individual states within the U.S. would fare on the TIMSS assessments. They linked the two sets of assessments by having common groups of U.S. students (from 12 states) take both NAEP and TIMSS assessments in the same year. About 1,800 students completed the NAEP and TIMSS assessments in each subject area. They used both projection procedures (a regression-based procedure) and statistical moderation (forming the link by using the means and standard deviations of the two assessments for the common group of students) to link the NAEP and TIMSS scales. Their validation analyses indicated the projection method did not work well, but the statistical moderation worked well for the 12 states involved in the analyses.

Based on the statistical moderation linking functions calculated by Johnson et al. (2005), Phillips (2007) addressed the question, “How would other countries perform if their TIMSS results could be expressed in terms of NAEP achievement levels?” (p. 2). Specifically, he analyzed data from the 1999 TIMSS Mathematics and Science Assessment to predict achievement level results for students from other countries on the 2000 NAEP Mathematics and Science assessments. Although these secondary analyses compared students who were tested in different years, Phillips asserted the results “should be considered rough, ballpark estimates and should be used only for broad policy understandings” (p. 3).

The results from the Phillips (2007) study indicated that for Grade 8 Mathematics, the top five TIMSS countries in 1999 had at least 87 to 96 percent of their students at or above the NAEP “Basic” achievement level, 61 to 73 percent at “Proficient” or above, and 23 to 34 percent at the “Advanced” level. The achievement level results for the U.S. were substantially lower—65 percent, 27 percent, and 6 percent were at or above Basic, Proficient, and Advanced levels, respectively. The mean score for students in the U.S. placed them at the Basic level, but the top six TIMSS countries were at the NAEP Proficient level. Twelve of the 38 countries (32 percent), had mean scores that fell in the below Basic performance category.

For Science, the achievement level profiles reflect lower outcomes across the entire group of participating countries than what was observed in the mathematics assessments reported above. For the top five TIMSS countries, the percentages of students at or above Basic ranged from 74 to 80 percent, at or above Proficient ranged from 39 to 51 percent, and at or above Advanced ranged from 6 to 15 percent. The results for the U.S. were 59 percent, 30 percent, and 6 percent for at or above the Basic, Proficient, and Advanced categories, respectively. The mean NAEP scores for other countries classified only two as Proficient (Chinese Taipei and Singapore). Almost half of the 38 countries (18 or 47 percent) had mean scores that fell below Basic.

Phillips (2007) used the linking function to the 2003 TIMSS assessment and found similar results. He also showed how here the 2003 TIMSS “international benchmarks” (p. 13) fell on the NAEP scale and compared them to the locations of the NAEP achievement levels. He concluded that the benchmarks most similar across the two assessments were close for Mathematics, but not for Science, with the NAEP achievement levels being substantially higher.

An overall conclusion reached by Phillips (2007) was:

If a nation’s average performance is at the proficient level, then it indicates that the typical student in that country is reaching a level of performance that meets

U.S. standards. Interpreted in this way, we find that the United States is a nation that is not meeting its own expectations. (p. 20)

However, some persons have interpreted the results from the Phillips study as indicating the Grade 8 NAEP achievement levels in Mathematics and Science were set too high (Bracey, 2007). Our own view of the Phillips findings is that with substantially larger percentages of students in other countries meeting our NAEP achievement levels, especially in Mathematics, it becomes very difficult to reasonably argue that the current achievement levels for grade 8 Mathematics and Science were set too high. This interpretation of the findings from Bracey, which is at odds with Phillips' and our perspective, highlights for us the important point that ultimately, policymakers, educators, and the public will make their own determinations about the reasonableness of the U.S. achievement levels on the mathematics and science achievement scales.

This page left intentionally blank

Method

National and International Assessment Data

Our research plan involved linking the Grade 8 NAEP Mathematics achievement levels to the corresponding TIMSS and PISA reporting scales, when we could find matches for test administrations in the same year. All analyses focused on Grade 8 results. To provide timely comparisons, we only considered test administrations beginning in 2000 or later were considered.

For NAEP, Grade 8 Mathematics was administered in 2000, 2003, and 2005; and Grade 8 Science was administered in 2000 and 2005. TIMSS was administered to Grade 8 students in 2003; PISA was administered to 15-year olds in 2000 with a focus on Reading, and in 2003 with a focus on Mathematics. We focused on Mathematics and on Grade 8, (15-year-olds)²⁶ because of the opportunities to match up grade levels and ages among NAEP, TIMSS and PISA samples, and because we felt that the Grade 8 results (rather than Grade 4 results) would be more interesting to policymakers. Phillips (2007) did not use data collected in the same year when making comparisons across assessments. We felt that comparisons of achievement across countries would be best made if the year of test administration was common. We wanted to interpret the findings from the study with any growth in specific countries that may have taken place between the two test administrations.

We chose to make two comparisons: (a) 2003 NAEP Mathematics and 2003 TIMSS Mathematics (this comparison had the advantage that the data were the most recent we could use in mathematics and still match up the samples with test administrations in the same year); and (b) 2003 NAEP Mathematics and 2003 PISA Mathematics (this comparison too had the advantage of being relatively up-to-date and the year of administration of the two assessments was the same, and the same year as the NAEP-TIMSS comparison).

Linking of the NAEP Achievement Levels to the TIMSS and PISA Score Reporting Scales

We carried out the linking of score scales by using equipercentile equating of the three achievement levels from the NAEP reporting scale (see, for example, Waltman, 1997). We made the assumption that the national sample that produced the NAEP score distribution was essentially equivalent to the national sample used to produce the U.S. score distribution on TIMSS and PISA. In both comparisons, students in the national samples were administered tests in the same year. We preferred to do the linking when student samples were administered the assessments in the same year, which is one difference from the Phillips (2007) study. Doing so avoided the confounding of findings from the study with any achievement growth that may have occurred by the national student samples over time.

Placing the NAEP achievement levels on the international scales was straightforward. For a given NAEP subject administration (e.g., 2003 NAEP, Grade 8 Mathematics), we determined the percentiles corresponding the achievement levels for Basic, Proficient, and Advanced in the national NAEP score distribution. The percentiles were then applied to the U.S. score distribution on the international assessment to determine the corresponding achievement levels on the TIMSS and PISA score reporting scales. The achievement levels themselves appear in Table 1. An illustration of how we mapped the NAEP achievement levels onto the international scales is presented in Figure 1.

²⁶ All of the NAEP and TIMSS mathematics data came from students in the eighth grade; PISA data in mathematics came from 15-year-olds.

Table 1. Achievement Levels on the International Reporting Scales

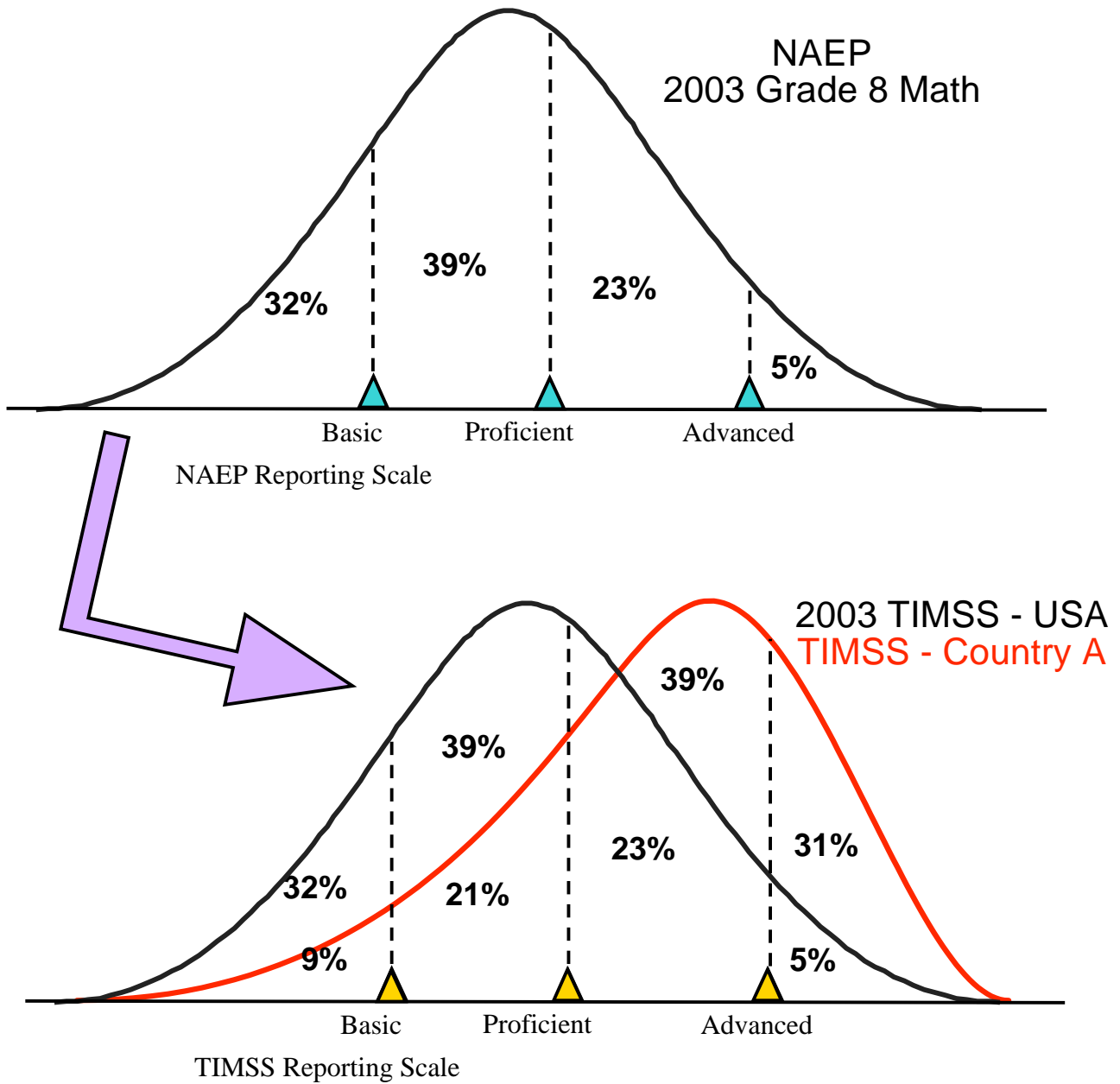
Assessment	Subject	Basic	Proficient	Advanced
PISA	Math (2003)	438	538	638
TIMSS	Math (2003)	467	550	632

Analyses

To conduct the linkings, we determined the NAEP achievement levels and their corresponding percentile ranks on the NAEP scale. Then, we found the corresponding achievement levels on the international scales using the U.S. samples and the percentile ranks, and applied these achievement levels to determine the percent of students in each participating country who were at or above the Basic, Proficient, and Advanced levels. With this information, we were able to compile the tables and figures that appear in the results section.

For TIMSS, we actually repeated the mapping five times and then averaged the results to obtain the three achievement levels: For each examinee we had five TIMSS scores (or plausible values) and these were used, one at a time for each examinee, to generate five U.S. score distributions on the TIMSS score reporting scale, NAEP percentiles corresponding to the achievement levels were applied to these distributions, and then the TIMSS achievement levels were obtained, one set for each distribution. Finally, the achievement levels were averaged across the five sets of estimates to obtain the best estimates of the achievement levels on the TIMSS reporting scales.

Figure 1. Illustration of Placing NAEP Achievement Levels on an International Score Scale



One of the interesting validity checks on our assumptions about content overlap and our approach to linking was provided by the state of Indiana. Indiana participated in the TIMSS-Mathematics Assessment in addition to NAEP in 2003 and was treated as a “country” in the analyses. Table 2 highlights the comparison of findings, and shows actual Indiana NAEP performance on the 2003 Mathematics Assessment compared to the state’s performance on the 2003 TIMSS Mathematics Assessment, using the NAEP achievement levels mapped to the TIMSS reporting scale. The results are close (the biggest difference was 3.8 percent and the average difference was 2.1 percent), and certainly close enough to support the types of comparisons we were interested in making in this study.

Table 2. Indiana NAEP-TIMSS Comparisons

Assessment	% At or Above Basic	% At or Above Proficient	% Advanced
NAEP-Math (2003)	74	31	5
TIMSS-Math (2003)	72.4	27.2	4

Results

The results of our analyses are summarized in Tables 3 through 8. The results are broken down by test (TIMSS or PISA) and achievement level. We begin with the 2003 NAEP-TIMSS comparison.

The results for the 2003 NAEP-TIMSS Mathematics comparison, which involved 47 countries, are summarized in Tables 3 to 5, for the Advanced, Proficient or Above, and Basic or Above achievement levels, respectively. The top five TIMSS countries had from 21 percent (Japan) to 41 percent (Singapore) of their students estimated to be Advanced. The U.S. ranked 11th with respect to this achievement level with about 5 percent of the students falling into this category. For Proficient or Above, over 60 percent of the students from the top five countries were classified into this level. For the U.S., the percentage was about 29 percent. The vast majority of students from the top five countries were classified as Basic or Above (ranging from about 87 percent for Chinese Taipei to 94 percent for Singapore, while about 68 percent of students from the U.S. were classified at Basic or Above. Thus, for all three achievement levels, the highest-performing TIMSS countries had noticeably larger percentages of students at or above each level.

The results for the 2003 NAEP-PISA Mathematics comparisons, which involved 30 countries, are summarized in Tables 6 to 8. For the Advanced achievement level, the top five countries had at least 13 percent of their students classified into this category, while the U.S. (ranked 26.5 of 30) had 5 percent of their students in this category. The top five countries had about half of their students meeting the Proficient or above standard, while only 29 percent of U.S. students met this standard. For at or above Basic, 28 out of the 30 countries had at least half of their students meet this standard, but once again the gap between the U.S. and the top performing countries was notable (68 percent for the U.S. versus 83 percent for Japan, which ranked fifth—the rankings are unstable because of the closeness of many countries, but the percentages of students are much more stable).

Table 3. 2003 NAEP Mathematics vs. TIMSS Mathematics for the Advanced Level

Participating Country	Rank	% Students Advanced
Singapore	1	40.6
Chinese Taipei	2	35.1
Korea, Rep. of	3	31.7
Hong Kong, SAR	4	26.5
Japan	5	21.1
Hungary	6	9.3
Netherlands	7	7.8
Belgium (Flemish)	8	7.3
Estonia	9	7.2
Slovak Republic	10	6.3
United States	11	5.4
Australia	12	5.2
Russian Federation	13	4.9
Israel	14.5	4.6
Malaysia	14.5	4.6
England	16.5	4.3
Lithuania	16.5	4.3
Latvia	18	3.9
New Zealand	19	3.8
Serbia	20	3.3
Romania	21	3.2
Bulgaria	22	2.8
Scotland	23	2.7
Sweden	24	2.6
Slovenia	25	2.5
Italy	26	2.1
Armenia	27	1.8
Cyprus	28	1.0
Moldova, Rep. of	29	0.9
Basque Region, Spain	30	0.8
Egypt	32	0.6
Jordan	32	0.6
Macedonia, Rep. of	32	0.6
Indonesia	34	0.5
Iran, Islamic Rep. of	36	0.3
Norway	36	0.3
South Africa	36	0.3
Chile	39	0.2
Palestinian Nat'l. Auth.	39	0.2
Philippines	39	0.2
Lebanon	42	0.1
Morocco	42	0.1
Saudi Arabia	42	0.1
Bahrain	45.5	0.0
Botswana	45.5	0.0
Ghana	45.5	0.0

Tunisia	45.5	0.0
---------	------	-----

Table 4. 2003 NAEP Mathematics vs. TIMSS Mathematics for At or Above Proficient

Participating Country	Rank	% Students At or Above Proficient
Singapore	1	76.8
Hong Kong, SAR	2	73.0
Korea, Rep. of	3	69.8
Chinese Taipei	4	66.1
Japan	5	61.7
Belgium (Flemish)	6	46.5
Netherlands	7	44.3
Hungary	8	40.5
Estonia	9	38.8
Slovak Republic	10	30.6
Malaysia	11	30.1
Russian Federation	12	29.8
Australia	13.5	29.1
Latvia	13.5	29.1
United States	15	28.8
Lithuania	16	27.7
Israel	17	26.7
England	18	25.6
Scotland	19	24.9
Sweden	20	24.2
New Zealand	21	24.0
Romania	22.5	21.3
Serbia	22.5	21.3
Slovenia	24	21.2
Armenia	25	20.7
Italy	26	19.4
Bulgaria	27	18.9
Basque Region, Spain	28	16.2
Cyprus	29	13.0
Moldova, Rep. of	30	12.7
Norway	31	10.1
Macedonia, Rep. of	32	9.4
Jordan	33	7.8
Egypt	34	6.3
Indonesia	35	6.0
Lebanon	36	4.2
Palestinian Nat'l. Auth.	37	3.9
Iran, Islamic Rep. of	38	3.4
Chile	39	3.3
Philippines	40	2.8
Bahrain	41	2.3
South Africa	42	2.1
Tunisia	43	1.3
Morocco	44	0.8
Botswana	45	0.6
Saudi Arabia	46	0.3
Ghana	47	0.1

Table 5. 2003 NAEP Mathematics vs. TIMSS Mathematics for At or Above Basic

Participating Country	Rank	% Students At or Above Basic
Singapore	1	93.8
Hong Kong, SAR	2	93.7
Korea, Rep. of	3	91.7
Japan	4	90.0
Chinese Taipei	5	86.5
Belgium (Flemish)	6	84.2
Netherlands	7	83.7
Estonia	8	82.6
Hungary	9	78.3
Latvia	10	71.5
Russian Federation	11	70.5
Malaysia	12	69.8
Slovak Republic	13	69.6
Australia	14.5	68.7
Sweden	14.5	68.7
United States	16	68.1
Scotland	17	67.4
Lithuania	18	66.8
England	19	65.4
Slovenia	20	64.4
Israel	21	64.2
Basque Region, Spain	22	63.2
New Zealand	23	63.1
Italy	24	59.9
Armenia	25	57.5
Romania	26	55.6
Bulgaria	27	55.4
Serbia	28	55.3
Moldova, Rep. of	29	49.0
Norway	30	48.9
Cyprus	31	48.6
Macedonia, Rep. of	32	37.6
Jordan	33	32.9
Lebanon	34	31.1
Indonesia	35	26.7
Egypt	36	26.5
Iran, Islamic Rep. of	37	23.0
Palestinian Nat'l. Auth.	38	21.4
Bahrain	39	20.2
Tunisia	40	17.8
Chile	41	17.3
Philippines	42	16.4
Morocco	43	12.6
Botswana	44	8.2
South Africa	45	6.0
Saudi Arabia	46	4.2
Ghana	47	2.1

Table 6. 2003 NAEP Mathematics vs. PISA Mathematics for the Advanced Level

Participating Country	Rank	% Students Advanced
Belgium	1	16.8
The Netherlands	2	15.2
Korea, Rep. of	3	15.1
Japan	4	15.0
Finland	5	13.5
Switzerland	6	13.0
New Zealand	7	12.5
Australia	8	11.5
Canada	9	11.4
Czech Republic	10	10.6
Germany	11	9.0
Denmark	12	8.7
Sweden	13	8.6
Israel	14	8.3
Great Britain	15	8.1
Austria	16	7.9
France	17	7.8
Slovak Republic	18	6.4
Norway	19	5.8
Hungary	20	5.7
Ireland	21.5	5.5
Luxembourg	21.5	5.5
Poland	23	5.3
United States	26.5	5.0
Spain	26.5	5.0
Greece	26.5	5.0
Italy	26.5	5.0
Portugal	26.5	5.0
Turkey	26.5	5.0
Mexico	30	0.0

Table 7. 2003 NAEP Mathematics vs. PISA Mathematics for At or Above Proficient

Participating Country	Rank	% Students At or Above Proficient
Finland	1	52.9
Korea, Rep. of	2	52.6
Japan	3.5	50.6
The Netherlands	3.5	50.6
Belgium	5	49.7
Canada	6	48.3
Switzerland	7	46.7
Australia	8	45.9
New Zealand	9	45.1
Czech Republic	10	41.7
Israel	11	41.5
Denmark	12	40.7
France	13	40.0
Germany	14	39.5
Sweden	15	38.3
Great Britain	16	38.1
Austria	17	37.5
Ireland	18	34.3
Slovak Republic	19	33.9
Norway	20	32.6
Luxembourg	21	32.2
Hungary	22	31.3
Poland	23	30.2
United States	24	29.0
Spain	25	28.2
Italy	26	22.4
Portugal	27	20.9
Greece	28	16.2
Turkey	29	13.3
Mexico	30	5.0

Table 8. 2003 NAEP Mathematics vs. PISA Mathematics for At or Above Basic

Participating Country	Rank	% Students At or Above Basic
Finland	1	90.0
Korea, Rep. of	2	86.8
Canada	3	85.8
The Netherlands	4	85.0
Japan	5	82.7
Switzerland	6	81.5
Australia	7	81.3
New Zealand	8	80.3
Israel	9	80.1
Denmark	10	79.9
Belgium	11	79.6
Czech Republic	12.5	78.4
France	12.5	78.4
Sweden	14	77.5
Ireland	15	77.4
Great Britain	16	76.9
Austria	17	75.5
Slovak Republic	18	74.2
Germany	19	73.5
Norway	20	73.0
Luxembourg	21	72.4
Poland	22	71.6
Hungary	23	70.9
Spain	24	70.7
United States	25	68.0
Portugal	26	62.8
Italy	27	61.6
Greece	28	53.5
Turkey	29	41.0
Mexico	30	27.1

This page left intentionally blank

Discussion

In this study, we compared results from NAEP and international mathematics assessments in 2003. A focus of our analysis was on whether the NAEP achievement levels were set too high, particularly at the Advanced level, as some critics have claimed. With respect to the Advanced achievement level, the results suggest the answer is “No.” The NAEP-TIMSS comparison indicated that the highest-performing countries had substantially larger percentages of students scoring Advanced, relative to the U.S. For example, Singapore, the highest-performing country had about 41 percent of their students classified as Advanced, compared with only about 5 percent of U.S. students. Clearly, this achievement level is not too high, if we are talking about world-class standards.

With respect to the NAEP standard of Proficient, again, many countries outperformed the U.S. by having much larger percentages of students at or above this level. For example, the corresponding percentages for at or above Proficient were 77 percent for Singapore and 29 percent for the U.S. Japan, ranked fifth, had about 62 percent of their students at or above Proficient. Although these countries had substantially higher proportions of students at or above Proficient, it should be noted that even these countries did not achieve 100 percent proficient, which is the goal of the *No Child Left Behind Act (NCLB)*. Stoneberg (2007) argued that proficiency with respect to *NCLB* refers to grade-level expectations, and proficiency with respect to NAEP refers to a higher level of achievement. This point may explain the political and philosophical differences between state and NAEP standards for proficient, even though the same term is used. To reinforce his point, Stoneberg quoted from Loomis and Bourque (2001), in their description of NAEP standard setting, who stated:

It is important to understand clearly that the Proficient achievement level does not refer to “at grade” performance. Nor is performance at the Proficient level synonymous with “proficiency” in the subject. That is, students who may be considered proficient in a subject, given the common usage of the term, might not satisfy the requirements for performance at the NAEP achievement level. (Loomis and Bourque, 2001, cited in Stoneberg, 2007).

Stoneberg also argues that for NAEP-state comparisons, the percentage of students at or above Basic should be used. Turning to our results, for at or above Basic, 94 percent of Singapore's students met this mark, as did 90 percent of Japan's students, compared with 68 percent of U.S. students. Thus, relative to the question of whether these NAEP standards are too high, it appears they are not, when taken within an *international* context. If these standards were set too high, the top-performing TIMSS countries would not have such notably larger percentages of students surpassing them. With respect to the use of the NAEP Proficient level as the standard of proficiency as defined in *NCLB* for adequate yearly progress, the international results presented here indicate none of the countries currently participating in TIMSS would come close to achieving 100 percent proficiency. The closest would be Singapore on the TIMSS Mathematics with 77 percent of their students being judged as Proficient or above.

Although many of the countries involved in the NAEP-PISA comparison were different from those involved in NAEP-TIMSS, the results were similar. For all three achievement levels, the percentages of students from the top-performing countries were noticeably higher than those for the U.S. Ten countries had 10 percent or more of their students in the Advanced category, compared with 5 percent for the U.S. For Proficient, 23 of the 30 countries had a larger percentage of students at or above this level than did the U.S. Thus, for 2003 Mathematics, there is consistency in the results across TIMSS and PISA and there does not appear to be evidence the NAEP achievement levels were set too high.

Limitations of This Study

Our linking of NAEP and TIMSS and NAEP and PISA results allowed us to provide rough estimates of how well students from other countries would perform with respect to some specific Grade 8 NAEP achievement levels. However, the scores from these assessments were not on the same scale and were not formally equated and so there are several limitations that should be considered. First, the corresponding samples of NAEP examinees (NAEP versus the corresponding sample on the international assessment administered in the same year) were not strictly "randomly equivalent." For example, different exclusion rules and accommodations are being used in NAEP, TIMSS, and PISA, and the testing for the three assessments is not done at the same time in the school year. Also, PISA students are a bit older than the corresponding students in the eighth-grade NAEP samples. Finally, while the test content is relatively similar and the balance of item formats about the same (see, Ginsburg, 2005), the linking might be best described as one that could build a concordance table, much like the linking that is done to produce comparable scores on the SAT and the ACT.

As for test content, Scott (2004) and others have concluded that the content is relatively similar in Mathematics between NAEP and TIMSS. The balance of item formats in the assessments is about the same, too.

The National Center for Education Statistics (NCES), the National Assessment Governing Board (NAGB), and others have done a lot of work suggesting sufficiently similar samples and test content are present to attempt to make some comparisons albeit with qualifiers about the potential for differences and their impact on the findings. And note, that strict compliance with the assumptions of random groups equating is not essential because (1) the goal of the study is not to precisely rank the countries, and (2) the goal is not to strictly equate but to establish a concordance table that only uses three matched points on each scale. In considering these limitations, we agree with Phillips (2007) who stated

....Such comparisons are not perfect, always require more research, and should be done with caution. However, such cross-country comparisons result in the cross-fertilization of information and help inform debate. In general, comparisons are useful in providing information to policymakers and the general public to help them achieve broad understandings that they otherwise would not have (p. 1).

In interpreting the results presented here, it should be noted that rankings of countries are rather unstable. According to Phillips (2007) and our own estimates of standard errors from the TIMSS data, the standard errors of the percentages were of the order of 1.5 or higher. But the trends themselves in the findings are clear without particular emphasis on the findings at the advanced level.

Relating Our Findings to Phillips (2007)

The Phillips (2007) study was intended to help policymakers make sense of international test results by placing the well-known NAEP achievement levels on the reporting scale for TIMSS. He showed how the U.S. performance can be compared to countries using the NAEP achievement levels on the international score-reporting scale for TIMSS. His focus was on facilitating the interpretations of results from TIMSS, PISA, and other international studies. He is right of course, in his approach, but he might have gone on to say, that studies like his show sizable numbers of countries are performing better in grade 8 mathematics than the U.S., and therefore, it would be difficult to argue convincingly that our U.S. achievement levels are too high. If they were, then considerably fewer countries would be exceeding our achievement levels. The evidence is strong that the achievement levels are achievable by higher numbers of students in other countries than observed in the U.S.

The Phillips findings are consistent with our own, and serve one other purpose from our perspective. Phillips uses a complex linking approach (statistical moderation) because of a need to match scale score to scale score across two assessments. He arrived at very similar conclusions to the ones we observed, and our conclusions are based on more recent international data in the area of mathematics.

From both the present study and that of Phillips (2007), it can be seen that data from international assessment initiatives are very useful for understanding the relative achievement of U.S. students and for improving our interpretations of NAEP scores. The results of the present study suggest that NAEP Mathematics standards may be high, but from our perspective, they do not appear to be unreasonably high. In fact, many students from countries that compete economically with the U.S. outperformed U.S. students in all three achievement level categories. By comparing results from future NAEP, TIMSS, and PISA assessments, U.S. policymakers, educators, and the public will be able to see if more students, in both the U.S. and abroad, are able to meet the standards NAGB has set for them, and make their own judgments about whether or not the standards set were too high.

This page left intentionally blank

References

- ACT, Inc. (April, 2005a). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Process report*. Iowa City, Iowa: Author.
- ACT, Inc. (May, 2005b). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Special studies report*. Iowa City, Iowa: Author.
- Bracey, G. (May 3, 2007). A test everyone will fail. *Washington Post*, downloaded June 16, 2007, from <http://www.washingtonpost.com>.
- Dossey, J.A., McCrone, S.A., and O’Sullivan, C. (2006). *Problem Solving in the PISA and TIMSS 2003 Assessments* (NCES 2007-049). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics. Retrieved Aug. 1, 2007 from <http://nces.ed.gov/pubsearch>.
- Ginsburg, A. (February, 2005). Benchmarking NAEP’s methodology against TIMSS and PISA: Issues to consider in evaluating NAEP. Presentation at the TWG for the NAEP evaluation project, Washington, D.C.
- Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., and Zwick, R. (2000). A response to “setting reasonable and useful performance standards” in the National Academy of Sciences “Grading the Nation’s Report Card.” *Educational Measurement: Issues and Practice*, 19(2), 5–14.
- Johnson, E., Cohen, J., Chen, W.H., Jiang, T., and Zhang, Y. (2005). *2000 NAEP—1999 TIMSS Linking Report* (NCES 2005–01). Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.
- Loomis, S. C., and Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, N.J.: Erlbaum.
- National Assessment Governing Board. (2007). *Achievement levels*. Downloaded from <http://www.nagb.org/> on June 16, 2007.
- Neidorf, T. S., Binkley, M., Gattis, K., and Nohara, D. (May, 2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments: Technical report*. Washington, D.C.: U.S. Department of Education, Institute of Educational Sciences.
- Nohara, D., and Goldstein, A. A. (June, 2001). *A comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)*
- Pellegrino, J. W., Jones, L. R., and Mitchell, K. J. (1999). *Grading the nation’s report card*. Washington, D.C.: National Academy Press.
- Phillips, G. W. (2007). *Expressing international education achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS*. Washington, D.C.: American Institutes for Research.
- Rothstein, R. (2006). “Proficiency for All” is an oxymoron. *Education Week*, Nov. 29, 2006.
- Scott, E. (2004). *Comparing NAEP, TIMSS, and PISA in Mathematics and Science*. Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.
- Stoneberg, Bert D. (2007). Using NAEP to Confirm State Test Results in the No Child Left Behind Act. *Practical Assessment Research and Evaluation*, 12(5). Downloaded from <http://pareonline.net/getvn.asp?v=12&n=5> on Aug. 16, 2007.

- Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement*, 34(2), 101–121.
- Zenisky, A. L., Hambleton, R. K., and Sireci, S. G. (2007). *Comprehensive evaluation of NAEP: Utility study final report* (Center for Educational Assessment Research Report No. 624). Amherst, Mass.: University of Massachusetts, Center for Educational Assessment.

Chapter 4:

A Study of the Utility of the National Assessment of Educational Progress

April L. Zenisky, Ronald K. Hambleton, and Stephen G. Sireci

Center for Educational Assessment

University of Massachusetts Amherst

This page intentionally left blank

Contents

List of Figures and Tables.....	4-v
Executive Summary.....	4-ix
Introduction.....	4-1
Purpose of the Utility Study.....	4-3
Review of Previous Research on NAEP Reports.....	4-5
NAEP Reporting: Summary of Current Policies and Practices.....	4-7
NAEP on the Web.....	4-9
Overview of Research.....	4-9
A Brief Review of NAEP on the Web.....	4-9
NAEP Web Site Usage.....	4-11
Site User Observations.....	4-12
Focus on the NAEP Data Explorer (NDE).....	4-17
Summary of Research on NAEP on the Web.....	4-24
NAEP Data Displays.....	4-27
Overview of Research.....	4-27
Current NAEP Data Display Methods.....	4-27
NAEP Data Displays: Impressions from the NAEP State Coordinators.....	4-28
Identifying Sources of Confusion in Current NAEP Displays.....	4-29
Revising Data Displays for NAEP.....	4-31
Summary of Research on NAEP Data Displays.....	4-44
Reporting Interests of NAEP Audiences.....	4-47
Overview of Research.....	4-47
Coordinators' Reporting Experiences in the States.....	4-47
Coordinators' Experiences with NAEP Sub-Audiences.....	4-48
Focus Group Findings.....	4-49
Reporting Interests from Site Usage Statistics.....	4-50
Summary of Research on NAEP Reporting Interests.....	4-50
Summary and Conclusions.....	4-53
References.....	4-57
Appendixes.....	4-61
Appendix A: Navigating 'The Nation's Report Card' on the World Wide Web.....	4-63
Appendix B: Do-It-Yourself NAEP Data Analysis on the Web.....	4-77
Appendix C: NAEP Web Site Usage: March 2005 to February 2006.....	4-101
Appendix D: NAEP Web-Based Score Reporting Evaluation.....	4-131
Appendix E: Do Mathematics Educators Use and Understand NAEP Score Reports?.....	4-141
Appendix F: State Reading Content Specialists and NAEP Data Displays.....	4-165
Appendix G: Displaying NAEP Results Effectively.....	4-201

This page left intentionally blank

Figures and Tables

Figures

Figure 1: The NAEP Home Page (captured June 26, 2006).....	4-13
Figure 2: The www.nationsreportcard.gov home page (Dec. 27, 2006)	4-14
Figure 3: Accessing the Cross-State Comparison Maps	4-16
Figure 4: NAEP Data Explorer Start Page	4-18
Figure 5: Analysis Selection Page, Quick Start Mode	4-20
Figure 6: NAEP Clickable State Comparison Map (Scale Scores)	4-30
Figure 7: Achievement Level Percentages for Five States.....	4-32
Figure 8: Bar Graph of Achievement Levels.....	4-33
Figure 9: Line Graph Illustrating Performance Gaps Between Two Student Groups.....	4-34
Figure 10: NAEP Pantyhose Chart.....	4-35
Figure 11: Clustered Bar Chart.....	4-36
Figure 12: Score difference graph	4-37
Figure 13: NAEP Item Map.....	4-39
Figure 14: NAEP Line Graph.....	4-40
Figure 15: NAEP Stacked Column Chart.....	4-42
Figure 16: NAEP Achievement Levels Table.....	4-43
Figure A-1: The NAEP Home Page (June 26–27, 2006).....	4-67
Figure A-2: The www.nationsreportcard.gov Home Page.....	4-69
Figure A-3: Illustration of the Link to the NAEP Home Page.....	4-70
Figure A-4: Links to the Cross-State Comparison Maps.....	4-72
Figure A-5: Accessing the Cross-State Comparison Maps.....	4-73
Figure A-6: Cross-State Comparison Maps.....	4-74
Figure B-1: NAEP Data Explorer Start Page.....	4-79
Figure B-2: Analysis Selection Page, Quick Start Mode.....	4-81
Figure B-3: Analysis Selection Page, Quick Start Mode with Options “Grayed Out”.....	4-82
Figure B-4: NDE Results Page, Average Scale Score.....	4-86
Figure B-5: NDE Results Page, Achievement Levels (Discrete).....	4-87
Figure B-6: Significance Testing Window from the NDE.....	4-89
Figure B-7: NDE Significance Test Results.....	4-90
Figure B-8: Graph Creation Window.....	4-91
Figure B-9: Sample NDE Graph (full graph mode).....	4-92
Figure B-10: Analysis Selection Page, Advanced Mode.....	4-93
Figure B-11: Format Table, Advanced Mode.....	4-94
Figure B-12: Regression Option Screen, Advanced Mode.....	4-95
Figure B-13: Results of Regression Analysis.....	4-96
Figure C-1: Frequency of Page Views and Visits to the NAEP Home Page March 2005– February 2006.....	4-106
Figure C-2: Screen Capture of the State Profile Access Page, with Alabama and Go Button Highlighted.....	4-110
Figure C-3: Use of the NAEP Data Tool, the Question Tool, and the State Profiles.....	4-111
Figure D-1: Selected Web Resources on http://nces.ed.gov/nationsreportcard/ by Category.....	4-133
Figure E-1: Average Mathematics Scale Scores, Grade 8: Various Years 1990-2005.....	4-146
Figure E-2: Percentage of Students at or above Basic and at or above Proficient in Mathematics Grade 4 Various Years, 1990-2005.....	4-147

Continues next page

Figure E-3: Cross-State Comparisons of Average Mathematics Scale Scores, Grade 4 Public Schools: 2005.....	4-148
Figure E-4: Cross-State Comparisons of Average Mathematics Scale Scores, Grade 4 Public Schools: 2005.....	4-149
Figure E-5: Example Multiple Choice Item from Grade 4 Mathematics.....	4-151
Figure E-6: Example Multiple Choice Item Results from Grade 4 Mathematics.....	4-152
Figure E-7: Example Scale Score Results from Grade 4 Mathematics.....	4-153
Figure E-8: Example Multiple Choice Item Results from Grade 8 Mathematics.....	4-155
Figure E-9: Example Scale Score Results from Grade 8 Mathematics.....	4-156
Figure E-10: Overall Cross-District Comparisons of Average Mathematics Scale Scores, Grade 4 Public Schools: 2005.....	4-157
Figure E-11: Percentages of Students at or above each Achievement Level for Mathematics, Grade 8.....	4-158
Figure E-12: Grade 4 Item Map.....	4-160
Figure F-1: Average Reading Scale Scores, Grade 4: Various Years, 1992–2005.....	4-169
Figure F-2: Percentages of Students at or above Basic and at or above Proficient in Reading, Grade 8: Various Years 1992–2005.....	4-171
Figure F-3a: Cross-State Comparisons of Average Reading Scale Scores, Grade 4 Public Schools: 2005.....	4-173
Figure F-3b: Cross-State Comparisons of Percentage of Students at or above Proficient in Reading, Grade 4 Public Schools: 2005.....	4-174
Figure F-4: Cross-State Comparisons of Average Reading Scale Scores, Grade 8 Public Schools: 2005.....	4-175
Figure F-5a: Example Grade 4 Reading Passage.....	4-176
Figure F-5b: Example Multiple Choice Item Grade 4 Reading.....	4-177
Figure F-6: Example Multiple Choice Item Results from Grade 4 Reading.....	4-178
Figure F-7: Example Scale Score Results from Grade 4 Reading.....	4-179
Figure F-8a and F-8b: Example Constructed Response Item and Results from Grade 4 Reading.....	4-181
Figure F-8c: Scoring Guide.....	4-182
Figure F-8d: Extensive—Student Response.....	4-183
Figure F-8e: Essential—Student Response.....	4-184
Figure F-8f and 8g: Partial and Unsatisfactory—Student Response.....	4-185
Figure F-9: Overall Cross-District Comparisons of Average Reading Scale Scores, Grade 4 Public Schools: 2005.....	4-186
Figure F-10: Average Reading Scale Scores and Percentages of Students within each Achievement Level, Grade 4 Public Schools: By State 2005.....	4-188
Figure F-11: White—Hispanic Scale Score Comparison on Grade 4 Reading.....	4-190
Figure F-12: White—Hispanic Gap in Average Reading Scores, Grade 4 Public Schools.....	4-192
Figure F-13: Grade 8 Item Map.....	4-194
Figure F-14: Average Reading Scale Scores, Grade 4 Public Schools: By Urban District, Various Years, 2002–2005.....	4-195
Figure F-15: Percentages of Students at each Achievement Level for Reading, Grade 4.....	4-197
Figure G-1: Average Reading Scale Scores, Grade 4: Various Years, 1992–2005.....	4-204
Figure G-2: Percentage of Students at or above Basic and at or above Proficient in Mathematics, Grade 4 Various Years 1990–2005.....	4-205
Figure G-3: Percentage of Students, by Mathematics Achievement Level, Grade 4: Various Years, 1990–2005.....	4-206

Continues next page

Figure G-4: Percentage of Students in Achievement Levels in Mathematics, Grade 4 Various Years, 1990–2005.....	4-207
Figure G-5: Cross-State Comparisons of Average Reading Scale Scores, Grade 4 Public Schools: 2005.....	4-209
Figure G-6: Cross-State Comparisons of Percentage of Students at or above Proficient in Reading, Grade 4 Public Schools: 2005.....	4-210
Figure G-7: Cross-State Comparisons of Average Reading Scale Scores, Grade 8 Public Schools: 2005.....	4-211
Figure G-8: Average Reading Scale Scores and Percentage of Students within each Achievement Level, Grade 4 Public Schools: By State, 2005.....	4-213
Figure G-9: Overall Cross-District Comparisons of Average Mathematics Scale Scores, Grade 4 Public Schools: 2005.....	4-214
Figure G-10: Percentages of Students at each Achievement Level for Reading Grade 4.....	4-215
Figure G-11: Example Scale Score Results Grade 4 Reading.....	4-217
Figure G-12: Grade 8 Item Map.....	4-219
Figure G-13: Trends in Average Mathematics Scale Scores and Score Gaps for Students Ages 9, 13, and 17, by Gender: 1973–2004.....	4-221
Figure G-14: Gaps in Average Mathematics Scale Scores, by Gender, Grades 4 and 8: 1990–2003.....	4-222
Figure G-15: Average Mathematics Scale Scores and Score Gaps for White—Hispanic Students, Grade 8: Various Years, 1990–2005.....	4-223
Figure G-16: Achievement-Level Results in Mathematics, by Gender, Grade 8: Various Years, 1990–2005.....	4-224
Figure G-17: White—Hispanic Gap in Average Reading Scores, Grade 4 Public Schools: By Urban District, Various Years, 2002–2005.....	4-225

Tables

Table 1: Utility Study Activities by Research Question.....	4-4
Table 2: Recommendations by Research Question Area and Topic	4-53
Table C-1: Frequency of Visits and Views to the NCES Web Site by Platforms and Browser Use	4-104
Table C-2: Most Popular Pages on the NAEP Site (March 2005 to Feb. 2006)	4-108
Table C-3: Most Common Entry Pages for the NAEP Site	4-112
Table C-4: Most Commonly Loaded Pages on http://www.nationsreportcard.gov , Oct. 2005–Feb. 2006	4-113
Appendix C-1: Most Popular Pages on the NAEP Site, by Month (March 2005 to February 2006)	4-117
Appendix C-2: Most Common Entry Pages for the NAEP Site by Month	4-124
Appendix C-3: Most Commonly Accessed Pages on the www.nationsreportcard.gov Site by Month.....	4-127
Table D-1: Summary of Web Evaluation Methods	4-137

This page left intentionally blank

Executive Summary

Tens of millions of dollars are spent on the National Assessment of Educational Progress (NAEP) each year so that policymakers, educators, and the general public are informed of the academic knowledge and skills of our nation's students and of changes in educational achievement over time. For NAEP to accomplish its goal of making "available reliable information about the academic performance of American students in various learning areas"²⁷ the assessments must be technically sound and the results must be interpreted appropriately, per the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). A substantial amount of research has been conducted on the technical aspects of NAEP. In contrast, however, very little research has been conducted on how well NAEP results are reported and interpreted. Clearly, if the intended audiences do not understand NAEP results, or if NAEP results are misinterpreted, the entire enterprise would be a failure from a validity perspective, regardless of the technical merits of NAEP.

In this report, we summarize two years' worth of research on the current *reporting* of NAEP results. The key questions guiding our evaluation of NAEP reporting were:

1. How do users of NAEP information regard the available NAEP information that is on the Web, including the online, interactive data tools?
2. How well are the current paper versions of NAEP reports and displays functioning with their intended audiences? Are stakeholders able to understand the information they are being presented with and use it to answer basic questions?
3. What are the reporting interests and preferences of NAEP audiences?

Details regarding our study methods, results, and interpretations of the results are contained in this report. General findings, as reported below, include recommendations for both policy and operational reporting efforts:

- Results reported for NAEP are comprehensive and are targeted to important audiences including policymakers, state and local education officials, educators, and the general public. Through the present research activities and those reported in literature reviewed, we found that these audiences understood NAEP results and were pleased with the depth and breadth of information provided. However, there was evidence of a relatively high level of confusion among many NAEP users surrounding both scale score and achievement level reporting for NAEP due to stakeholders' uncertainty about the relationship between NAEP results for the nation and the states, and states' reporting of their own *NCLB* assessment results. We recommend:
 - Operational: *Create additional score report designs.*
 - Operational: *Carry out focus-group work to eliminate confusion with various elements of current NAEP score reports.*
- NAEP results should continue to use the World Wide Web as a primary reporting mechanism. We recommend:
 - Policy: *Develop a comprehensive program of ongoing usability research for the NAEP Web site.*
 - Policy: *Carry out targeted studies of the Initial Release Site.*

²⁷ Downloaded on March 4, 2007 from <http://www.nagb.org/>.

- Policy: *Focus NAEP Web site research on the most commonly accessed pages within the NAEP presence on the Web, including the State Profiles, Question Tool, subgroup results, the Initial Release Site, and the NAEP Data Explorer (NDE).*
 - Operational: *Reorganize aspects of the NAEP Web site to reflect empirical findings about ease of use, audience interests, and current accepted Web development practices.*
 - Operational: *Consider minor revision to the NAEP home page with respect to the left navigation bar, the placement of the home page link, and the presence of multiple search boxes.*
 - Operational: *Continue empirical evaluation of the NDE with NAEP audiences.*
 - Operational: *Review the appearance, functionality, and layout of the NDE variable selection page.*
 - Operational: *Consider reformatting the data output window to reflect how other statistical software packages structure results.*
 - Operational: *Streamline the NDE's regression, statistical significance, and graphing functions.*
- Graphical presentations of results are an integral part of all NAEP reporting efforts, and can be an effective way to depict a wide range of assessment results quickly and clearly. It is recommended that NAEP results:
 - Policy: *Pay attention to deciding what relevant information is to be reported and what graphical display most clearly and effectively represents the information to be communicated.*
 - Policy: *Use focus groups for evaluating graphical presentations with NAEP audiences.*
 - Operational: *Use graphs when possible rather than tables.*
 - Operational: *Explicitly define, label and provide context for the NAEP score scale.*
 - Operational: *Report percentages of students within each achievement level rather than "at or above."*
 - Operational: *Provide users with clear notes defining statistical significance and appropriate interpretations.*
 - Operational: *Review legends, graph axes, and footnotes for accuracy and clarity.*
 - Data presented in NAEP results should be readily comprehensible to the various audiences it serves. To improve the comprehensibility of NAEP results, we recommend:
 - Policy: *Carry out studies of NAEP-state alignment to provide users with additional information for understanding NAEP's state-level scores and achievement level results relative to performance on state testing programs.*
 - Policy: *Pursue and publicize strategies such as item maps and skill profiles to add further context to results.*
 - Operational: *Clarify the reporting of achievement levels to communicate meaning associated with performance since confusion still appears to exist between scale scores and achievement level score reporting. If both score reporting approaches are to exist, then more attention should be given to distinguishing them in the minds of NAEP score users.*
 - Operational: *Report scale scores in context (including the range of possible scores and examples of skills/knowledge exhibited by individuals at that score level), because scale scores in isolation were not as informative as performance levels (in NAEP, below Basic, Basic, Proficient, and Advanced).*

- Operational: *Concentrate reporting efforts at the state level (including score gaps of importance) and work with NAEP state coordinators to ensure the appropriateness and usability of data tools for reporting.*
 - Operational: *Develop strategies to communicate the practical significance and appropriate interpretations of differences in scale score results for NAEP reporting groups (e.g., gender, race, and ethnicity).*
 - Operational: *Prepare a cognitive analysis of the different aspects of the experience of analyzing and using NAEP data and reports.*
- With respect to overall policy recommendations for NAEP score reporting, we recommend:
 - Policy: *Carry out systematic studies of planned and current ongoing reporting strategies (data displays, Web pages or tools) with stakeholder groups prior to operationalizing the use of these reporting strategies.*
 - Policy: *Develop formal procedures for incorporating research findings on reporting strategies into operational reporting efforts.*
 - Policy: *Revisit and revise aspects of the NAEP Web site to reflect empirical findings about ease of use, audience interests, and current accepted Web development practices.*
 - Policy: *Consider ways to incorporate stakeholder interest by developing materials for targeted audiences.*

NAEP score reporting has improved substantially in the last 15 years due, in part, to criticisms from stakeholders, implementation of new research findings about score reporting, and the program's goal of making NAEP results accessible to a wider set of audiences. We hope the findings and recommendations in this report will encourage NAGB and NCES to continue to develop NAEP score reporting and expand its accessibility to an ever-increasing number of users.

This page left intentionally blank

Introduction

The National Assessment of Educational Progress (NAEP) is uniquely designed to assess and report on the academic proficiency of American elementary, middle, and high school students in its role as “The Nation’s Report Card.” Through its various assessments and special studies, NAEP monitors changes in student academic achievement over time at both the state and national levels, and disseminates results to a wide range of intended audiences including policymakers, educators, researchers, and members of the general public. Few (if any) other testing programs have the scope and substance to influence national education policy as NAEP can.

NAEP provides policymakers, educators, researchers, and members of the public with information about the reading, mathematics, science, geography, civics, economics, United States history, and writing knowledge and skills of elementary, middle, and high school students. It monitors changes in student achievement over time at both the state and national levels. Considerable statistical and psychometric sophistication is used in test design, data collection, test data analysis, and scaling (see, for example, Beaton and Johnson, 1992; Johnson, 1992; Mislevy, Johnson, and Muraki, 1992).

Substantial time and millions of dollars have been spent over the past two decades (since ETS was awarded the NAEP contract in 1984) overcoming complex, technical problems associated with NAEP and its matrix sampling design. Many state assessment programs have also benefited from the technical efforts of NCES and ETS on NAEP.

Until recently, however, far less attention has been given to the ways in which the complex NAEP data are organized and reported, and accessed and used on the NCES Web site. Increasingly, the Internet is becoming the principal means by which interested parties can access assessment information, and there appears to be a distinct contrast present between (1) the efforts and success in producing sound technical assessments, drawing samples, administering the assessments, and analyzing the assessment data, and (2) the efforts and success in disseminating the assessment results. For example, in the National Research Council’s *Grading the Nation’s Report Card* (Pellegrino, Jones, and Mitchell, 1999), the topic of score reporting was addressed in commentary dispersed throughout the book and not highlighted in a dedicated chapter as was done with curriculum frameworks, test design, standard setting, and other key topics.

Explicit in the mission of NAEP is its charge to communicate NAEP’s results to various stakeholder groups. Score reporting for NAEP involves disseminating assessment results of the different NAEP assessments to interested audiences and is an impressive effort that involves a number of members of the NAEP Alliance as well as NCES and NAGB. Furthermore, NAEP reports are not provided for individuals but rather in aggregate for various groupings of students based on geography (such as the nation, students in national public schools, the states, and Census Bureau regions) and demographic and other categories (such as gender, membership in racial or ethnic categorizations, language status, and parents’ education level). This is no small undertaking.

Unfortunately, across agencies and contexts for testing, score reporting is an area that unfortunately is often a postscript at best, to the test development process. This is regrettable because clear and logical dissemination of test results promotes valid score interpretation and advances the intended consequences of an assessment program, as underscored in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). Specifically, Standard 5.10 states:

When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what

the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used. [p. 65]

Other *Standards* with great relevance to the reporting of NAEP include the following:

Standard 1.1: A rationale should be presented for each recommended interpretation and use of test score, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation. [p. 17]

Standard 1.2: The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described. [p. 17]

Standard 13.14: In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores. [p. 148]

Standard 13.15: In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of those differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation. [p. 148]

Standard 13.19: In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions. [p. 149]

These particular *Standards* address the nature and role of score reporting and reinforce the relationship between the methods used to communicate test results to stakeholders and validity. The *Standards* (AERA, et al., 1999) define validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9) and “the most fundamental consideration in developing and evaluating tests” (p. 9). Given this weight, it seems critically important to prioritize the present evaluation of the dissemination of results within the context of conducting an evaluation of NAEP.

Concerns specific to NAEP data reporting have been documented in numerous studies and articles over the past 15 years, ranging from inquiries about the data needs of constituent groups to formal research studies of graphical displays of results. These topics have been addressed in Levine, Rathbun, Selden, and Davis (1998), Hambleton and Slater (1995), Jaeger (1992, 2003), Koretz and Deibert (1993), Linn (1998), Linn and Dunbar (1992), Wainer (2000b, 1997, 1996), Wainer, Hambleton, and Meara (1999), De Mello (2004), Ogilvy Public Relations Worldwide (2004), the National Research Council (2001), Simmons and Mwalimu (2000), and Hambleton (2002). NAEP’s score reporting practices are the focus of unprecedented scrutiny due to the *No Child Left Behind* legislation and receive keen interest from many policymakers, educators, and the public with respect to state-to-state comparisons of educational achievement gains and status.

Purpose of the Utility Study

The purpose of this study was to evaluate the utility of NAEP reports. Considerable thought has gone into NAEP'S reporting practices over the past 15 years. Evidence from the literature indicates that (a) reporting efforts have been adjusted over time to reflect the growing interest among different constituent groups in NAEP results (particularly with respect to the introduction of achievement levels reporting), and (b) program leaders from both the U. S. Department of Education's National Center for Education Statistics (NCES) and the National Assessment Governing Board (NAGB) have tried to make these changes informed by both empirical studies and opinion research, but questions remain about how well NAEP reports are understood by interested users. How effective are NAEP's reporting practices given the current reporting strategies and policy goals of NAEP? Are NAEP's audiences able to understand NAEP results, as communicated in print and on the Web? Can audiences make reasonable inferences from these reports and avoid inappropriate ones? What do they think about the information they are being given? Is it important to them?

This study systematically considers NAEP reporting efforts given the increasing use of Web-based communications and five years past implementation of *No Child Left Behind*. If the intended users of NAEP results are unable to understand the reported results or they experience frustration with the reporting methods, then the impact of NAEP may be considerably weakened, per the *Standards* (AERA, et al. 1999).

In addition to evaluating the current utility of NAEP reports, another goal of this study is to offer empirically developed recommendations that can be used to enhance the reporting of NAEP results as the assessments themselves and the technology used to communicate results evolve.

The following research questions were identified to guide this utility study of reporting practices associated with the National Assessment of Educational Progress in the age of *NCLB*:

1. How do users of NAEP information regard the NAEP information available on the Web, including the online, interactive data tools?
2. How well are the current paper versions of NAEP reports and displays functioning with their intended audiences? Are stakeholders able to understand the information they are being presented with and use it to answer basic questions?
3. What are the reporting interests and preferences of NAEP audiences?

Table 1 includes a listing of the studies undertaken to address each of these research questions. Some studies provided information relevant to more than one research question.

Table 1. Utility Study Activities by Research Question

Research Question	Study Focus	Method(s)	Participants/Source Materials	Full Study
1.) How do users of NAEP information regard the NAEP information available on the Web, including the online, interactive data tools?	Usability: NAEP Web site	Observations (directed) Observations (undirected)	13 state/district education personnel/policymakers	Appendix A
	Usability: NAEP Data Explorer	Observations (directed) Observations (undirected)	9 state/district education personnel/policymakers 5 Ph.D. education researchers	Appendix B
	NAEP Web site usage statistics	Document review	Documents from NCES and Webtrends	Appendix C
	Web site evaluation methodologies	Literature review	Psychometric and usability literature	Appendix D
2.) How well are the current paper versions of NAEP reports and displays functioning with their intended audiences? Are stakeholders able to understand the information they are being presented with and use it to answer basic questions?	NAEP graphical/data displays (Math)	Focus group	8 state math content specialists	Appendix E
	NAEP graphical/data displays (Reading)	Focus group	8 state reading content specialists	Appendix F
	NAEP graphical/data displays (both Math and Reading)	Focus group	4 state/local education personnel	Appendix G
3.) What are the reporting interests and preferences of NAEP audiences?	Reporting interests, Math	Focus group	8 state math content specialists	Appendix E
	Reporting interests, Reading	Focus group	8 state reading content specialists	Appendix F
	Reporting interests, general	Focus group	4 state/local education personnel	Appendix G
	Reporting interests, general/Web	Interview questions	9 state/district education personnel/policymakers	Appendix A
	Reporting interests, Web use	Document review	Documents from NCES and Webtrends	Appendix C

In the remainder of this section we briefly describe recent research on NAEP reporting as well as the reporting policies and practices established by NAGB and NCES. Following that, we report on our research describing the findings and suggestions for improvements and future research associated with studies of the NAEP Web presence. This topic area is concerned with the NAEP Web site, with particular focus on how users interact and seek out information. The next section focuses on NAEP data displays along with recommendations emerging from that aspect of the study. These displays depict the statistics and graphics used by NAEP to communicate results to stakeholder groups. The displays of interest are drawn from NAEP reports published within the last three years. Then, we provide an overview of our research on the reporting needs of different NAEP constituent groups, as well as next steps in that regard. The last section of this document offers a summary of the overall findings of this report and our recommendations for future practice.

Review of Previous Research on NAEP Reports

Both with respect to NAEP and other large-scale assessment programs, research suggests that scales and score reports issued by different testing agencies are not fully understood by their intended audiences. As noted by Hambleton and Zenisky (in preparation), the myriad reporting scales used on countless tests are confusing to many prospective users of test data. Results from recent studies of adult literacy indicate that only 13 percent of U.S. adults scored in the *Proficient* range for quantitative literacy.²⁸ The challenge of presenting results to different stakeholders is further compounded by evidence indicating that many involved in testing are unfamiliar with the seriousness of score reporting challenges or with the relevant literature guiding the process of score reporting. A recent report by Goodman and Hambleton (2004) highlighted multiple problematic elements of individual score reports distributed to students, parents, and teachers by various states and national test publishers at the time.

The topic of score reporting is not an issue to be considered from psychometric and statistical perspectives in a vacuum devoid of consideration of the target users. A well-designed report of test results involves reflection on the intended audience and the information to be communicated. It requires expertise in testing, graphic design and layout, public relations, and psychology. Books by Cleveland (1994), Tufte (1990, 1997, 2001, 2006), and Wainer (2000a) provide numerous examples of how data can be represented well in graphical form, and how ill-considered figures can mislead and obscure interesting results. A further framework for studying the issues associated with good reporting practices is cognitive load theory. Cognitive load can be understood as the processing ‘burden’ on working memory during problem solving, thinking and reasoning (including perception, memory, language, etc.) and can be broken out into intrinsic cognitive load, germane cognitive load, and extraneous cognitive load (Pass, Renkl, and Sweller, 2003). For users of NAEP reports and the NAEP Web site, understanding the information displayed presupposes not only statistical knowledge about test scores in general, but also programmatic information about NAEP and prose, document, and quantitative literacy to be able to process data presented in different ways. Another approach to cognitively managing the kinds of pictorial and verbal information that NAEP reporting materials include is described in recent work by Mayer and Moreno (2003). Interdisciplinary methodologies such as these offer implications for learning in a multimedia context and appear particularly promising.

The publications and practices employed by NAEP have, however, received some attention from educational researchers. In studies dating from the 1990s, some reviews of NAEP reports have shown some graphical missteps (e.g., Hambleton and Slater, 1994; Wainer,

²⁸ U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2003 *National Assessment of Adult Literacy*.

Hambleton, and Meara, 1999), but recent work has also found clear improvements in the reporting methods and clarity of displays (Hambleton and Meara, 2000). In Hambleton and Slater (1994), many policymakers interviewed were unable to read major sections of the NAEP Executive Report of the 1992 National and State Mathematics Results. Problems included (1) confusion about the basic NAEP score scale (what in the world is the meaning of a score of 300?), (2) failure to distinguish anchor points and achievement levels, (3) lack of knowledge by policymakers of even basic statistics, which caused problems in interpreting significant differences, confidence bands, and other statistical information, (4) confusing graphics (such as the "panty-hose chart"). Recommendations made on the basis of this research included a) charts, figures, and tables should be stand-alone and understandable without in-depth text-based explanations, b) the critical need for field-testing of data displays, c) the desirability of high-quality, high-resolution graphics to ensure quality in later reproductions, d) minimal complexity in graphs, figures and tables, e) simplified, jargon-free, visually interesting executive summaries focused on a specific message, and f) careful consideration of intended audiences for individual publications. The Wainer, Hambleton, and Meara (1999) study was similarly disappointing in its findings that policymakers and educators were misinterpreting displays from the 1994 NAEP Executive Report in Reading (though the Reading Report of 1994 appeared to be a substantial improvement over earlier reports, according to the researchers themselves).

Hambleton and Meara (2000) reviewed more than 500 newspaper accounts of NAEP results from 1990 through 1998, reporting clear changes in how NAEP was reported over that eight-year span, with the trend toward more informational materials being issued (including content information and sample items, as well as figures, graphs, and tables) and more newspaper-like documents (as contrasted with technical report-style tomes). The results that the media reported seemed to be interpreted and explained accurately for the most part, although the authors noted that some confusion about the meaning of the achievement levels persisted, especially when media representatives tried to expand reporting efforts beyond the NCES/NAGB-provided materials (for example, describing *Basic* students as "basically competent"). Assessment jargon also emerged as a common source of misunderstanding for journalists (and hence, their readers).

Increased uses of anchor points, achievement levels, benchmarking, market-basket displays, etc., have been increasingly used and studied for use, as they represent promising and useful approaches for improving NAEP displays to further enhance the NAEP reports (Jirka, 2007). Hambleton (1998, 2002) and Jaeger (2003) discussed some of these approaches in the NAEP context, and the National Research Council (2001) explored the possibility of district-level and market-basket reporting for NAEP, as did Mislevy (1996).

Some attention has also been paid to identifying NAEP's audiences and the kinds of information these specific stakeholder groups require. Jaeger (2003) listed seven groups of NAEP audiences, including various individuals at the federal level (the Executive Branch and the Congress, including staffers), the state level (the Executive Branch and the state legislatures), local districts, local schools, the general public, members of the press, and educational research personnel.

Simmons and Mwalimu (2000) explored the reasonableness and information value of NAEP achievement levels reporting to governors' and states' legislative staff members, state assessment personnel, public and private educators, administrators, parents, business leaders, and education policymakers. With respect to the criteria of reasonableness, findings varied between the different constituent groups, with the policy and content descriptions of the NAEP achievement levels being found to be largely acceptable by legislative staff members. However, state assessment personnel and educators were sometimes confused between state and NAEP performance categories. They also tended to believe the expectations of performance in the

Proficient and *Advanced* categories were unrealistically high, which raised concerns for them about the appropriateness of the NAEP levels in general. Also, using focus groups, Simmons and Mwalimu (2000) found that although users liked the achievement levels, they found them hard to interpret. This finding led Simmons and Mwalimu to conclude that a great deal more effort should be made in aiding interpretation of NAEP results, including the provision of contextual information and (when possible) connections to state results.

Levine, et al. (1998) detailed the findings of surveys and focus groups involving hundreds of participants from a wide array of stakeholder groups at the state and local levels (both legislative and educational), as well as the press, the public, and national business organizations. Among their main findings was that these audiences were interested in receiving results in a timely fashion (recognizing that faster delivery of key results might mean less exhaustive analysis using all of the extensive background variables available in NAEP). Also, these groups expressed a clear preference for results for reading and writing, and also wanted subscale results for mathematics. Levine, et al. (1998) also found support for annual administration of NAEP and considerable interest in linkages between NAEP and both international and state assessments.

Recently, in looking at NAEP reports from 2003 and select NAEP Web pages, a study by Ogilvy Public Relations Worldwide (2004) identified a number of positive strategies that NAEP employs in its reporting practices. However, the report also provided numerous additional recommendations for improvements to meet the reporting needs of the constituent groups. These recommendations included streamlining and simplifying the language of and overall document styling for NAEP summary reports and for the NAEP Highlights documents. With respect to the NAEP information on the Internet, an initial release site for the direct communication of broad-interest results was among the main points. The Ogilvy report is a significant addition to the literature on NAEP for including consideration of elements of the NAEP Web site, and that report is complemented by technical usability studies completed by De Mello (2004) and Government Micro Resources, Inc. (2004, 2005a, 2005b).

NAEP Reporting: Summary of Current Policies and Practices

NAEP reporting, as with other aspects of the testing program, is a collaborative effort of NAGB and NCES, in which policy for reporting is set by NAGB with a subset of the board's members serving on the Reporting and Dissemination Committee. Among the recent activities of this committee is the development of a formal Policy Statement and Guidelines on Reporting, Release, and Dissemination of NAEP Results.

The most current version of the NAEP Policy Statement (National Assessment Governing Board, 2006a) is available online (<http://www.nagb.org/release/policy06.doc>), and describes in significant depth the policy principles for NAEP regarding report preparation and content, the public release of NAEP results in both paper and Web-based formats, and dissemination and outreach. The report preparation and content policies define the primary audience for NAEP as the American public, and state that the initial public release of the NAEP results is a printed summary report and a dedicated Web site now known as the Initial Release Site (<http://www.nationsreportcard.gov>). This section also governs reporting efforts with respect to the use of straightforward data reporting and requires that results should be provided for the nation, states, and school districts (disaggregated by subgroup as data permits). Achievement levels, average scale scores, and percentile distributions are among the reporting statistics that should be included. As to the public release of NAEP results, this section establishes the role of NAGB in scheduling data releases, including the manner of releases. Lastly, dissemination and outreach policies include prescriptions for the distribution of results through the media, the

Internet, and other publications. These principles also provide for the results of NAEP assessments to be distributed widely to business, education, labor, civic, and other groups, with materials appropriate for the intended recipients.

The Guidelines can be found at <http://www.nagb.org/release/guidelines06.doc>, and were developed with the intention of offering “additional direction for the content and organization of the initial release of NAEP results in print and on the World Wide Web” (National Assessment Governing Board, 2006b). These Guidelines set expectations about the content and format of executive summaries, navigation of reports as a whole, the information design and layout of the reports, and presentation of results (notably, with both text and data or graphics components) for the printed NAEP reports. The use of sample questions to illustrate achievement levels is recommended. As to the Initial Release Site for Web dissemination, the navigation structure of the site is outlined in the Guidelines (including a listing of elements to be included, such as report cards, state profiles, methodology, and information centers for parents, researchers, the media, and educators), and the appearance of the home page is detailed. Other elements of the Guidelines for Web reporting focus on design and layout, Section 508 accessibility for persons with disabilities, and public relations efforts for promoting the www.nationsreportcard.gov site.

In addition to the formulation of the Policies and Guidelines for reporting, NAGB has paid considerable attention to specific aspects of NAEP reports in recent years. As reflected in the Policies and Guidelines documents, considerable effort has been put forth in ensuring that NAEP reports use lay language, increase the user-friendliness of the design, distribute results more widely, and schedule briefings of results more frequently. The use of communications firms such as Hager Sharp and Ogilvy Public Relations Worldwide (each with different roles and levels of involvement at different stages in reporting discussions) is indicative of the importance with which NAEP reporting is regarded among program leaders. NAGB has also expressed considerable interest in learning more about how NAEP is reported throughout the news media via the use of media reviews. In these reviews, appearances of NAEP in print, online, and television news outlets are tallied and summarized post-release. Other discussions of reporting involving NAGB in recent years have focused on challenges associated with achievement level reporting and the release plans for different assessments.

NAEP on the Web

Overview of Research

As noted previously, an increasingly key strategy for communicating NAEP's results to its constituents is the NAEP presence on the World Wide Web (<http://nces.ed.gov/nationsreportcard/>). By making vast quantities of data and information available to be accessed at the convenience of the site's users, NAEP is responding to the processing needs of its stakeholders and being proactive to ensure its continued significance as the Nation's Report Card. At the same time, as with traditional, paper-based score reporting (both with respect to NAEP and otherwise), Web-based reporting methods need to be critically evaluated for their information value and ease of use.

The primary research question guiding this aspect of the Utility study concerns how users of NAEP information regard the NAEP information available on the Web, including online data tools. Activities in this portion of the Utility study were designed to address this question by considering the NAEP Web site with respect to the following dimensions:

- a) frequency of stakeholder use of the NAEP Web site;
- b) the preferences and types of information accessed by different user groups;
- c) user impressions of the navigability and overall accessibility of the Web site; and
- d) the usefulness and functionality of interactive data tools on the site.

As a priority in the *Comprehensive Evaluation of NAEP*, gathering insight from users of the Web site relative to these aspects of the site can provide NAEP with a broad picture of the user experience, and enable NCES and NAGB to continue to offer data and information in useful ways to NAEP's constituents via the Web. Furthermore, given the prominence of large-scale testing in American education and the increasing use of the Web to transmit information to interested parties, consideration of Web-based reporting practices with respect to the *Standards* (AERA, et al., 1999) is an important area for study.

A tremendous amount of information about NAEP is available on the Web. There are numerous techniques that can be employed to evaluate Web sites such as NAEP's (U.S. Department of Health and Human Services, 2006), and several of these strategies were used in the course of preparing this portion of the Utility study.²⁹ First, usage statistics of visitors to the NAEP Web site and their interests (collected by the NCES vendor Webtrends' tracking software) were examined. Next, users were observed while navigating the NAEP Web site in one-on-one interviews, where participants explored different sections of the NAEP Web site using the Treasure Hunt approach (i.e., users complete specific tasks) employing the think-aloud strategy, and were observed while doing so. Finally, one of the data tools available on the NAEP Web site, the NAEP Data Explorer, was examined in some depth by individuals in one-on-one, undirected observations, also with a think-aloud component. In the remainder of this section, the findings from each of these aspects of the Utility study are summarized and their implications for the NAEP Web site and broader NAEP reporting strategies are discussed.

A Brief Review of NAEP on the Web

For the National Assessment of Educational Progress (NAEP), the Internet is a primary means by which interested audiences can access test results. Much information about the NAEP testing program, including multiple years' worth of results for a number of content areas

²⁹ Our review of web site evaluation methods (Appendix D) identified eight methods used in web site evaluation studies, including tracking software, online/paper survey, one-on-one interviews, one-on-one contextual observations, one-on-one 'Treasure Hunt' observations, think aloud or Delphi protocol, eye-tracking, and focus groups.

including core subjects such as Mathematics and Reading, is currently available on the NAEP Web site (<http://nces.ed.gov/nationsreportcard/>). While the Web is not the only means by which NAEP results are being disseminated to NAEP's audiences, the increasingly key role of the NAEP Web site as a source for quick access to information about NAEP is unsurprising in today's world. NAEP's presence on the Internet is evolving and expanding steadily. Evidence for this includes the creation of an Initial Release Web site (<http://www.nationsreportcard.gov>) for special events such as the 2005 fourth and eighth grade Mathematics and Reading results release as well as ongoing efforts with respect to developing Web-based data analysis tools such as the NAEP Data Explorer (NDE).

From the main NAEP Web site, stakeholders interested in NAEP data can get access to a wide range of information. These available resources fall into four main categories: programmatic Web pages, static data-oriented Web pages, interactive or media tools, and downloadable PDFs of paper-based NAEP reports that have been released over the years. For the purposes of categorizing the resources on the site, we define these categories as follows.

- *Programmatic Web pages* are text-based resources accessible by branching off the main NAEP page. These programmatic pages are explanatory in nature and do not contain assessment data/results.
Examples: Overview (<http://nces.ed.gov/nationsreportcard/about/>), Frequently Asked Questions (<http://nces.ed.gov/nationsreportcard/faq.asp>), the NAEP Inclusion Policy (<http://nces.ed.gov/nationsreportcard/about/inclusion.asp>), the Site Map (<http://nces.ed.gov/nationsreportcard/sitemap.asp>)
- In contrast, *static data-oriented Web pages* provide assessment findings in structured tables, charts, and text formats that web site users cannot manipulate.
Examples: Long-Term Trend Summary Data Tables for 2004 (accessed via http://nces.ed.gov/nationsreportcard/ltr/results2004/2004_sdts.asp), State Profiles (accessed via <http://nces.ed.gov/nationsreportcard/states/>), the Trial Urban District Assessment results for 2005 (accessed via http://nces.ed.gov/nationsreportcard/nrc/tuda_reading_mathematics_2005/t0018.asp?printver=)
- *Interactive or media tools* are defined by a high degree of user choice in generating what results or analysis are called up to be displayed on a page: we refer here to the use of multimedia and clickable data resources which, for example, might allow users to manipulate the format (tables or graphs), information (scale scores, proficiency levels, percentiles), and type of results displayed (national, state, subgroups, gaps, etc.).
Examples: NAEP Data Explorer (<http://nces.ed.gov/nationsreportcard/nde/>), NAEP Question Tool (<http://nces.ed.gov/nationsreportcard/itmrls/>), the State Comparisons Tool (<http://nces.ed.gov/nationsreportcard/nde/statecomp/>)
- A number of pages on the NAEP site contain links to numerous *downloadable PDFs*, which package information in easy-to-print formats for user review, often in traditional technical report-style layouts with tables of contents.
Examples: the NAEP frameworks documents for assessed subject areas (accessed via <http://nces.ed.gov/nationsreportcard/frameworks.asp>), 150+ Report Cards and other reports in Arts, Civics, Geography, Mathematics, Reading, Science, U.S. History, Writing (1990 – 2005) (accessed via <http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031>)

While the above examples and the accompanying links are only a very small fraction of what is currently available on the NAEP Web site, they provide a sampling of the kinds of resources the users of the site can access at their convenience.

NAEP Web Site Usage

As described in Appendix C, one year's worth of data from the NAEP Web site, as collected by the vendor Webtrends was examined to establish a baseline understanding of the use of the NAEP Web site. This review of several dimensions of Web usage statistics for the NAEP Web site was illuminating in several respects, as it provided a very broad summary of the kinds of pages and information that visitors to *The Nation's Report Card* Web site seek out. At the same time, in reviewing data collected about visitors to any Web site, caution must be taken not to over-interpret results, particularly with respect to the one-year snapshot of use presented here. First, this Web site, like most, is an evolving entity that is constantly maintained and updated. Between March 2005 and February 2006, a number of new features were added and older ones were completely revamped, and several major assessment results were released, so that there is something of an ebb and flow to the counts of visits and views to the site month to month. At best, reviewing data from a single year provides a general pattern site use. In addition, as the NAEP site is a U.S. government site, there are data collection limitations that require a high level of anonymity and aggregation of the results.

Some noteworthy findings from this study include information about the kinds of computers and Web browsers used by visitors to the site (almost 10 percent use the Mozilla Firefox browser), which is significant in that maintaining and improving the functionality of the NAEP site and the tools found there involves being informed about the technology needs of various users, so that all site visitors have a satisfactory user experience and can access information. Furthermore, the NAEP home page was viewed nearly half a million times in the one-year period evaluated. Users exhibited a consistently high level of interest in the state profiles. By this data, individual state profiles were accessed over 230,000 times between March 2005 and February 2006. The results for the interactive online tools available on the NAEP site (the Question Tool, the Data Tool/Data Explorer) indicated that users were increasingly interested in these features, as exhibited by a steady rise in the frequency of use. For the NAEP Data Explorer, activated for only the last five of the twelve months considered here, the main NDE page was viewed over 17,000 times. Among the Initial Release Site results for the five months for which it was active in the year considered here, state results were again among the pages most accessed, as were student group results.

This review of NAEP site usage data identified a number of key directions for further operational review of the use of the NAEP site:

- Given the high volume of use of the State Profiles, these should be a priority for future conversations with stakeholders: What information are visitors to these profiles most interested in? Is there additional data that should be included? Is the information represented and displayed in the most useful/effective ways?
- Use of the Question Tool is growing, and to the extent possible it would be informative to learn more about which NAEP stakeholder groups are and are not accessing this tool, and why.
- The high level of interest in the displays of results for student subgroups likewise suggests areas for further study. What information on those pages are visitors focusing on? How are subgroup results displayed, and do different user groups understand and use different displays appropriately?

- Efforts might turn to the Initial Release Site and obtaining user impressions of that page.
- Lastly, use of the interactive NAEP Data Explorer tool seems to be growing, and it will certainly be informative to learn more about how users work with the tool and for what purposes.

Ultimately, developing an understanding of what pages and information are of interest to the aggregate of visitors to the NAEP site (as provided in this report) has much practical value for subsequent utility study activities, such as observations of individual users navigating the Web site and focus groups convened to discuss aspects of the site.


Site User Observations

The goal of our Site User Observations was to elicit opinions about the NAEP Web site (<http://nces.ed.gov/nationsreportcard>) from a range of individuals involved in educational assessment. Participants were asked to provide feedback about multiple aspects of the NAEP site, including the content and navigation of the home page (and links from that page), the State Profiles pages, and the site in general. This study took place June 26 and 27, 2006, in San Francisco, Calif., during the Council of Chief State School Officers' Conference on Large-Scale Assessment, and involved individuals from state and district education offices and policymakers. During the meeting, participants were asked a) to spend time on the NAEP home page and give their impressions of the organization and structure of that page and to navigate links off that page of their own choosing, while providing running commentary on what they saw and did (these were *undirected* observations, designed to gather information about user interests and perceptions without constraining users), b) to reflect on the information and navigational ease of the NAEP State Profiles pages, and c) to complete several brief but specific common tasks (*directed* observations), all while providing impressions of the experience via a think-aloud protocol. Examples of these tasks included finding information on NAEP's policies on accommodations, providing a brief summary of selected results from recent assessments including Mathematics 2005 and Science 2005, identifying the NAEP achievement levels, and accessing released NAEP items.

A summary is provided below of the findings from the site user observations (results in full are in Appendix A):

The NAEP Home page. As the starting place for many user visits to the NAEP Web site, the NAEP home page (<http://nces.ed.gov/nationsreportcard>) was generally well perceived. Opinions on the appearance of the home page were largely positive ("not too busy— there's a lot of stuff but there's a lot to NAEP", "a table of contents feel"). Some participants asked questions about how items were selected for inclusion or prominence on the home page's "real estate" noting the presence of some content overlap, such as with the 2005 Science results. As shown in a capture of the home page as viewed during this study (Figure 1), Science 2005 appeared on both the "big splash" but also under "New and Noteworthy." For some users this apparent redundancy raised questions about the "management of news about NAEP", meaning that those results were what "we're supposed to care about and look at, not other stuff."

Figure 1. The NAEP Home Page (captured June 26, 2006)



Institute of Education Sciences U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION STATISTICS

NewsFlash Staff Contact Site Index Help

Search NCES

Publications & Products

Surveys & Programs

Data Tools


Tables & Figures

Fast Facts

School, College, & Library Search

Annual Reports

What's New?



ABOUT NAEP...

- [overview](#)
- [current activities](#)
- [long-term trend](#)
- [high school transcript study](#)
- [special studies](#)
- [selected schools](#)
- [parents](#)
- [researchers](#)
- [media](#)
- [educators](#)
- [policymakers](#)

SUBJECT AREAS...

- [civics](#)
- [geography](#)
- [mathematics](#)
- [reading](#)
- [science](#)
- [u.s. history](#)
- [writing](#)
- [other subjects](#)


Search N.

HELP SITE MAP CONTACT US GLOSSARY NEWSFLASH

SAMPLE QUESTIONS | ANALYZE DATA | STATE PROFILES | PUBLICATIONS

National Assessment of Educational Progress

THE NATION'S REPORT CARD




Results of the 2005 National and State Science Assessment

NOW AVAILABLE

INSIDE NAEP

Results of the National Indian Education Study Now Available!



Explore the [results](#) of the National Indian Education Study (NIES), Part I: *Performance of American Indian and Alaska Native Fourth- and Eighth-Grade Students on the 2005 NAEP Reading and Mathematics Assessments.*

NEW & NOTEWORTHY

Results of the 2005 national and state science assessment were released May 24, 2006. View the archived [webcast](#) of the data release event held in Washington, DC. Read the [transcript](#) of the StatChat—an online discussion with NCES Associate Commissioner Peggy Carr—about the results.

The 2007 NAEP Secondary Analysis Research Program is receiving applications until July 27, 2006. See [more information and application forms.](#)

Are you going to the Council of Chief State School Officers (CCSSO) National Conference on Large-Scale Assessment? Check out a [list of NAEP and NAEP-related sessions.](#)

Last updated 16 June 2006 (AA)

NCES Headlines

- ▶ [NEW REPORT! - Dropout Rates in the U.S.: 2002 & 2003](#)
- ▶ [NEW REPORT! - Average Freshman Graduation Rates](#)

- ▶ [Profile of Undergraduates in Education Institutions: 2003-04](#)
- ▶ [Documentation for the NCES Comparable Wage Index Files](#)


[Pubs/Products](#) | [Surveys/Programs](#) | [DataTools](#) | [Tables/Figures](#) | [FastFacts](#) | [School/LibrarySearch](#) | [Annuals](#) | [What's New?](#) | [Kids Site](#)

Search NCES

[NewsFlash](#) | [Staff](#) | [Contact](#) | [Site Index](#) | [Help](#)

[Privacy & Security Policy](#) | [Statistical Standards](#) | [RSS](#) | [FedStats.gov](#)

[Institute of Education Sciences](#)
[U.S. Department of Education](#)



ies NATIONAL CENTER FOR EDUCATION STATISTICS

1990 K Street NW, Washington, DC 20006, USA, Phone: (202) 502-7300 ([map](#))

Participants who clicked on the 2005 Science results in the middle of the page were transported to the Initial Release Site (IRS). While a direct evaluation of the IRS was not a part of this study, users who accessed the IRS did note that they liked the look and feel of the page (shown below in Figure 2) as compared to the main NAEP Home Page (Figure 1).

Figure 2. The www.nationsreportcard.gov home page (Dec. 27, 2006)



Across the participants in this study, other comments and suggestions about the NAEP home page (Figure 1) were received and are listed in brief below:

- Left navigation menu:
 - *Policy: Revisit the audience categorizations, as users wanted to know how groups were identified for inclusion, considered 'educators' too broad a grouping, and asked why there was no link for the general public.*
 - *Operational: Group High School Transcript Study and Long-Term Trend under Special Studies, rather than separate links.*
 - *Operational: Prioritize Reading, Mathematics, and (maybe) Writing in the left menus, as they are the content areas of greatest importance per NCLB, then list "Other Subjects."*
- The presence of multiple search boxes:
 - *Operational: Situate the NAEP search more prominently, because users were confused between the regular NAEP search and the two NCEs searches.*

- NCES/IES links/page header:
 - *Operational: Reduce size and prominence of header, as users found this distracting.*
- The repeated labeling of items as “New” and/or “Noteworthy” on the home page:
 - *Policy: Consolidate aspects of the page or develop alternate terminology to highlight different elements/information across the NCES/IES header and the NAEP-specific portion.*

One significant navigation issue relating to the home page independently raised by nearly all participants involved how to return to the NAEP home page once users had followed several links to elsewhere on the site. Most users were able to return to the home page by clicking ‘back’ on the Web browser but expressed frustration at not seeing a clearly marked “Home” link. Though such a link does exist, most users wholly overlooked it. One participant asked, “Why is it in the middle? Society has trained us to expect a site to look a certain way.... Look up top and left for home page buttons,” while another commented, “Links should look like links.” Below is the recommendation regarding the NAEP home page link:

- The link to the NAEP home page:
 - *Operational: Redesign the link to look more like something clickable, make it more obvious or prominent on the page, and move the link to the top and left.*

State Profiles. As part of the semi-directed portion of the observations, all users were asked to click on the “State Profiles” link on the NAEP home page, and on the page that loaded from that link, to select any state’s results to explore and provide feedback on. In most cases, users chose their home state. Overall, users liked the content and layout of the State Profiles and found the information contained there to be consistent with their knowledge or experience. They also appreciated the inclusion of historical performance on NAEP scale scores and achievement levels, as well as both tables and graphs. Some reactions to and issues raised with respect to the State Profiles are provided below:

- Content of profiles:
 - *Operational: Update the state data in the State Profiles (at the time of the study (in June of the 2005–06 school year) the state data were Common Core of Data school information from the 2003–04 academic year).*
 - *Operational: Denote results on the page as reflecting significant change within state, from administration to administration.*
 - *Operational: Include the percent of students below Basic in achievement level results (“It would be good to fully illustrate all categories”).*
 - *Operational: Provide background information, as in student characteristics (Number enrolled, percent in Title I schools, etc.), racial or ethnic background, and school or district characteristics when National Public is chosen as a jurisdiction from the drop down menu.*

Another navigation issue that emerged through these observations was the difficulty encountered by participants in accessing the Cross-State Comparison Maps from within the State Profiles. The Profiles include a link that suggested to users that they could click and get the maps, but the subsequent page that loaded was a table of checkmarks that were not obviously links to participants (Figure 3). This resulted in more than a few participants commenting that they didn’t understand why they were not seeing maps, and only those users who guided the mouse (often by accident) over the checkmarks realized that those were clickable. Despite the

presence of instructions at the top of the page (which were not likely read by participants), the checkmarks in the table did not clearly appear to the participants in this study to be links.

- The link to the Cross-State Comparison Maps:
 - *Operational: Revise links on cross-state comparison maps to conform to Web standards for accessibility.*

Figure 3. Accessing the Cross-State Comparison Maps

State Profiles
The Nation's Report Card (home)

Cross-State Comparisons, Average Scale Scores: Massachusetts

Cross-state comparisons are available in two formats below. Select jurisdiction... ▾

Click on a symbol in the chart below to view a Scalable Vector Graphics ([requires SVG viewer](#)) comparison map.

For more information about SVG graphics and instructions on how to cut and paste SVG images please see our [help page](#).

✓ participated
✗ did not participate or did not meet minimum participation requirements

Subject	Grade	Year								
		1990 ⁿ	1992 ⁿ	1994 ⁿ	1996 ⁿ	1998	2000	2002	2003	2005
Mathematics	4		✓		✓		✓		✓	✓
	8	✗	✓		✓		✓		✓	✓
Reading	4		✓	✓		✓		✓	✓	✓
	8					✓		✓	✓	✓
Science	4						✓			✓
	8				✓		✓			✓
Writing	4							✓		
	8					✓		✓		

ⁿ Accommodations were not permitted for this assessment

Also received through the observations were some comments specific to findings and using data and information from the Trial Urban District Assessment program. These comments and accompanying suggestions are provided below:

- TUDA profiles:
 - *Policy: Link TUDA on the home page or more prominently from one of the main links, because they found it difficult to find information about TUDA (not listed under*

Special Studies and as previously mentioned participants found the search function frustrating).

- *Policy: Provide more information about the urban districts, including a map analogous to the one on the main State Profiles page with the urban districts involved in NAEP listed so that people could click on and get information similar to what is available for states via the State Profile page.*

Conclusions. Overall, participants expressed positive impressions of the NAEP Web site. They were impressed both by the quantity of information and the quality of effort put into the NAEP site. At the same time, participants raised several navigational issues that have important implications for the user experience. As to the NAEP home page, users identified several elements that might be reconsidered, including the content and structure of the left navigation bar and the presence of three search text boxes (and only one of them NAEP-specific). The ability of users to quickly and easily return to the NAEP home page emerged as a source of difficulty in navigating the site: while there is a “home” link most users did not notice it and found it frustrating to have to click the “Back” button on the browser multiple times in order to return to the home page. Participants also had suggestions about the content, appearance, and navigation of the State Profiles pages. They asked that more information, including the denoting of statistically significant change within states and the reporting of discrete achievement level percentages for all categories, be provided. While the cross-state comparisons maps were greatly liked by participants who appreciated their interactive nature and their usefulness in illustrating between-state differences, the difficulties in accessing the maps curtailed their use. Another access issue was identified with respect to finding TUDA information on the site.

Focus on the NAEP Data Explorer (NDE)

The current online NAEP Data Explorer built on an earlier data tool available via the NAEP Web site, the NAEP Data Tool (NDT). The NDE was released to the public at the time of the Mathematics and Reading 2005 release, in October 2005. As part of efforts to evaluate this aspect of the NAEP Web site, two utility study activities were undertaken. First, several participants in the Web site user observations that took place during the 2006 Council of Chief State School Officers’ (CCSSO) conference in San Francisco were asked their impressions of the NDE (these participants were state and district education personnel). A second, targeted study of the NDE took place at the annual meeting of the Northeastern Educational Research Association (NERA) conference, held in Kerhonkson, N.Y., in October of 2006. This study of the NDE was targeted in the sense that the population of interest in this case was postdoctorate educational researchers, in order to learn more about the usability of the NDE among users with advanced training in educational statistics and data manipulation. The results of both of these sets of observations, described in brief below, and summarized in Appendix B, provide considerable insight into the user experience and likewise identify several important recommendations for improvement.

CCSSO Observations. Of the U.S., state, and district personnel in this study, this updated version of the NDE was unfamiliar to most. Of those who reported knowing about it, one commented that she “didn’t realize it was this good.” One expressed some frustration at not being aware of the NDE’s existence and a perceived lack of publicity about some of NAEP’s interactive tools, citing her role as an assessment person in a large urban district.

In this portion of the site usage observations, participants were directed to the data NDE welcome page (<http://nces.ed.gov/nationsreportcard/nde/>; Figure 4), and given the instruction to briefly familiarize themselves with the tool and take a few minutes to run a few analyses of their

choosing. All users selected “Quick Start” after scanning the welcome page, and none opted to read either the Quick Start or Advanced introductions or clicked on the tutorial or help links on the welcome page, though later in using the tool several commented about the helpfulness of what one user termed the “info dots” (clickable blue circles with a white letter “i” leading to information/help).

Figure 4. NAEP Data Explorer Start Page



Participants were permitted to choose the demographic variables, jurisdictions, and types of results they wanted to explore. A sampling of the analyses done by several participants is listed below.

- Grade 8, Reading, Colorado, Parents' Education, ELL status
- Grade 8, Math, West region, Students with Disabilities
- Grade 8, Math, Texas, Ethnicity
- Grade 8, Reading, San Diego district, ELL status
- Grade 8, Math, Indiana, Parents' Education Level
- Grade 4, Reading, Wyoming, All Students
- Grade 4, Math, Missouri and DC, Gender
- Grade 4, Reading, Houston, Nation, and Louisiana, Race/ethnicity, All years

In general, the participants reported quite favorable impressions of the NDE. They found it to be “powerful,” “intuitive and great,” and “neat.” They also liked the flexibility provided by the tool to “let you choose what you want to look at” and to “build your own data.” Features such as being able to export to Excel, making graphs, and being told that the NDE was processing while it was gathering the requested data were also much appreciated. One participant wished that an NDE-type tool were available for individual state *NCLB* assessment results.

Several suggestions for improvement were also made in the course of the users’ experience with the NDE. One disconnect that was seen by many of the users was that they did not understand why some jurisdictions and variables “grayed out” when they chose a specific grade or grade or subject area combination.³⁰ For some users with a high degree of familiarity with NAEP’s different national and state samples, this may not be an issue, but across the range of experience with NAEP seen in this group of users, this was a source of confusion. Furthermore, in selecting criteria, including demographic variables and jurisdictions for analysis, one participant expressed a preference for a greater level of flexibility than currently permitted. Specifically, this individual wanted all jurisdictions and variables to appear in both Box 3 and Box 4 (see Figure 5 below), rather than feeling forced into a jurisdiction in Box 3 and a variable in Box 4. For this participant, if that flexibility were allowed, that would “open the possibility of answering more questions with data, it would allow users to define their own cross-tabulations.” For several other users, crosstabs analyses were attempted but participants reported being unable to figure it out and were observed abandoning those analyses; another user had the same feeling about testing for statistical significance. One additional suggestion was to devise a way to graph two variables at once.

³⁰ For example, when a user selects Grade 12 Civics, the only jurisdiction options are National and National Public, because results for individual states are not available for that analysis with this tool.

Figure 5. Analysis Selection Page, Quick Start Mode

1 Grade: ⓘ

Grade 4
 Grade 8
 Grade 12

2 Subject: ⓘ

Civics
 Geograph
 U.S. Histo
 Mathemat
 Reading
 Science
 Writing

3 Jurisdiction(s): ⓘ

National
 State/Jurisdiction
 Urban District
 Region

4 Variable(s): ⓘ

Major Reporting Groups
 All Students (Overall Results)
 Gender
 Natl School Lunch Prog eligibility (3 categories)
 Parental education level (from 2 questions)
 Parental education level (from 2 questions)(2005)
 Parental education level (from 6 questions)
 Public or nonpublic school (5 categories)
 Public or nonpublic school (7 categories) (2002+)
 Race/ethnicity used in NAEP reports after 2001
 Region of the country (2003 and later)
 School identified as charter (National Public)
 School location (2005)
 School location (9 categories) (2005)
 Student classified as having a disability
 Student is English Language Learner (2 categories)
 Student is English Language Learner (3 categories)
 Gaps and Changes in Gaps

5 Year(s): I want to see results for ⓘ most recent year only all years available 16 variables available; 0 selected

What next? Choose a grade, subject, and at least one jurisdiction and variable.

[About NAEP Data Explorer](#) [Important Legal Information](#)

NERA Observations. Among this group of participants, the NDE was largely unfamiliar (three had no prior use while the other two reported only having used it once or twice). As before, these individuals were asked to access the data NDE welcome page, and then told to briefly familiarize themselves with the tool and take a few minutes to run a few analyses of their choosing, followed by completion of a directed task. Below are summaries of the user experience and possible changes suggested by participants (full results in Appendix B). Each point was classified into one of four categories: as being a *Comment* (no suggestion for improvement), as addressing the *Appearance* of elements of the NDE, as relating to the content of the *Information* displayed, or as concerning the *Functionality* of the NDE. Comments were also defined as being directed at Overall Impressions, the Opening Page/Usage Agreement, Variable Selection, Data Analysis and Appearance of Results, Help Links, Statistical Significance, Graphing, Exporting Results to Excel, Advanced Mode, or Regression. Where appropriate, specific improvement suggestions are included:

Comments: The participants in this study were very positive in their perceptions of the NDE during and after participating in the study. Broadly speaking, these users indicated that this was a well-designed tool that provided visitors to the site with specific data-based questions about NAEP performance an opportunity to step outside of the bounds of a testing program's customary paper reports and answer those questions independently. All emphasized that they perceived that the NDE would likely be most appreciated by those with specific research questions (versus random exploration of results), largely due to the sheer quantity of data and

data analysis options available. Features that participants reported valuing included the small help icons strategically placed on the page (denoted onscreen by a small blue circle with a white “i”), the inclusion of the option to obtain statistical significance, and many of the sophisticated features associated with Advanced mode.

Appearance: Users generally found the appearance of the NDE agreeable. At the same time, several suggestions to further enhance the look (and consequently, the usability) of the tool were mentioned by participants, including the following points:

Opening page/usage agreement:

- *Operational: Make the link to the tutorial more prominent.*
- *Operational: Reduce the amount of text on this page.*

Selection of variables:

- *Operational: Reformat the variable selection page to more clearly delineate the sequence of choices.*

Help links:

- *Operational: Reformat/re-orient the label of the Tips button to make it clearer that this is a Help function.*

Regression:

- *Operational: Reformat regression results tables to report the unstandardized regression coefficients first, then the standardized coefficients.*

Information: The feedback received from participants that was described as informational in nature was drawn from questions that participants had about different content (not functional) elements of aspects of the NDE. In many of these cases, users suggested additional clarifying text or links be prominently positioned to aid other users.

Opening page/usage agreement:

- *Operational: Add a simplified, prominently linked, non-interactive FAQ tutorial.*

Opening page/usage agreement:

- *Operational: Create and post a link specifying the features of Quick Start and Advanced modes so that users have a quick reference for making a mode choice.*

Selection of variables:

- *Operational: Make data availability for some grade/subject area combinations more explicit, because most users do not understand why some states/jurisdictions and background variables are “grayed out” in the selection process.*

Selection of variables:

- *Operational: Provide hyperlink text and answer choices for background variables so users can be make informed decisions about using the questions for carrying out analyses using background data.*

Data analysis and appearance of results:

- *Operational: Institute a mouseover or popup link to NAEP score information, such as a scale score range and a list of the achievement levels.*

Data analysis and appearance of results:

- *Operational: Include sample size data when possible.*

Graphing:

- *Operational: Provide simple help text to explain the difference between a full graph and a scroll window when selecting these options to make a graph of results.*

Statistical significance:

- *Operational: Implement additional, simplified documentation explaining the statistical significance tests used (how the test was built, what was the statistic type used was, and what was the alpha level).*
- *Operational: Report Cohen's d and provide text categorizing the effects as small, medium, or large.*

Advanced mode:

- *Operational: Add an explanatory link or popup with clarifying information to understand the counts of variables at the bottom of the page in Advanced mode would be helpful.*

Regression:

- *Operational: Provide a direct hyperlink to further explanatory text explaining the contrast coding carried out in the analyses.*

Functionality: In addition to providing feedback on the appearance and content of the NDE, users of course also reacted to different aspects of how the NDE worked in the course of carrying out analyses. Some of these comments involved features working in a way that the participants described as counterintuitive to what they expected or needed, while others entailed suggestions for things they did not see but would like to have added.

Selection of variables:

- *Operational: Add option for comparing results across grades in Advanced mode.*

Data analysis and appearance of results:

- *Operational: Allow users to request all results at once or provide a checklist letting them choose all, some, or just one, rather than have them click a radio button to change results (scale scores, achievement levels, etc.).*

Data analysis and appearance of results:

- *Operational: Append new results to existing analyses (as in SPSS), rather than having results 'lost' when users clicked on the radio button to change result type, if possible.*

Data analysis and appearance of results:

- *Operational: Make links appear as links, and things that are not links should be distinct and not link-like. Specifically, headers in results tables are blue, which is a typical color for Web links.*
- *Operational: Permit users to sort results tables by the headers.*

Statistical significance:

- *Operational: Embed the significance results in the regular display of results, perhaps appended to the end of previous results.*

Graphing:

- *Operational: Embed the graphs in the regular display of results, perhaps appended to the end of previous results.*

Exporting results to Excel:

- *Operational: Expand the beta-testing of this feature, including with respect to compatibility across various operating system/browser combinations, and when problems are detected program a popup to appear notifying the user of plug-ins needed/action to be taken.*

Regression:

- *Operational: Embed the regression results in the regular display of results, perhaps appended to the end of previous results.*

Among the most frequently received feedback regarding functionality were the comments made about the process of graphing, statistical significance, and regression. In each case, when users selected those options, they expected a seamless process within the open working window, rather than a popup with a lengthy and sometimes not-well-documented series of decisions to be made. Users perceived this process to be redundant to the process of variable selection that they had undertaken at the outset of their analyses.

Conclusions. The NAEP Data Explorer clearly offers visitors to the NAEP Web site a tremendous opportunity to run analyses and answer data questions that in many cases are idiosyncratic to the user and might not otherwise be easily answered through other means of NAEP score reporting. In addition, it permits users to customize results, obtain graphics, and test for statistical significance as needed. This is a unique resource among educational testing programs and represents another way in which NAEP is on the cutting edge of score reporting practices today.

The feedback from participants in the two sets of user observations reported here is remarkably consistent. Both groups reported a high level of overall satisfaction with the tool, finding it useful and somewhat simple to learn to use. As with any Web site or data analysis tool, there was a learning curve, but for the most part users did not encounter significant navigation or logical difficulties in carrying out analyses. Where there were difficulties, these were identified as concerning the NDE's *Appearance*, the *Information* displayed, and the NDE's *Functionality*.

As to *Appearance*, users appreciated the "What's Next?" bar on the Selection page. In some cases, users suggested minor redesigns of links and clickable icons. Similarly, certain aspects of the appearance of the variable selection screen and some result displays were found to be confusing to some users, and these should be followed up on with additional user groups to further determine if edits are warranted.

The participants here were most impressed with the volume of *Information* available for analyses and reported in the results screens of the NDE. Suggestions concerning the content of the pages within the tool involved additional contextual information (e.g., the range of the NAEP score scale, the text of the background variable questions, sample sizes). Other information feedback received has the potential to impact the users' perceptions of the functionality of the NDE: when users saw some jurisdictions and variables "grayed out" without explanation, they thought they perhaps had done something wrong, and in some cases wanted to start over. More, prominently placed explanation is needed, in this case. Users also wanted to know more about the analyses performed and the tests carried out, commenting that if they were to try and use these results in a professional setting (a technical report or a conference presentation) they would need to understand the statistics in full.

Concerning *Functionality*, users identified several possible modifications to the current operating infrastructure of the NDE. Comparing student performance across grades was one such use. In addition, the opening of a new browser window and the sequence of decisions involved in making graphs, testing for statistical significance, and carrying out regression analyses were considered somewhat cumbersome by participants. They wanted the process to be somewhat simplified and to occur within the main browser window. On a related point, the users requested that new results within a session be appended to the bottom of previous results. This is perhaps due to their familiarity with how data analysis software packages such as SPSS and SAS work, where as users carry out new analyses output is organized sequentially within one output window (unless the user chooses to close an output window and open a new one), so this may be a different way of managing results for users.

Summary of Research on NAEP on the Web

Three activity areas comprised the research on the NAEP Web site: analysis of Web traffic data, observations of user behavior on the main NAEP site, and observations of engagement with the NAEP Data Explorer. The main findings and implications of this research are described below.

NAEP Web site Usage. Future research should focus on the State Profiles, the Question Tool, results for subgroups, the www.nationsreportcard.gov initial release site, and the NAEP Data Explorer. Each of these elements was found to generate a comparatively high level of traffic, and as these aspects of the site are seen by many site visitors, learning more about a) the reasons why these features are increasingly popular and b) how well these features meet the information needs of users are important directions for future research.

The Main NAEP Site. The NAEP Web site provides a comprehensive look at the NAEP testing program. The results of this study indicate that there are several ways in which the navigability of the site can be enhanced, especially with respect to how users explore the site beginning with the NAEP home page. Users are used to Web sites being structured in a certain way, and continued research for NAEP may consider additional studies of how the NAEP home page does and does not conform to common Web practice for home page links, menu structure, search boxes, and navigation bars (such as the IES/NCES bar at the top of the screen). Users also provided feedback on the State Profiles, and as noted previously, since these are among the most-visited pages on the site, further research as well as content and navigation changes may be warranted.

NAEP Data Explorer. The NDE is among the most sophisticated features of the NAEP Web site, allowing users to carry out highly complex analyses of NAEP data via the Web. Participants in this research with a range of data handling skills found the NDE relatively easy to use but offered recommendations about the user experience and specific functions of the NDE. As a Web-based data analysis application, users commonly followed a “dive right in” approach to using the tool but wanted to be able to access specific help documentation as questions arose. The streamlining of certain important NDE functions such as regression, graphing, and statistical significance was also requested. To the extent possible in a Web application, users asked that functionalities and output be consistent with other data tools that they are familiar with, such as SPSS and SAS. The findings relating to the NDE may particularly be informed by consideration relative to the *Standards* (AERA, et al., 1999), especially because of the self-directed nature of the tool. Some comments received from study participants indicated that more attention to issues of measurement error (Standard 13.14) might be warranted. In all, these findings offer valuable insight into the user experience of working with the NDE. With additional study, minor modifications, and increased use, NAEP’s NDE can be a model for states and other testing agencies with respect to open access for data understanding, which ultimately should be a primary goal of large-scale score reporting efforts.

The Initial Release Site. Though not a primary focus of this research, some users of the main NAEP site did access this page, and found it to be a quite easily navigated vehicle for communicating NAEP results. A concerted research study to look at the effectiveness and usage of the www.nationsreportcard.gov site should be developed, and where possible lessons learned from that site might be applied to the main NAEP site.

Research. As Web-based score reporting can no longer be said to be in its infancy, a comprehensive, ongoing program of research to evaluate the information value and organization of pages across the entire site and to critically review the operation of the interactive tools on the site is in order. Particularly informative in this regard may be the www.usability.gov Web site and accompanying publication (U. S. Department of Health and Human Services, 2006), as well

as the cognitive processing work of Mayer and Moreno (2003) and Pass, Renki, and Sweller (2003). This agenda could be developed cooperatively by NCES and NAGB in collaboration with the members of the NAEP Alliance.

This page left intentionally blank

NAEP Data Displays

Overview of Research

Audiences for test results are increasingly being provided with large amounts of data about how students are doing, with the expectation that this data will be used to assess student achievement and develop instructional strategies and improvement plans. Test results have the potential to help schools, districts, and states make data-based decisions about instruction and student progress. However, reporting test results to any stakeholder group is challenging because of the need to consider the density, accuracy, and possible misinterpretations of the information to be communicated. As noted earlier, the *Standards* (AERA, et al., 1999) are clear about the need for score reports to communicate rationales for recommended score interpretations and uses (e.g., Standards 1.1, 1.2, and 5.10), as well as information about measurement error (Standard 13.14) and contextual information for groups or group differences (Standards 13.15 and 13.19).

In this portion of the evaluation of NAEP relating to the effectiveness of NAEP reporting methods, the purpose of this aspect of the study was to explore the extent to which state and local education administrators were familiar with current methods of displaying NAEP results, what kinds of inferences they might make on the basis of those displays, and potential design improvements. Guiding this study were the following questions.

- 1) How are NAEP results displayed, particularly in electronic communications with respect to principles of good reporting (e.g., Goodman and Hambleton, 2004)?
- 2) What do users understand and not understand in selected NAEP data displays?
- 3) Are there specific recommendations that can be made to improve the development of alternative displays that may alleviate misunderstandings/misconceptions where they exist?

Current NAEP Data Display Methods

The first step in this research was to review current NAEP materials to develop an understanding of the display methods currently in use. From several recent releases of NAEP mathematics and reading results, including 2004 Long-Term Trend (Perie, Moran, and Lutkus, 2005), 2005 National and State NAEP (Perie, Grigg, and Dion, 2005; Perie, Grigg, and Donahue, 2005), and 2005 Trial Urban District Assessment (Lutkus, Rampey, and Donahue, 2005; Rampey, Lutkus, and Dion, 2005), the following displays were identified as a representative sampling of the tables and figures commonly seen throughout the materials:

- line graphs,
- stacked and clustered column or bar charts,
- clickable state comparison maps of average scale scores and percents of students at or above achievement levels,
- tabs from the NAEP question tool with item text, student item performance, a distractor analysis,
- “pantyhose” charts, and
- item maps.

NAEP Data Displays: Impressions from the NAEP State Coordinators

One important source of information about the extent to which NAEP data displays were understandable to different audiences was provided by feedback from a sample of NAEP state coordinators. Among the responsibilities of the NAEP state coordinators are several critical reporting roles in the states, including promoting understanding of NAEP and its relevance to the state program, enhancing states' capacity to use NAEP data, and promoting assessment literacy. The work of the coordinators is essential to NAEP reporting in the states (and thus to learning about the ways in which NAEP display methods are understood among NAEP's audiences). In January 2006, nine coordinators were interviewed by telephone for the purpose of discussing NAEP reporting strategies. These individuals were asked to think about the graphics, tables, and narratives of NAEP that they were familiar with, and to identify those that were in their opinions the most and least effective for communicating with interested audiences. Several reported that their use of different reporting mechanisms depended on the intended audience or user: visual displays such as the cross-state comparison maps and bar graphs (for both scale scores and achievement level results) were cited by several as a primary mechanism for communicating a lot of information quickly, though coordinators from three states indicated that tables were more familiar and useful for the users in those states. In addition, NAEP state coordinators reported they were often in the position of providing information to parties such as state education leaders and education commissioners as well.

The coordinators were also asked for specific examples of reporting challenges they had encountered. Several coordinators mentioned achievement levels as a particular source of confusion, with respect to 1) connecting the labels of the NAEP achievement levels to what they mean for student performance, 2) how NAEP achievement levels are similar to and more commonly different from the states' own performance categories, and the implications of this for explaining differences with state and NAEP results, and 3) the use or phrasing of "at or above" for reporting. With respect to scale scores, the lack of knowledge of the range of the NAEP score scale was cited as problematic. Five of the nine coordinators were very explicit in raising the issue of communicating when differences were significant, cautioning users about overinterpreting differences in mean scores, and wanting to develop strategies for communicating what is really educationally significant. In their experience, this was a difficult concept for users of the NAEP results to understand. The coordinators noted that often *any* difference is interpreted as real and greatly meaningful, and with the statistics of NAEP, sometimes one-point differences are statistically meaningful and other times they are not.

The clickable cross-state comparison maps were mentioned by one coordinator as emerging as a recent reporting challenge due to the tendency of users to attach too much significance to statistically significant differences without due consideration of the states being compared. Following up on that, many coordinators spoke about how different stakeholder groups in their states were quite interested in establishing how states compared. In some cases, the coordinators found themselves urging caution in these kinds of interpretations because of differences amongst the states. One coordinator noted that comparing nearly any two states on NAEP performance can be problematic as the students and curricula across different states may be disparate in ways that render comparisons between those states' NAEP results less meaningful. Another coordinator indicated that because different states often tend to focus on different, specific subgroup populations, these groups can emerge as a common factor in reporting across states and are used in reporting as one yardstick of between-state educational accomplishment. It is important to note, however, that research by Stoneberg (2005) identified further complications to making state comparisons when such relationships are considered without taking into account information about the magnitude of standard errors.

In sum, the data from the coordinators about their experience with stakeholders' understanding of NAEP displays provides a useful source of insight about its usefulness in NAEP's operational reporting efforts at the state level. As further NAEP reporting research is carried out, the perspectives of these coordinators can clearly inform additional study. In addition, the information reported here suggests that there remain many questions for states about what NAEP is and how its results can be put in perspective relative to the *NCLB*-mandated assessments used in the states.

Identifying Sources of Confusion in Current NAEP Displays

Two complementary sets of displays to be used in focus group studies were compiled from the feedback of the coordinators and review of current NAEP publications. One set drew from the Mathematics content area for use with a focus group of individuals with expertise in that academic domain, while the other set reflected results from Reading for reading personnel.

Two focus groups were then convened. The procedures for both focus groups were identical (full study reports are included in Appendices E and F). Briefly, as each display was projected on the screen, participants were asked to reflect on it for a few minutes, and then were asked questions about the data display by one of the two meeting facilitators. Questions ranged from those that were informational in nature ("What was the average score for eighth graders in 2005 in math [reading]?") to opinion ("What, if anything, do you find confusing or not clear about this display?"). The focus group discussion format was appropriate for this study because it served to stimulate broad conversation among the participants and facilitators about the data displays, building on what was being displayed on the screen, and allowed the participants to answer some of the more difficult data interpretation questions collaboratively.

The findings from the two focus groups seem to have important implications for NAEP. This line of research clearly indicated that, with respect to some materials for certain audiences, there may be a need to revise the NAEP score reports to make them more user-friendly. It highlighted the need for more explanatory materials for persons using the NAEP reports. Even educators with quantitative skills experienced some difficulty with many of the common NAEP score reports. This is an important consideration for a testing program such as NAEP, in which the audiences for the data and data products differ widely. Even within an audience ("teachers", "the public"), the range of interests and comprehension levels vary. There seemed to be two types of knowledge that helped these users work with the NAEP data displays included in this study: first, a broad familiarity with test scores and the jargon of assessment (e.g., standard errors, scale scores, etc.), and second, a familiarity with common NAEP terms and reporting mechanisms (e.g., "at or above", the NAEP achievement levels, the interactive online tools).

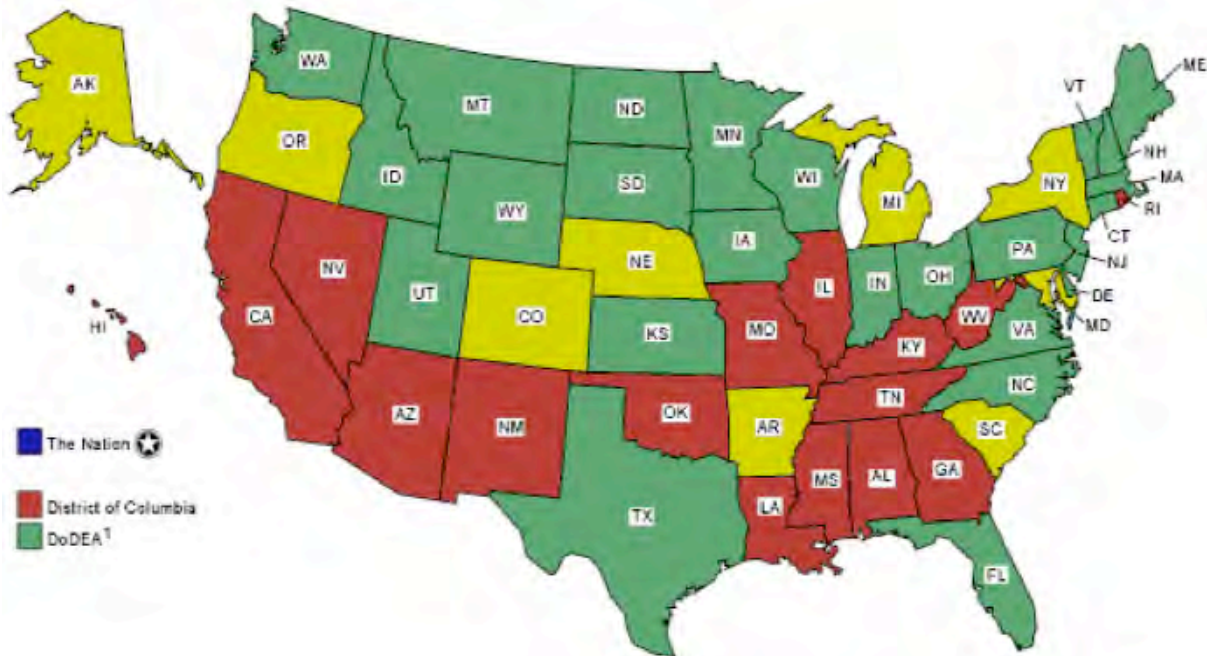
Numerous suggestions for revising the layout of several data displays were made by participants from both groups, particularly with respect to footnotes and arrangement of keys and legends within figures. Participants sought additional information about the practical meaning of some displays, especially when results were labeled as statistically significant. The clickable state maps (Figure 6) were cited as a particularly accessible display for quick interpretations.

Figure 6. NAEP Clickable State Comparison Map (Scale Scores)

Cross-state comparisons of average mathematics scale scores, grade 4 public schools: 2005

NAEP Mathematics Grade 4 - Mathematics
 Difference in Average Scale Score Between Jurisdictions
 for All students [TOTAL] = All students
 2005

Color



- Focal state/jurisdiction (National Public)
- Has a higher average scale score than the focal state/jurisdiction
- Is not significantly different from the focal state/jurisdiction
- Has a lower average scale score than the focal state/jurisdiction
- Sample size is insufficient to permit a reliable estimate

¹ Department of Defense Education Activity schools (domestic and overseas).

NOTE: View complete data with standard errors for [grade 4](#).

Source: Perie, M., Grigg, W., and Dion, G. (2005). *The Nation's Report Card: Mathematics 2005* (NCES 2006-453). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

In summary, the research results from these two focus groups studies of NAEP displays highlight (1) the utility of focus groups for gaining insights about the NAEP score reports, and (2) the importance of either revising the NAEP score reports to make them more user-friendly or the need for more explanatory materials for persons using the NAEP reports. Two focus groups of eight persons each are not a sufficient basis for initiating major report changes, but this research suggests the need for substantially more study. Future directions for research include conducting one-on-one conversations with users of the NAEP Web site about selected data

displays and using suggestions from the focus group to redesign some displays for tryout with focus groups.

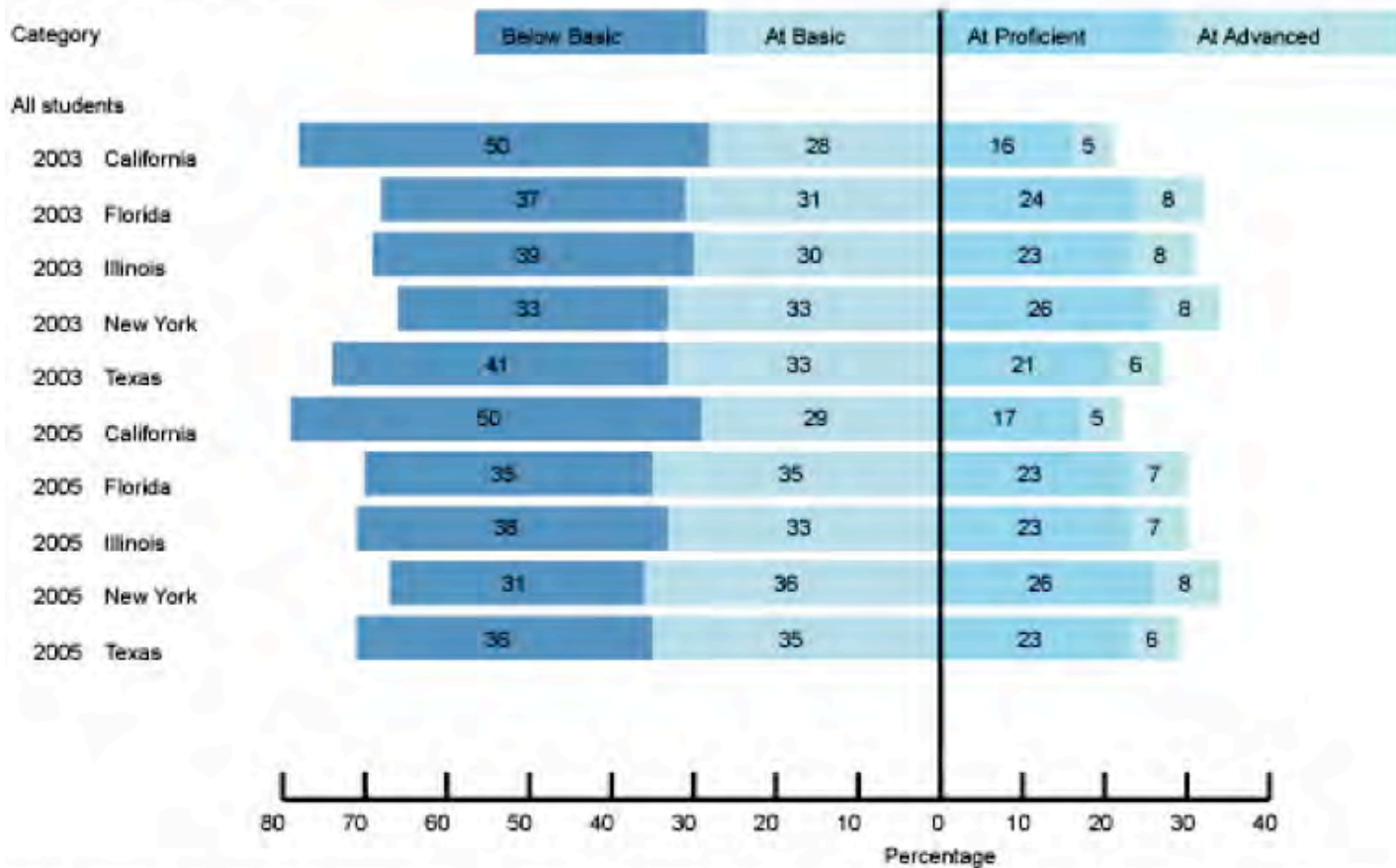
Revising Data Displays for NAEP

Based on the results of these two focus group studies, a third focus group meeting was convened to focus the discussion on specific findings of the previous two groups and to explore in greater depth specific recommendations for improving NAEP data displays. These findings are detailed in Appendix G and are summarized briefly here.

Among the style of displays found to be the most useful were those in Figures 7 and 8 for illustrating achievement levels, and Figure 9 for displaying differences in scores for reporting groups. There was a general agreement that figures such as these that were less data rich, but easy to read and interpret, would be more useful for all audiences.

Figure 7. Achievement Level Percentages for Five States

Percentages of students at each achievement level for reading, grade 4
 All students [TOTAL]
 By jurisdiction, 2003 and 2005



NOTE: Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP).

Source: NAEP Data Explorer (<http://nces.ed.gov/nationsreportcard/nde/>)

Figure 8. Bar Graph of Achievement Levels

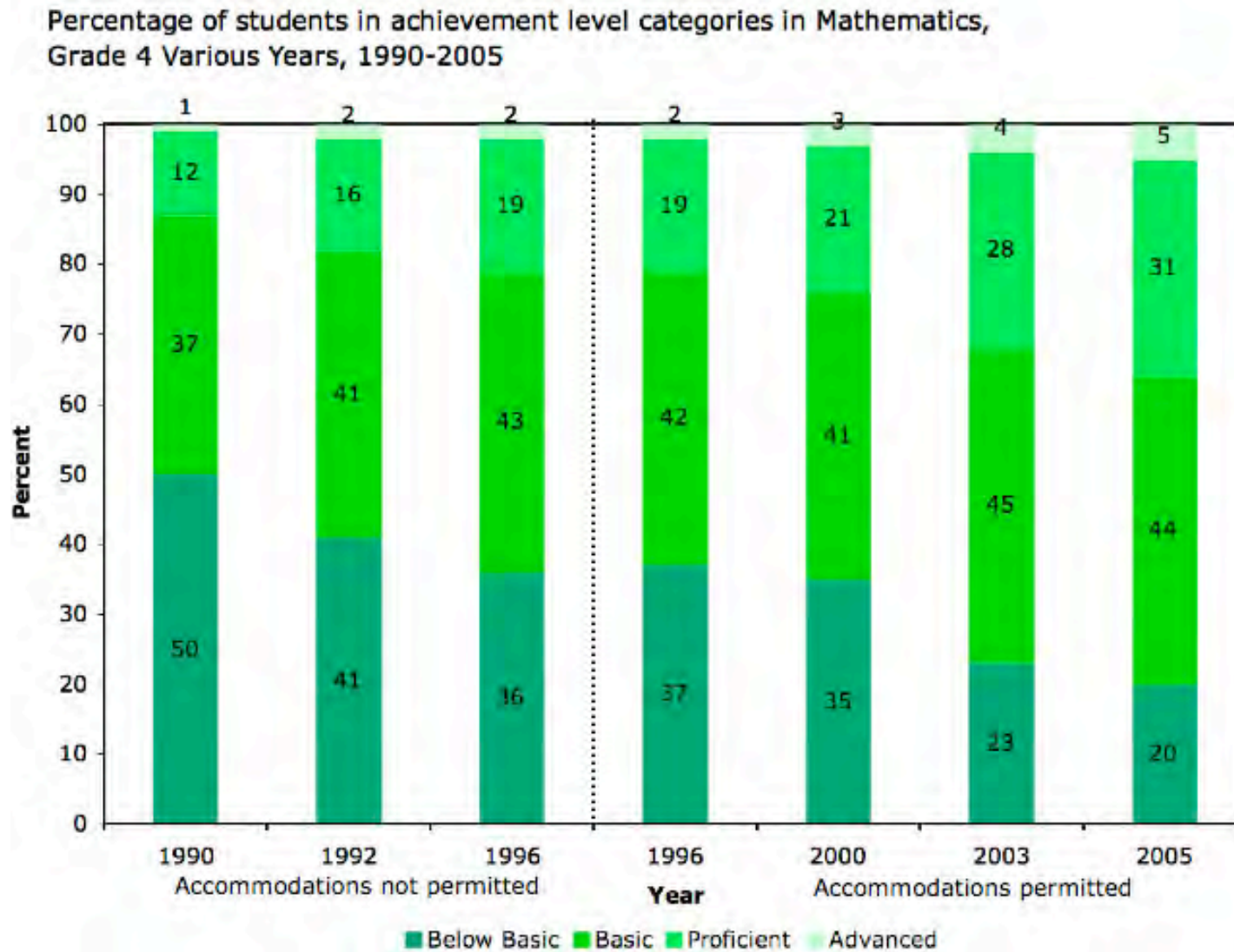
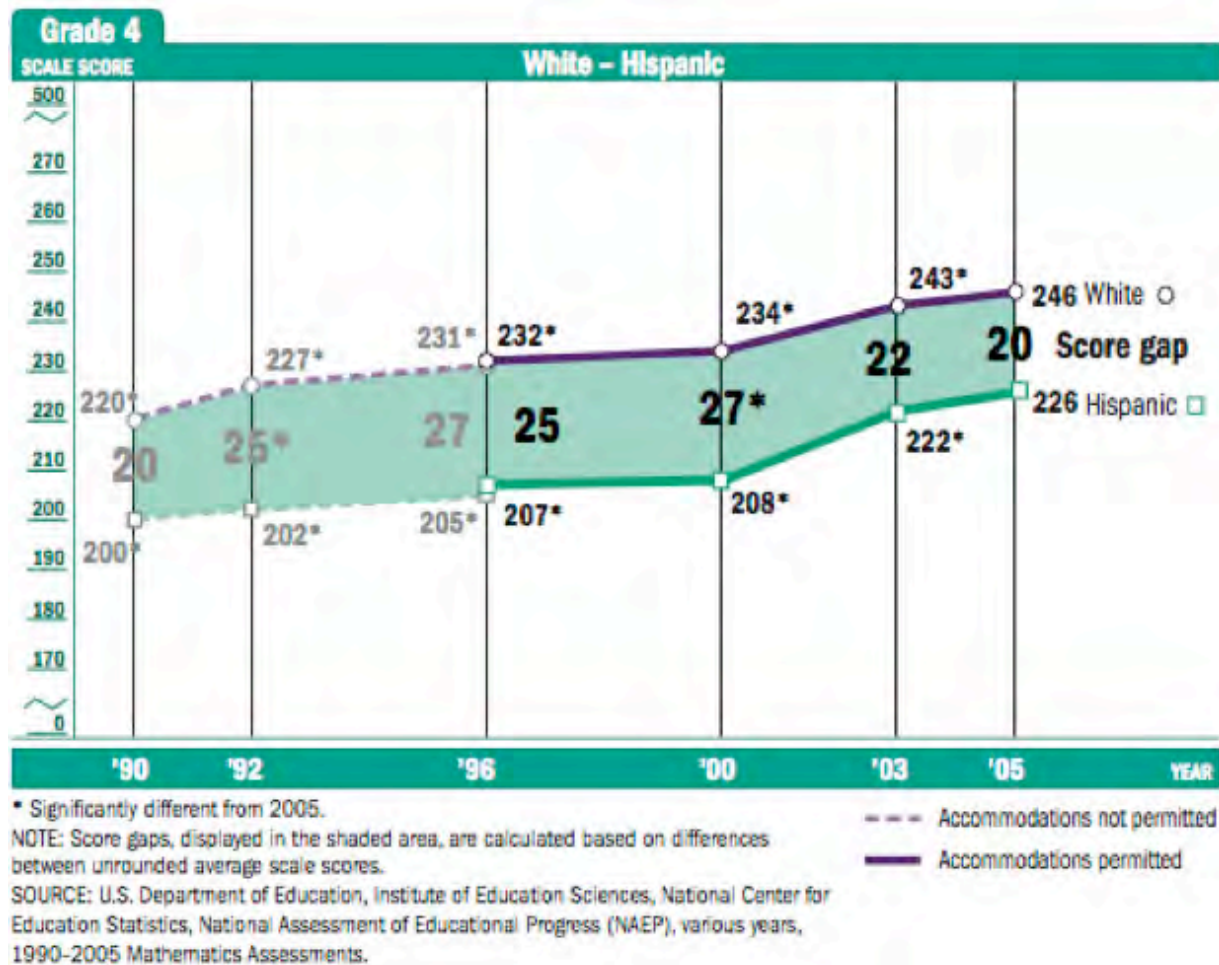


Figure 9. Line Graph Illustrating Performance Gaps Between Two Student Groups



Source: Perie, M., Grigg, W., and Dion, G. (2005). *The Nation's Report Card: Mathematics 2005* (NCES 2006-453). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

For the displays that participants characterized as somewhat less intuitive (such as Figures 10, 11, and 12), their attention was focused on the footnotes and display keys. In Figure 10, the words “statistically significant” only appear in the solid white boxes, and participants noted that they felt that were expected to infer that the boxes with up or down arrows were similarly statistically meaningful. The participants were also quick to point out any inconsistencies in terminology and layout in and across graphs, especially with respect to information presented on the vertical and horizontal axes. One difficulty noted by participants was that the graphs were not consistent when displaying units of time for each jurisdiction represented (such as shown in Figure 11). The axes should be consistent across graphs and level appropriate spacing for years in which NAEP is not administered. However, when the axis is measuring percent and it sums to over 100 percent as in the figures reporting at or above a given level, this should be emphasized and clearly explained on the graph (example in Figure 7).

Figure 10. NAEP Pantyhose Chart

Overall cross-district comparisons of average reading scale scores, grade 4 public schools: 2005

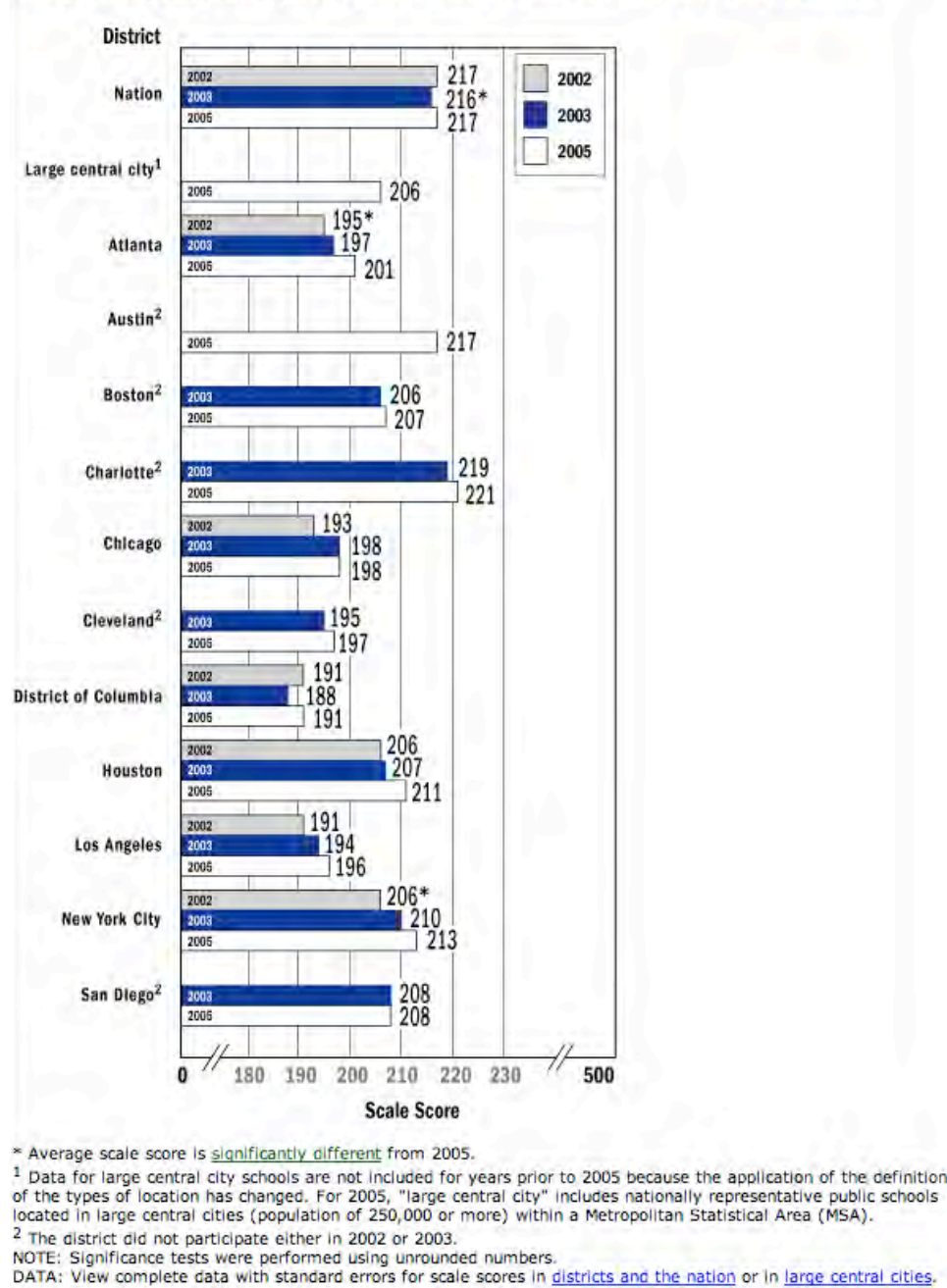


DATA: View complete data with standard errors for scale scores in [districts and the nation](#) or [large central cities](#).

Source: Lutkus, A.D., Rampey, B.D., and Donahue, P. (2006). *The Nation's Report Card: Trial Urban District Assessment Reading 2005* (NCES 2006-455r). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

Figure 11. Clustered Bar Chart

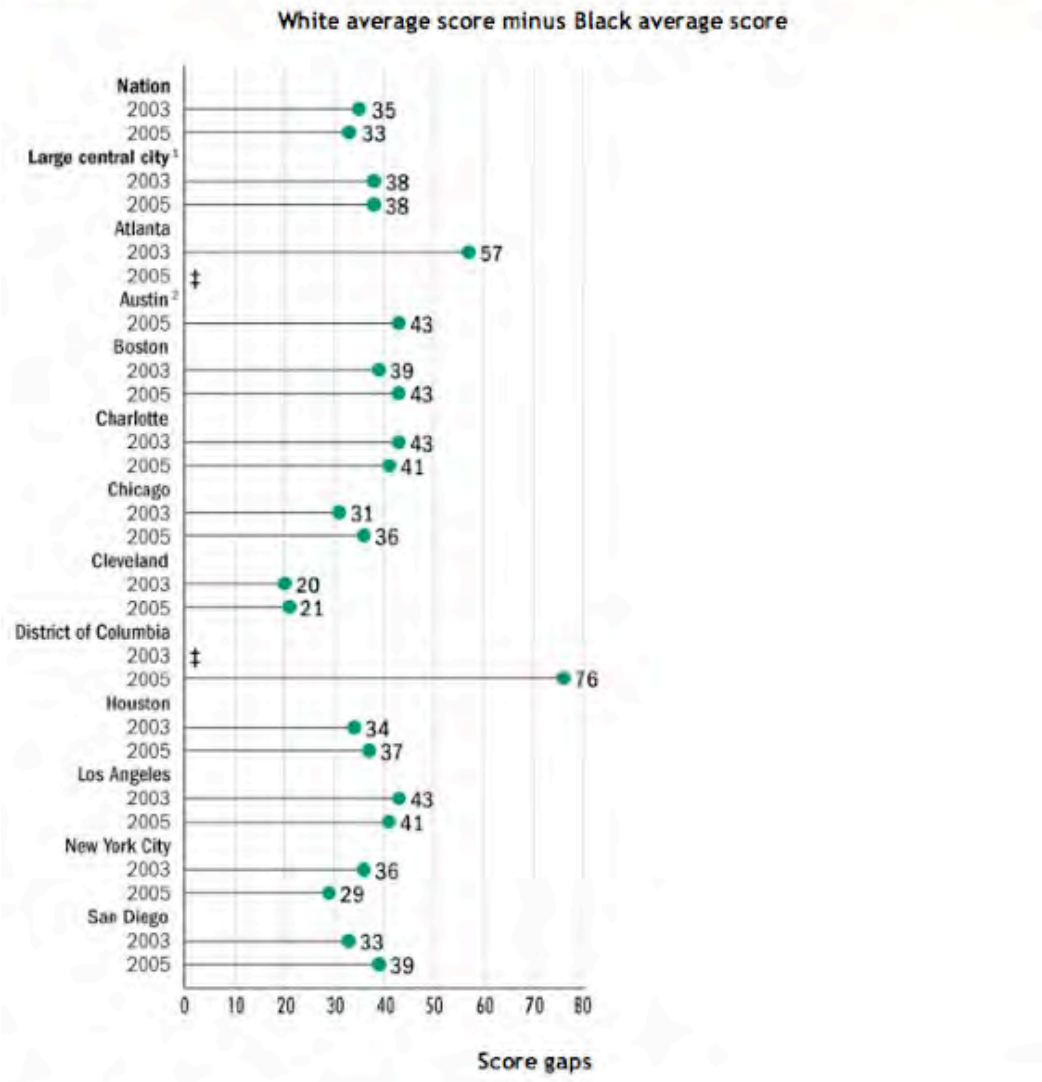
Average reading scale scores, grade 4 public schools: By urban district, various years, 2002-2005



Source: Lutkus, A.D., Rampey, B.D., and Donahue, P. (2006). *The Nation's Report Card: Trial Urban District Assessment Reading 2005* (NCES 2006-455r). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

Figure 12. Score difference graph

White-Black gap in average mathematics scores, grade 8 public schools: By urban district, various years, 2002-2005



† Reporting standards not met.
¹ Some of the TUDA districts include a few public schools located outside of large central cities as defined by the Census Bureau (population of 250,000 or more within metropolitan areas). These schools are included in the category of "large central city" for all years, but were not included in previously published results. As a result, some numbers reported on this website may differ slightly from those reported earlier on the web and in print.
² The district did not participate in 2003.
 NOTE: Score gaps are calculated based on differences between unrounded average scale scores.

Source: Rampey, B.D., Lutkus, A.D., and Dion, G. (2006). *The Nation's Report Card: Trial Urban District Assessment Mathematics 2005* (NCES 2006-457r). U.S. Dept. of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

As noted previously, participants found it challenging to interpret several of the footnotes and legends found on the displays. They questioned the inconsistency of the terminology throughout the various figures (e.g. some figures used the term "district" and others used "jurisdiction"). There were also questions about how NAEP reports statistically significant results. Some of the figures made it explicit that significant results were being reported and others provided no information and participants had to make assumptions.

Overall, the participants expressed interest in seeing more information on subgroup performance. The groups also wanted to see information about the sample sizes and subgroup sample sizes for the data reported (a request which speaks to Standards 13.15 and 13.19 from the AERA, et al. *Standards* (1999)). This group of participants also thought item maps (a sample is shown in Figure 13) were valuable for supplementing the interpretability of displays of scale scores.

The NAEP score scale itself was a source of some confusion. All of the participants reported they were somewhat familiar with NAEP yet none of them seemed familiar with the NAEP scale. When figures displaying scale score data were shown (e.g., Figure 14), numerous questions were raised about the comparability of the NAEP scale to the scales used by state testing programs. Initially the participants found these figures confusing, but when presented with the item map they recognized that they could use it to interpret the meaning of the scale scores and scale score differences between subgroups (which, was a particular area of interest). Participants expressed concern that people may not understand how the item map could be utilized but recommended that it be displayed with graphs that report score scale data to add context to the scores.

Figure 13. NAEP Item Map



¹ Each grade 8 reading question in the 2005 reading assessment was mapped onto the NAEP 0–500 reading scale. The position of a question on the scale represents the average scale score obtained by students who had a 65 percent probability of successfully answering a constructed-response question, or a 74 percent probability of correctly answering a four-option multiple-choice question. Only selected questions are presented. Scale score ranges for reading achievement levels are referenced on the map. For constructed-response questions, the question description represents students' performance at the scoring level being mapped.

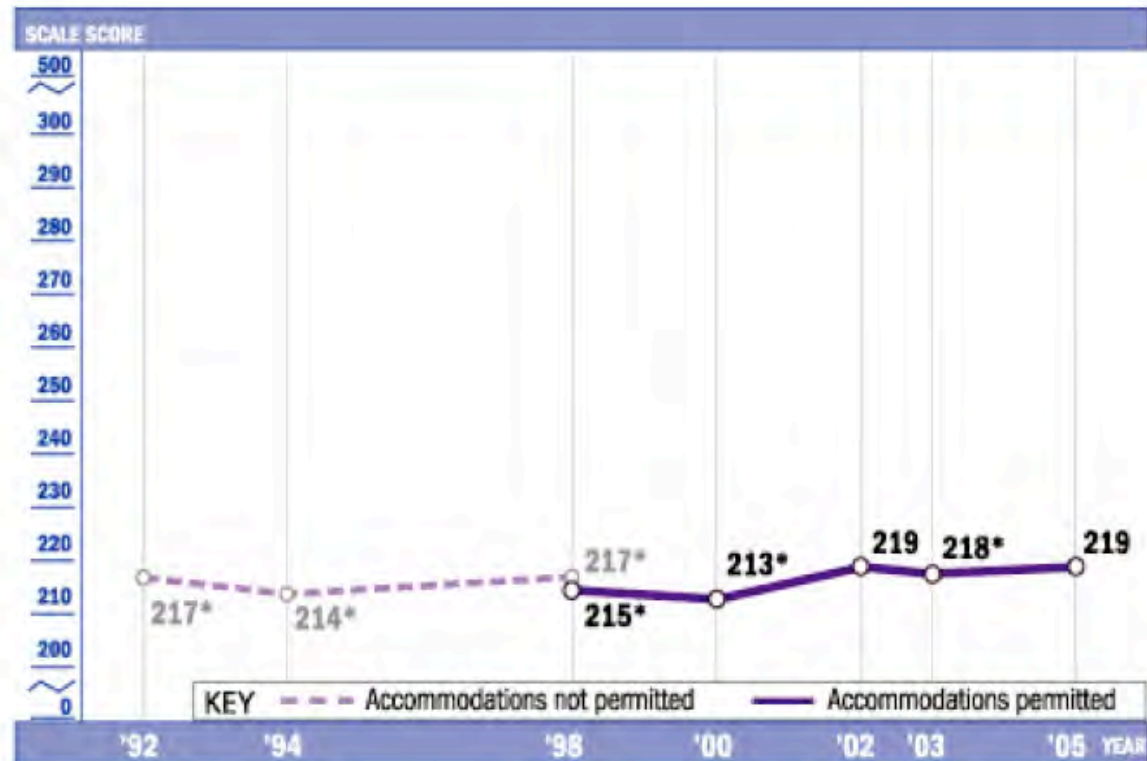
NOTE: Regular type denotes a constructed-response question. *Italic* type denotes a multiple choice question.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment.

Source: Perie, M., Grigg, W., and Donahue, P. (2005). *The Nation's Report Card: Reading 2005* (NCES 2006–451). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

Figure 14. NAEP Line Graph

Average reading scale scores, grade 4: Various years, 1992-2005



* Significantly different from 2005.

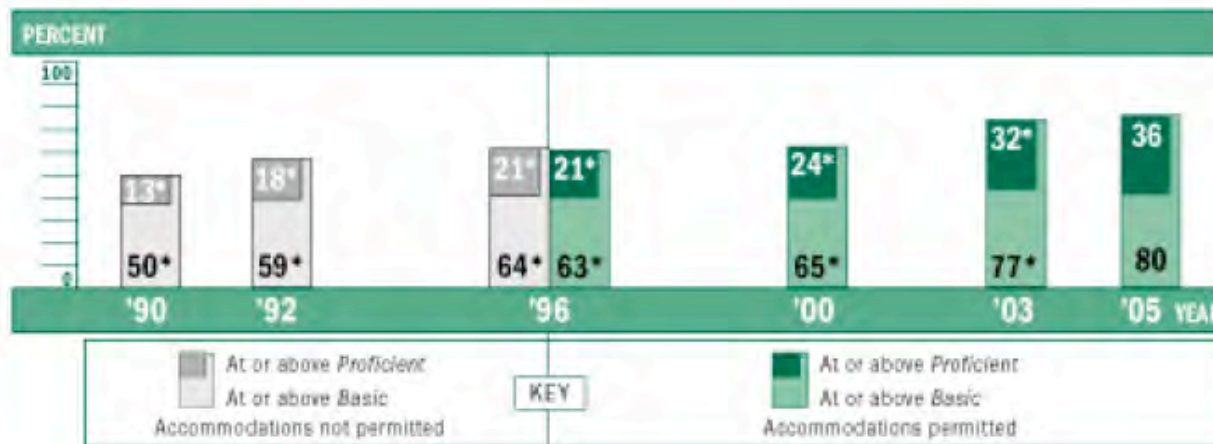
NOTE: The dashed and solid lines represent results based on administrations when accommodations were not permitted and when accommodations were permitted, respectively. View complete data with standard errors for [grade 4](#).

Source: Perie, M., Grigg, W., and Donahue, P. (2005). *The Nation's Report Card: Reading 2005* (NCES 2006-451). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

One area that caused confusion among the group was the terminology and reporting mechanisms used by NAEP. The participants found it difficult to interpret the graphs in which the scores were reported at or above a given achievement level. This finding has also been reported by Hambleton and Slater (1996). Although Figures 15 and 16 were displaying similar information to Figures 7 and 8, all of the participants agreed that Figures 7 and 8 were considerably easier to read and interpret simply because the scores were reported within each level and the percentages summed to 100 percent. Overall the participants were more interested in the percentage of students performing within each performance level and found it more intuitive to interpret information displayed in this manner.

Figure 15. NAEP Stacked Column Chart

Percentage of students at or above *Basic* and at or above *Proficient* in mathematics, grade 4: Various years, 1990-2005



* Significantly different from 2005.

NOTE: The gray shaded boxes represent results based on administrations when accommodations were not permitted. View complete data with standard errors for grade 4.

Source: Perie, M., Grigg, W., and Dion, G. (2005). *The Nation's Report Card: Mathematics 2005* (NCES 2006-453). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

Figure 16. NAEP Achievement Levels Table

Percentage of students, by reading achievement level, grade 8: Various years, 1992–2005				
Assessment year	Below <i>Basic</i>	At or above <i>Basic</i>	At or above <i>Proficient</i>	<i>Advanced</i>
<u>Accommodations</u> not permitted				
1992	31 *	69 *	29	3
1994	30 *	70 *	30	3
1998	26	74	33 *	3
<u>Accommodations</u> permitted				
1998	27	73	32	3
2002	25 *	75 *	33 *	3
2003	26 *	74 *	32 *	3
2005	27	73	31	3

* Significantly different from 2005.

NOTE: Rows will not sum to 100 percent because of cumulative categories. View complete data with standard errors [grade 8](#).

Source: Perie, M., Grigg, W., and Donahue, P. (2005). *The Nation's Report Card: Reading 2005* (NCES 2006–451). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

In sum, the participants in this focus group provided interesting and extensive feedback about the displays as well as broader issues for NAEP reporting. All of the participants offered a tremendous amount of positive feedback about several of the innovative displays and the Question Tool, and expressed a preference for the timeliness and flexibility of online score reporting for accessing results at their convenience. The following are some suggestions for improvements:

- Include sample sizes for each administration, as well as the sample sizes for subgroups where applicable.
- Focus reporting efforts on subgroup differences and aiding interpretation of scale score differences between groups.
- Monitor the consistency of layout and terminology within and across graphs to the extent possible. The data should be displayed in a straightforward manner with explicit and easy to understand footnotes and legends. These should also include links to additional information when it is applicable (e.g. if a graph is reporting scale score, there should be a link to an item map to add meaning to the scores).
- Continue to increase the use of color in data displays; however, in several of the displays it was difficult to distinguish between the colors. Adding more contrast to the colors would clarify several of the graphs.

Summary of Research on NAEP Data Displays

An old adage states that a picture can express a thousand words, and this is the case when the topic is reporting results for a large-scale assessment program such as NAEP. In principle, graphs and figures are an effective way to communicate information about student test performance. Indeed, as noted by Wainer (2000a), well-constructed graphical displays of data put the data front and center to illustrate a clear and complete story. For NAEP, there are many stories of interest. That said, this line of research provides considerable information about the extent to which current NAEP displays are understood by segments of the intended audience, and also offers a blueprint for additional study with other stakeholder groups. Presented below are the specific recommendations for NAEP data displays and continued research in this area.

Reporting Strategies and Content. Between earlier studies of NAEP reports (e.g., Wainer, Hambleton, and Meara, 1999; Hambleton and Slater, 1996) and the current investigations, clear improvement in NAEP data displays have been made, and reporting strategies such as item maps add context in ways that are consistent with the recommendations of the AERA, et al. *Standards* (1999). At the same time, the present studies have identified some broad preferences and recommendations for NAEP data display efforts.

- Visual Displays:
 - *Operational: Use graphs rather than tables whenever possible, as these allow users to make quick visual evaluations of patterns in data.*
- NAEP Scale Scores:
 - *Policy: Identify ways to assign meaning to scale score differences.*
 - *Policy: Promote use of NAEP item maps to help assign meaning to NAEP scale scores.*
 - *Operational: Clearly state the NAEP score scale range on graphs involving scale scores.*
- NAEP Achievement Levels:
 - *Operational: Display achievement level results reported as discrete rather than cumulative.*

- *Operational: Develop and try out strategies to fully explain “at or above” if achievement level results must be reported as cumulative.*

Structural Elements of Graphs. Though participants in the research studies were asked to focus on individual data displays, across all displays discussed much of the feedback received involved the elements of graphs that are intended to aid data users in making sense of the display.

- **Legends:**
 - *Operational: Review legends for consistency in labeling and principles of good layout and design to minimize confusion. For example, in the legend for the pantyhose chart display (Figure 10), participants had to infer that up arrows correspond to significantly better performance and down arrows indicate significantly lower performance.*
- **Axes of Graphs:**
 - *Operational: Clarify axes on graphs. Decisions about score points on axes should be made with ease of understanding in mind, and should reflect common understanding of percents and other information. For example, an axis in a figure generated by the NAEP Data Explorer displayed a horizontal axis labeled “Percentage” with 140 possible points on it and increased the difficulty some users had with making correct interpretations.*
- **Denoting Statistical Significance:**
 - *Operational: Revise footnotes to explain/denote statistical significance in an explicit and meaningful way.*

Display-specific Comments. Across all of the focus groups, participants were exposed to a wide range of current NAEP data displays, including line graphs, bar charts, state maps, and tables.

- **Clickable State Maps:**
 - *Operational: Make the color scheme more distinct.*
 - *Operational: Enhance the size of the legend.*
- **Pantyhose Chart:**
 - *Operational: Devise strategies to displaying information in the pantyhose chart to enhance the practical meaning that could be ascribed to this display (as no scale scores are shown, just significance test results).*
- **NAEP Question Tool:**
 - *Policy: Promote awareness of the Question Tool among stakeholders to enhance use.*
 - *Operational: Consider slight revision of the layout of Question Tool distracter analysis tab to rearrange information about standard errors less prominent (but still available).*
- **Horizontal/vertical bar charts:**
 - *Operational: When information for two or more groups (states, subgroups, etc.) and multiple years is included in a horizontal bar or vertical bar chart, even if data are not available for a group in each assessment year, denote that data are not available and maintain consistent spacing in chart development to facilitate reading of the graph.*
- **Reporting Gaps:**
 - *Operational: Use the line graphs for two-group comparisons, but label the lines clearly.*
 - *Operational: Avoid the difference score bar charts (as shown in Figure 12).*

Focus Groups. The focus group methodology allows for useful data about data displays to be gathered. NCES, NAGB, and contractors should carry out focus groups and similar research activities in order to identify potentially problematic aspects of displays prior to use in paper or Web-based reports.

Reporting Interests of NAEP Audiences

Overview of Research

One of the critical tasks in score reporting is to identify the intended audience for reporting efforts and, in the case of NAEP, the Policy Guidelines (National Assessment Governing Board, 2006a) explicitly define the primary audience for NAEP as the American public. The guidelines also specify that materials to disseminate NAEP should be developed for the interested general public, policymakers, teachers, administrators, and parents, and that NAEP results should be distributed to governors and chief state school officers, as well as to superintendents of TUDA districts. National and state organizations with interest in education should also be notified of NAEP results, and personnel from NCES and NAGB are encouraged to communicate information about NAEP with various national, state, and local organizations and media representatives. From the menu on the home page of the NAEP Web site, parents, researchers, media, educators and policymakers are five groups with dedicated information sections.

In this section of the Utility study, we further explore the reporting interests of NAEP audiences with respect to a) education officials and educators and b) other audiences of interest, as identified in previous studies (including Jaeger, 2003; Levine, et al., 1998; Simmons and Mwalimu, 2000). To do so, conversations with a subset of NAEP state coordinators provided a discussion of their reporting experiences and the reporting interests of the stakeholder groups that they interact with. Next, in the course of focus groups meetings on NAEP data displays (results reported in Appendixes), participants (educators and education officials) expressed their preferences about reporting methods and interests, and these are briefly summarized here.

Coordinators' Reporting Experiences in the States

Several NAEP coordinators were asked to reflect on the results of NAEP that they perceived of greatest interest to NAEP audiences, based on their experience. Year-to-year scale score trends within states were commonly mentioned by a number of coordinators, as were comparisons to national averages and rankings of states. Subgroup results were likewise identified as a considerable source of interest in different states, but the specific subgroup comparisons depend on the state. For example, where there are high Native American populations or many limited English proficient students, studies of score gaps with those groups take on critical significance.

One reporting angle that was cited as something that was not always a priority in the states' reporting efforts was the NAEP achievement levels (though these are widely focused on in the national NAEP reporting efforts, per NAGB's Policy Statement (2006a) and Guidelines (2006b)). Nearly all coordinators noted that in many cases neither the state tests and NAEP nor the states' *NCLB*-reported performance categories and the NAEP achievement levels can be connected in ways that are amenable to easy communication to users who are unfamiliar with the fine points of alignment. There seemed to be many questions about NAEP-state alignment, and this was brought up in some cases as an impediment to the relevance of NAEP in some states.

In addition, a broader lack of assessment literacy and statistical knowledge among audiences was named as one further challenge to communicating NAEP results in the states. Because of the complexity of NAEP in terms of administration procedures and in terms of the policy of reporting results as scale scores and achievement levels (which are different from those used in the individual states), NAEP results are not as familiar to many users. Two coordinators further raised the issue of the perceived relevance of NAEP among stakeholders,

noting that without student- or district-level results (except for the municipalities involved the Trial Urban District Assessment) the immediate meaningfulness of NAEP was not readily apparent to some potential audiences and therefore was an additional obstacle to effective reporting.

Coordinators' Experiences with NAEP Sub-Audiences

Of considerable interest in this study of the broader experience of reporting NAEP were answers to questions about interactions with potential consumers of NAEP data, including education officials (positions such as state education commissioners, state assessment directors, and curriculum leaders), educators, the media, policymakers (federal and state legislators and their aides), and the public. Coordinators were asked how often they spoke with members of those different groups, whether those interactions were generated by the coordinator or members of those groups, the kinds of information that the different groups expressed interest in, and how that information was communicated (via the Web, in-person meetings, etc.).

Politicians and Political Aides. Different states used NAEP results differently, and these variations were reflected in the range of responses obtained to questions about coordinator interaction with these individuals. A few mentioned having prepared materials for state legislators or their aides on occasion, and in several cases this contact was through media relations. One coordinator said that governors' briefings (done by NAGB) were conducted in conjunction with releases, and the state school board or board of regents would occasionally request information. In addition, it was indicated by several coordinators that the results of interest for politicians of this sort are often very straightforward, focusing on trends, state-national comparisons, and gaps. No coordinator reported having been asked in-depth or exploratory questions by these individuals: the reporting efforts with this group typically involve reformatting results or existing reports so they are informed in a simple, direct way, but not directing these individual to the NAEP Web site.

The Media. In a focus group of education writers convened by Levine, et al. (1998), findings of interest included some level of interest in reporting using the extensive NAEP background variables, a preference for executive summary-level technical documentation, and desiring a set schedule with considerable advance notice for NAEP data releases. They stressed the need for information about international comparisons (in which the data included both public and private schools), and specific information of interest to them given limited space in print including state NAEP results, comparing scores from recent years, and reporting information without numbers and graphs (though better graphics would also be helpful).

While a few coordinators reported that they communicated directly with members of the press, in other states, education departments preferred the NAEP coordinators to work with a media relations or public affairs office and let the personnel there handle press inquiries about NAEP. In the latter case, the coordinators prepare press releases and simplified, quick-reference "fact sheets" that could be distributed or referenced as needed. The timing of the most media interest in NAEP in the states was when there was a national release of results ("that's when the pressure is highest"), and the questions the media asked were often "following the lead of the state with respect to the press release." As that was true for many of the coordinators we spoke with, many coordinators reiterated the importance of the prerelease workshop for allowing coordinators and their states to access the data before release in order to prepare and double-check the message for the media. One challenge of working with the press, noted a coordinator, was that "sometimes they're off doing they're own research, and not always getting it right," explaining that the press in that state did not always understand the data, and then when they tried to break the information down further their confusion was transferred to their viewers or

readers. The coordinator had to try and clarify the message after the fact. This can be hard, according to another coordinator, who pointed out that the window of opportunity for communicating NAEP results after a release was very brief, and the job was to “make the message your own.” One particular source of problems in NAEP media coverage was communicating the concept of “statistical significance.”

The General Public. Little or no direct interaction with the public on anything NAEP-related was reported by most coordinators, with one mentioning very occasional questions from parents about why a student was participating in NAEP (but no questions about reporting). Some information about the reporting preferences of the public at-large was reported in Levine, et al. (1998), when the representatives of this NAEP audience indicated that fewer, larger releases of results would likely ensure that more people paid attention to the assessment. In terms of the contents of reports, this focus group identified the main question for NAEP to answer for them as “Are we improving as a nation educationally?” State comparisons and rankings were identified as important. Knowing what the higher-performing states were doing “right” to inform other jurisdictions was also identified.

Educators. Many coordinators reported limited educator-initiated communication about NAEP results. Of the coordinators who reported being in touch with educators, most of that contact involved workshops for educators at state teacher conventions or meetings held annually or semiannually or cases in which coordinators made concerted outreach efforts to connect with school administrators and teachers about NAEP in general. When coordinators were not talking with educators about administration procedures—a topic that coordinators said comprised the bulk of those conversations—the coordinators were working to introduce the NAEP Question Tool to educators and to show them ways of integrating that resource into their classrooms, thereby enhancing the relevance and use of NAEP. To the extent that educators raise questions about results with coordinators, the kinds of information they focused on were national-state comparisons, gaps and subgroup analysis, and the relevance of results for them (and many coordinators identified this as a challenge throughout the conversations because of the disconnect that potentially exists between NAEP and state assessments).

Focus Group Findings

At the conclusion of the three focus group studies, all of which primarily focused on NAEP data displays, participants in the different groups were asked several different questions about their interests in NAEP and reporting preferences as an audience for NAEP results (see Appendices E, F, and G). In discussing reporting methods in general, most found that executive summaries are generally useful for them, and the NAEP state snapshot reports for a content area at a given grade level was mentioned as being particularly useful for giving results clearly with just enough context. Speaking to the attendees’ knowledge of schools’ and teachers’ use of the results, executive summaries were identified as unlikely to be read, and participants stated that “schools want to know how their kids are doing compared to other schools and districts” but that “not all schools care as much about NAEP results because their kids don’t take it” and “districts want individual results but that’s not NAEP.” Another attendee described the state’s use for NAEP as being important at the state level but not at the district level, to “justify our state assessment is on target.”

The general consensus expressed by these individuals was that scale scores in isolation were not as informative as performance levels (in NAEP, below *Basic*, *Basic*, *Proficient*, and *Advanced*), because most focus group participants were not familiar with the NAEP scale or the specifics of how it differed from the various scales in use in their states for *NCLB* or other assessments in use. When presented in NAEP reporting efforts, scale scores should be placed in

context (including the range of possible scores and examples of skills or knowledge exhibited by individuals at that score level). The kinds of results that states are interested in often varies, with some choosing to emphasize state comparisons based on geographical proximity, while others were more concerned with demographic-group performance or state comparisons chosen to emphasize peers with demographically similar student populations.

Reporting Interests from Site Usage Statistics

As reported in Appendix C, the NAEP Web site's usage statistics also provide a level of insight into the frequency with which certain information about NAEP is accessed by visitors to the NAEP site. First, the State Profiles exhibited a consistently high volume of use. This suggests that among the results of considerable interest to the NAEP site are state-specific results and information, and as such appears to be a focus for future reporting efforts and an area for future research. Also, use of the Question Tool is growing. Certainly, anecdotal evidence from some of the coordinators indicated that these individuals were promoting use of the Question Tool among educators they were in contact with. A high level of interest in the displays of results for student subgroups was also noted in the Web site usage statistics study, and this likewise suggests that visitors to the NAEP site are interested in demographic results. Lastly, use of the interactive NAEP Data Explorer tool seems to be growing.

Summary of Research on NAEP Reporting Interests

The data collected from the current research activities, when taken in conjunction with the research by Jaeger (2003), Levine, et al. (1998), and Simmons and Mwalimu (2000), offers a compelling argument for talking with stakeholders about NAEP results that they are interested in and that are useful to them. A clear issue that emerged in much of the research on reporting preferences here was the NAEP achievement levels, and this also raises a larger point about the overall relevance of NAEP. In reporting national results and trends in the nation's performance (and that of demographic subgroups over time), the achievement levels can likely offer a clear yardstick for understanding the performance of American students, because nothing else fills that role for the nation at large at least at the present time. However, when talking with the NAEP state coordinators and personnel interested in NAEP at the state level, it seems clear that in many quarters a disconnect is present with regard to many stakeholders' understanding of the meaningfulness of NAEP's performance categories in individual states.

For many of the participants in this research, NAEP's achievement levels do not attain the familiarity or level of comprehensibility that the state performance categories do for many users of assessment data in the states, and this seems in part to be related to the infrequency of NAEP data releases. Thus, one recommendation for moving forward in this dimension of reporting is to continue to work with the NAEP state coordinators to draw on their experiences in the states and to develop a comprehensive strategy for communicating NAEP achievement level results in the states. Related to this is further work on NAEP-state alignment to understand the relationship between NAEP and the states' respective frameworks and assessments.

A second recommendation is to continue work on ways to add context to NAEP results. The item maps appear to be an important addition to the NAEP reports in recent years, but they are largely unfamiliar to many users of NAEP and often, additional explanatory text is needed to help users understand what is being communicated. The work of DiBello and Stout (2003) through the NAEP Secondary Analysis Grants program is one example of a promising direction for research, in which the focus is on detailed criterion-referenced interpretations of achievement levels performance aggregated across students. As noted by DiBello and Stout, profile scores

may “provide a foundation ... for evaluating similarities and differences between NAEP and the state accountability tests” (p. 4).

Third, the coordinators are at the forefront of reporting NAEP results in the states, and are in touch with which results are of importance in their respective states. The diversity of the states and the specific demographic conditions that exist in each of them lead to the need for continued evaluation of the reporting tools available for making subgroup and gap comparisons. The methods for reporting gaps were among some of the most misunderstood data displays in the earlier portion of the Utility study, but for many political, social, and educational reasons are among the results of greatest interest to many stakeholders.

This page left intentionally blank

Summary and Conclusions

Communicating test results is a critical part of what testing agencies must do, and the test referred to by the moniker of “The Nation’s Report Card” has drawn an especially hefty task in this respect. NAEP is a barometer of what American students know and can do. The NAEP assessment program is nationally representative of states and the nation, involves content areas ranging from the requisite (mathematics, reading and writing) to the revealing (civics, economics, and world history), and measures the academic proficiency of students in elementary, middle, and high school. Reporting NAEP results serves a wide variety of audiences with a considerable range of (a) interest in the results and (b) assessment literacy for interpreting and using the findings. Reflecting the technical excellence that characterizes other aspects of NAEP, it is clear that considerable investment of time and resources have gone into NAEP’s reporting practices. NAEP’s paper reports and executive summaries in many respects have become models for communicating test results to stakeholder groups, and the data tools available online via the NAEP Web site are significant innovations that few other testing programs can claim.

At the same time, the AERA, et al. (1999) *Standards* define a number of vital considerations for reporting efforts, and the results of this study identified a number of specific directions for improvement with respect to the reporting of NAEP results. Through the various interviews and focus groups, we arrived at 77 specific recommendations, which were described in previous sections of this report. An accounting of the recommendations stratified by Research Question Area and Topic is presented in Table 2.

Table 2. Recommendations by Research Question Area and Topic

Research Question Area	Topic	Number of Specific Recommendations
NAEP on the Web	NAEP Data Explorer: Information	11
	NAEP Data Explorer: Functionality	9
	NAEP Home page	7
	State and TUDA Profiles	7
	NAEP Data Explorer: Appearance	5
	High Traffic Links	5
NAEP Data Displays	Display-specific Comments	8
	Reporting Strategies and Content	6
	Structural Elements of Graphs	5
	Focus Groups	1
NAEP Audiences	Adding Context to Results	4
	State-Level Reporting	1
	Communicating Gaps in Performance	1
Overall Recommendations	Future Research	4
	Stakeholder Interests	2
	Incorporating Feedback: NAEP Web site	1
Total:		77 Recommendations

With respect to the NAEP Web site, the numerous operational recommendations identified through this research call for continued investigation into the usability and understanding of the most commonly accessed pages and careful consideration of the layout of the current NAEP home page. Ongoing empirical study of both the NDE with specific audience groups and the Initial Release Site is also recommended. The highest priority should be given to

recommendations associated with the NAEP home page, as this is the entry point for most users to NAEP's presence on the Web.

As to the data displays, reporting efforts should emphasize visual clarity, attention to detail in reporting, and identification of ways to add context to graphs so that they are more readily interpretable, and consistent with the tenets of the AERA, et al. (1999) *Standards*. Going forward, the suggestions contained within reporting strategies and content should be strongly considered for their broad potential to impact the use and understanding of NAEP reports.

Among stakeholder interests and needs, the primary theme identified involved boosting current efforts to add context to NAEP scores and improve audiences' understanding of the achievement levels, particularly to minimize confusion between NAEP-state reporting and enhance the relevance of NAEP results in the states. We reiterate the importance attached to the recommendations concerning ways to add context to NAEP reports.

In addition, several policy findings emerged that have broader implications for NAEP reporting going forward. These are provided below.

- Recommendation 1: *Carry out systematic studies of planned and current or ongoing reporting strategies (data displays, Web pages or tools) with stakeholder groups prior to the use of these reporting strategies operationally.*
- Recommendation 2: *Develop formal procedures for incorporating research findings into operational reporting efforts.*

Currently, NAEP score reporting research seems to be carried out independently under the auspices of different partners in the NAEP Alliance (including Educational Testing Service and GMRI, Inc., among others) and NAGB and its subcontractors. Particularly as new initiatives in reporting are rolled out, focus groups, one-on-one observations, "think-aloud studies," interviews, and other research techniques should be employed to identify potential limiting content, understanding, and functionality elements of NAEP reports and data tools. In addition, as research occurs on new and existing reporting methods, these findings should be used to inform and improve practices.

- Recommendation 3: *Revisit and revise aspects of the NAEP Web site to reflect empirical findings about ease of use, audience interests, and current accepted Web development practices.*

The steady rise in the availability of NAEP information on the Internet is an important, positive step for score reporting efforts, both in communicating NAEP to constituents and as a model for other testing programs. At the same time, as with all aspects of NAEP's reporting, the Web site requires a concerted ongoing program of research and maintenance. The findings of the research activities detailed here provide clear direction for improvement, and with further evaluation relative to industry standards, the NAEP Web site can continue to be a leader in online score reporting.

- Recommendation 4: *Consider ways to incorporate stakeholder interest by using audience-specific materials.*

As part of the broader picture of communicating NAEP results, the current research as well as the work of Jaeger (2003), Levine, et al. (1998), and Simmons and Mwalimu (2000) indicates that different NAEP audiences have different data needs and reporting interests. For example, some states have identified within-state subgroup differences on NAEP as a reporting priority, while others are more concerned with how their state compares to states that are demographically similar. Because NAEP is not reported for participating districts, schools, or individuals, the value of reporting at the state level is the one of greatest relevance for many constituents. Connecting NAEP to the state *NCLB* assessments is one recommended direction for this work. Particular effort should be devoted to incorporating stakeholder feedback into NAEP's reporting

documents, and following up to identify other potential avenues for information presentation or use.

In sum, the research described here has offered a number of substantive recommendations regarding NAEP score reporting, with respect to both a) small but easily-implemented changes and b) larger program concerns. While NAEP remains at the forefront of assessment practices (including score reporting efforts), much input from many members of different NAEP stakeholder groups is reflected throughout these findings about how NAEP is and perhaps should be communicated, and their participation in this and other research on NAEP score reporting is appreciated and must be encouraged. As the broader NAEP testing program continues to provide critically valuable data to the American public about the academic performance of the nation's schoolchildren, the findings described here support ongoing reflection on NAEP's practices as well as evidence for the high quality of NAEP's current reporting and dissemination strategies.

This page intentionally left blank

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Beaton, A.E., and Johnson, E.G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement*, 26 (2), 163–175.
- Cleveland, W.S. (1994). *The elements of graphing data* (2nd edition). Summit, N.J.: Hobart Press.
- De Mello, V. B. (2004). *NAEP state analysis project. Task 2.2. State Profile and Report Enhancement: Recommendations on State Web Profiles* (Contract ED-01-CO-0026/0019). Washington, D.C.: American Institutes for Research.
- DiBello, L. V., and Stout, W. (2003). *Skill profiles for groups of students at a given NAEP scale level—Development and demonstration*. Proposal for the NAEP Secondary Analysis Grant Program.
- Goodman, D. P., and Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17 (2), 145–220.
- Government Micro Resources, Inc. (2004). *NAEP Data Analyst usability study*. Manassas, Va.: Author.
- Government Micro Resources, Inc. (2005a). *Usability test results for the Initial Release Site*. Manassas, Va.: Author.
- Government Micro Resources, Inc. (2005b). *NAEP ambassadors and the Initial Release Site usability study*. Manassas, Va.: Author.
- Hambleton, R. K. (1998). Enhancing the validity of NAEP achievement level score reporting. In M. L. Bourque (ed.), *Proceedings of the Achievement Levels Workshop* (pp. 77–98). Washington, D.C.: National Assessment Governing Board.
- Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scale and reports more understandable? In R. W. Lissitz and W. D. Schafer (eds.), *Assessment in educational reform* (pp. 192–205). Boston, Mass.: Allyn and Bacon.
- Hambleton, R. K., and Meara, K. (2000). Newspaper coverage of NAEP results, 1990 to 1998. In M. L. Bourque and S. Byrd (eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 133–155). Washington, D.C.: National Assessment Governing Board.
- Hambleton, R. K., and Slater, S. C. (1994). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved?

Paper presented at the Joint Conference on Standard-Setting for Large-Scale Assessments, Washington, D.C.

Hambleton, R. K., and Slater, S. C. (1995). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved? In *Proceedings of the Joint Conference on Standard-Setting for Large-Scale Assessments*, pp. 325–343.

Hambleton, R. K., and Slater, S. C. (April, 1996). Are NAEP executive summary reports understandable to policy makers and educators? Paper presented at the annual meeting of the National Council on Measurement in Education, New York, N.Y.

Hambleton, R. K., and Zenisky, A. L. (in preparation). Reporting on score reports: The grade is D but improvement can be seen. *CLEAR Exam Review*. [Also Center for Educational Assessment Report No. 622. Amherst, Mass.: University of Massachusetts, School of Education.]

Jaeger, R. M. (1992). General issues in reporting of the NAEP trial state assessment results. In R. Glaser and R. Linn (eds.), *Assessing student achievement in the states* (pp. 107–109). Stanford, Calif.: National Academy of Education.

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working Paper 2003-11). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.

Jirka, S. J. (2007). *A review of the literature on score reports and data displays*. Unpublished manuscript.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29 (2), 95–110.

Koretz, D., and Deibert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991* (MR-385-NCES). Santa Monica, Calif.: RAND.

Levine, R., Rathbun, A., Selden, R., and Davis, A. (1998). *NAEP's constituents: What do they want? Report of the National Assessment of Educational Progress Constituents Survey and Focus Groups* (NCES 98-521). Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, 11, 23–47.

Linn, R. L., and Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29, 177–194.

Lutkus, A.D., Rampey, B.D., and Donahue, P. (2006). *The nation's report card: Trial urban district assessment reading 2005* (NCES 2006-455r). U.S. Department of Education,

- National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.
- Mayer, R. E., and Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38 (1), 43–52.
- Mislevy, R. J. (December, 1996). Implications of market-basket-reporting for achievement level setting. Paper presented at a workshop on Setting Consensus Goals for Academic Achievement, under the sponsorship of the National Research Council Committee on the Evaluation of NAEP, Washington, D.C.
- Mislevy, R.J., Johnson, E.G. and Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17 (2), 131–154.
- National Assessment Governing Board. (2006a). *Policy statement on reporting, release, and dissemination of NAEP results*. Retrieved Jan. 27, 2007, from <http://www.nagb.org/release/policy06.doc>.
- National Assessment Governing Board. (2006b). *Guidelines for The Nation's Report Card*. Retrieved Jan. 27, 2007, from <http://www.nagb.org/release/guidelines06.doc>.
- National Research Council (NRC). (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting*. Washington, D.C.: National Academy Press.
- Ogilvy Public Relations Worldwide. (2004). *NAEP: Reporting initial results. Analysis and recommendations for improvement*. Washington, D.C.: Author.
- Pass, F., Renki, A., and Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38 (1), 1–4.
- Perie, M., Grigg, W., and Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES 2006-453). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.
- Perie, M., Grigg, W., and Donahue, P. (2005). *The nation's report card: Reading 2005* (NCES 2006-451). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.
- Perie, M., Moran, R., and Lutkus, A.D. (2005). *NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics* (NCES 2005-464). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.
- Rampey, B.D., Lutkus, A.D., and Dion, G. (2006). *The nation's report card: Trial urban district assessment mathematics 2005* (NCES 2006-457r). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U.S. Department of Education.

- Simmons, C., and Mwalimu, M. (2000). What NAEP's publics have to say. In M. L. Bourque and S. Byrd (eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 184–219). Washington, D.C.: National Assessment Governing Board.
- Stoneberg, B. D. (2005). Please don't use NAEP scores to rank order the 50 states. *Practical Assessment, Research, and Evaluation*, 10 (9). Retrieved Jan. 27, 2007, from <http://pareonline.net/pdf/v10n9.pdf>.
- Tufte, E. R. (1990). *Envisioning information*. Cheshire, Conn.: Graphics Press.
- Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, Conn.: Graphics Press.
- Tufte, E. R. (2001). *The visual display of quantitative information*. Cheshire, Conn.: Graphics Press.
- Tufte, E. R. (2006). *Beautiful evidence*. Cheshire, Conn.: Graphics Press.
- U.S. Department of Health and Human Services. (2006). *Research-based web design and usability guidelines*. Washington, D.C.: Author.
- Wainer, H. (1996). Using trilinear plots for NAEP data. *Journal of Educational Measurement*, 33, 41–55.
- Wainer, H. (1997). Improving tabular displays: with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, 22, 1–30.
- Wainer, H. (2000a). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot* (2nd ed.). Hillsdale, N. J.: Lawrence Erlbaum Associates.
- Wainer, H. (2000b). Cholera, rocket ships, and Tom's veggies: Contemporary and historical ideas toward the effective communication of school performance. *Evaluation and Research in Education*, 14, 148–180.
- Wainer, H., Hambleton, R.K., and Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335.

Appendixes

This page intentionally left blank

**Appendix A: Navigating ‘The Nation’s Report Card’ on the World Wide Web:
Site user activities and impressions**

April L. Zenisky and Ronald K. Hambleton

*Center for Educational Assessment
University of Massachusetts Amherst*

Abstract

The Internet is increasingly a source that individuals use as a primary source of information. The purpose of this study was to elicit opinions about the NAEP Web site (<http://nces.ed.gov/nationsreportcard>) from a range of individuals involved in educational assessment. Participants were asked to provide feedback about multiple aspects of the NAEP site, including the content and navigation of the home page (and links from that page), the State Profiles pages, and the interactive tools such as the NAEP Data Explorer and the NAEP Question Tool. This study took place June 26 and 27, 2006, in San Francisco, Calif., during the Council of Chief State School Officers' Conference on Large-Scale Assessment. About a month prior to the conference, an electronic version of the attendee list was obtained from CCSSO, and from that list conference attendees associated with national policy organizations (45 attendees), state or regional policy organizations (16 attendees), the media (9 attendees), local school districts (65 attendees), and state departments of education (289 attendees) were identified, and about 50 of those individuals working with assessment data were invited by e-mail to participate. Individuals were asked to participate in a one-hour meeting during the CCSSO conference to explore the NAEP Web site and provide feedback on the user experience. Participants were offered an honorarium of \$100 for their participation. A total of 16 participants were recruited for the study, and 14 participants were present at their scheduled sessions.

During the meeting, participants met with a researcher from the University of Massachusetts Amherst in a one-on-one conversation. Participants were asked to do several things during the one-hour meeting. Some participants were asked to spend time on the NAEP home page and give their impressions of the organization and structure of that page and to navigate links off that page of their own choosing, all the while providing running commentary on what they saw and did. These were *undirected* observations, designed to gather information about user interests and perceptions without constraining users too heavily. Other participants were asked to start on the NAEP home page and complete several brief but specific common tasks (*directed* observations). Some tasks, such as finding out information about the 2005 NAEP Science results, involved rather obvious links on the NAEP home page, while others required participants to use different links on the page and perhaps explore more deeply to find the information.

Background and Purpose

With billions of pages' worth of information, the World Wide Web has become a go-to source for information on countless subjects. It is fast becoming something of a rule, rather than the exception, that many services, products, programs, and—increasingly—people are associated with one or more Universal Resource Locators (URLs) in some way, and more information about practically anything can be obtained with a brief sequence of clicks. The National Assessment of Educational Progress (NAEP), as *The Nation's Report Card* and a national indicator of student academic performance in a variety of content areas, provides a tremendous amount of information about its testing program on the Web. From results to content and resources specific to audiences including educators, policymakers, parents, and the media, users can read through many pages' worth of information about NAEP, print off reports for later review and distribution, and answer their own data questions using several interactive tools.

For visitors to the NAEP site (<http://nces.ed.gov/nationsreportcard/>), the purpose of this study was to elicit opinions about the NAEP Web site (<http://nces.ed.gov/nationsreportcard>) from a range of individuals involved in educational assessment. Participants were asked to provide feedback about multiple aspects of the NAEP site, including the content and navigation of the home page (and links from that page), the State Profiles pages, and the interactive tools such as the NAEP Data Explorer and the NAEP Question Tool.

Method

This study took place June 26 and 27, 2006, in San Francisco, Calif., during the Council of Chief State School Officers' Conference on Large-Scale Assessment. About a month prior to the conference, an electronic version of the attendee list was obtained from CCSSO, and from that list conference attendees associated with national policy organizations, state or regional policy organizations, the media, local school districts, and state departments of education were identified. This process yielded a large number of possible participants. From this pared-down list, several criteria were used to identify an initial sample of participants to invite to participate in the research study. Included in the CCSSO attendee list are individual job titles, and efforts were made to identify personnel broadly or specifically involved in working with test data in a variety of contexts or reporting results. In the case of district administrators, these were district assessment directors. However, at the state level, positions of interest were not content area specialists but assistant state testing directors and other administrative personnel³¹.

With the initial invitation list set consisting of over 100 names set, an e-mail invitation to participate was sent to 50 individuals. Individuals were invited to participate in a one-hour meeting during the CCSSO conference to explore the NAEP Web site and provide feedback on the user experience, and were offered an honorarium of \$100 for their participation.

During the meeting, participants met with a researcher from the University of Massachusetts Amherst in a one-on-one conversation. The meetings began with a brief description of the project and a review of what these individuals were going to be asked to do during their participation in the study. They were then provided with an informed consent document for their review and signature, and this was followed up by a brief background survey and completion of forms required for processing of the honorarium payments.

At this point the study began. A Dell Inspiron 6000 laptop computer with a 15.4" screen running Windows XP and Microsoft Internet Explorer was used for this study. The computers were connected to the Internet through a T1 connection. Participants were advised that they could choose to use the touchpad on the computer or a wireless mouse to manipulate the onscreen cursor as they navigated the NAEP Web site.

³¹ Subsequent utility study activities involve state testing directors and hence those individuals were not included in this study.

Part One: The NAEP Home Page

When asked to provide general comments on the appearance and navigational ease of the NAEP Home Page, users expressed a range of opinions. On the positive side, some users indicated that it was “not too busy—there’s a lot of stuff but there’s a lot to NAEP.” Several users thought that they could access anything they needed from this page, and that the page had a “table of contents feel” and appeared well-thought out in terms of organization. One participant also liked that on the monitor used for this study, almost the whole page was displayed and minimal scrolling was necessary to see the bottom. Others noticed that having the 2005 Science front and center was important, as they were aware that those results had been released recently.

One user commented that the assortment of links on the home page seems to be determined by which results are current. Given that comment from one user, others also raised similar questions about how items were chosen for the NAEP front page and how the page’s “real estate” was allotted. For example, with respect to the 2005 Science results, one user noted that not only did that appear on the “big splash” but also under “New and Noteworthy.” For some users this duplication raised questions for them about the “management of news about NAEP,” meaning that those results were what “we’re supposed to care about and look at, not other stuff.” Another user wanted obvious links to information about students with disabilities or English language learners on the home page.

In the following section are discussed some specific elements of the NAEP home page. These elements are highlighted in Figure A-1.

Figure A-1. The NAEP Home Page (June 26–27, 2006)

Institute of Education Sciences U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION STATISTICS

NewsFlash Staff Contact Site Index Help

Search NCES

Publications & Products Surveys & Programs Data Tools Tables & Figures Fast Facts School, College, & Library Search Annual Reports What's New? Kids Site

ABOUT NAEP...
[overview](#)
[current activities](#)
[long-term trend](#)
[high school transcript study](#)
[special studies](#)
[selected schools](#)
[parents](#)
[researchers](#)
[media](#)
[educators](#)
[policymakers](#)

SUBJECT AREAS...
[civics](#)
[geography](#)
[mathematics](#)
[reading](#)
[science](#)
[u.s. history](#)
[writing](#)
[other subjects](#)

Search N.

HELP SITE MAP CONTACT US GLOSSARY NEWSFLASH

SAMPLE QUESTIONS ANALYZE DATA STATE PROFILES PUBLICATIONS

National Assessment of Educational Progress

THE NATION'S REPORT CARD

Results of the 2005 National and State Science Assessment

NOW AVAILABLE

INSIDE NAEP

Results of the National Indian Education Study Now Available!

Explore the [results](#) of the National Indian Education Study (NIES). Part I: *Performance of American Indian and Alaska Native Fourth- and Eighth-Grade Students on the 2005 NAEP Reading and Mathematics Assessments.*

NEW & NOTEWORTHY

Results of the 2005 national and state science assessment were released May 24, 2006. View the archived [webcast](#) of the data release event held in Washington, DC. Read the [transcript](#) of the StatChat—an online discussion with NCES Associate Commissioner Peggy Carr—about the results.

The 2007 NAEP Secondary Analysis Research Program is receiving applications until July 27, 2006. See [more information and application forms](#).

Are you going to the Council of Chief State School Officers (CCSSO) National Conference on Large-Scale Assessment? Check out a [list of NAEP and NAEP-related sessions](#).

Last updated 16 June 2006 (AA)

NCES Headlines

- ▶ [NEW REPORT! - Dropout Rates in the U.S.: 2002 & 2003](#)
- ▶ [NEW REPORT! - Average Freshman Graduation Rates](#)
- ▶ [Profile of Undergraduates in Education Institutions: 2003-04](#)
- ▶ [Documentation for the NCES Comparable Wage Index Files](#)

[Pubs/Products](#) | [Surveys/Programs](#) | [DataTools](#) | [Tables/Figures](#) | [FastFacts](#) | [School/LibrarySearch](#) | [Annuals](#) | [What's New?](#) | [Kids Site](#)

Search NCES

[NewsFlash](#) | [Staff](#) | [Contact](#) | [Site Index](#) | [Help](#)
[Privacy & Security Policy](#) | [Statistical Standards](#) | [RSS](#) | [FedStats.gov](#)

Institute of Education Sciences
U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION STATISTICS

1990 K Street NW, Washington, DC 20006, USA, Phone: (202) 502-7300 ([map](#))

Left Navigation Menu. Of all the visual and navigational elements of the NAEP home page, this menu found on the left-hand side of the page was among the most commented on.

First, with respect to the options listed under “About NAEP” many participants had many comments about these options, and these reflected both positive feedback as well as suggestions for improvement or change. They liked that different audiences were explicitly listed on the menu, and especially that information for Parents was available. However, three people asked about the decision-making process that identified the groups for the left-menu, and one participant questioned the audience listing, noting that educators is a broad category and perhaps there might be a mouse-over menu for the educators heading (state people, content people, district people, teachers, etc.) Another question was also brought up about the lack of a general “community” or “public” link: “If you’re not a parent, nor an educator, nor a policymaker, where do you go?” Examples of these individuals cited by participants included community leaders, such as chamber of commerce personnel, economic development officers, and the like. “Are those people considered policymakers?”

One user asked what “Current Activities” meant: was that 2005 Results? 2006 administrations? 2007 plans? This user desired a simple link with dates of assessments and results releases in that place. Several others inquired about the inclusion of the High School Transcript Study—“It seems to be a limited-interest link,” and [given that the data were about five years out of date] “How relevant or important is that to get a coveted place on the home page on the fixed menu?” Participants also probed into how Long-Term Trend, the Transcript Study, and Special Studies are different—maybe those could be grouped and linked to via a mouse-over menu.

Next up, several participants questioned the ordering and selection of “Subject Areas” in the left menu. They recognized that the information contained there was ordered alphabetically, though they suggested that Reading, Mathematics, and (maybe) writing were what mattered most per *NCLB* and should be first, then the other subjects. This led some of them to click on Other subjects, and find links to not just the content areas omitted from the first page (Arts, Foreign Languages, etc.) but all content areas, which they found a little confusing. A suggestion was made to revise this list to reference NAEP results in terms of four main areas of results: Long Term Trend NAEP, State NAEP, National NAEP, and NAEP Special Studies.

For the participants who clicked on the 2005 Science results in the middle of the page, they were transported to the Initial Release Site (IRS). While a direct evaluation of the IRS was not a part of this study, three users did comment that they liked the look and feel of the left menu on the IRS (shown below in Figure A-2) as compared to the main NAEP Home Page (Figure A-1).

Figure A-2. The www.nationsreportcard.gov Home Page (captured Dec. 27, 2006)



Home National Assessment of Educational Progress

The Nation's Report Card™

Report Cards

Science

- Summary
- Urban District Results
- Student Group Results
- District Comparisons
- Sample Questions

Resources

Information For...

- Media
- Parents
- Educators
- Researchers
- Policymakers

Learn More

- The Nation's Report Card
- About Urban Districts
- Downloads & Tools
- Glossary
- Help

November 15 2006, WASHINGTON, D.C.

Science Report for the Trial Urban District Assessment Now Available

In many districts, the average scores for White, Black, Hispanic, and Asian/Pacific Islander students at both grades were either higher or not significantly different from the national average for their peers.

View an archived copy of the [webcast](#) of the press release held on November 15, 2006.

Read a transcript of the [StatChat](#) with Associate Commissioner Peggy Carr about the results.

StatChat questions and answers

See [Help](#) to learn how to best use this site.

NCES, IES links, menu bars. The next issue concerning the NAEP home page that arose for many participants involved the presence of the NCES and Institute of Education Sciences links and bars (Figure A-1). A number of participants indicated that the having the IES-NCES bar at the top of the page was sometimes distracting. They wanted to click on something there to get them back to the NAEP home page. One participant remarked, “[It’s] hard to ignore the IES links at top of page,” while another clicked on one of the IES links to get back to the home page and found that “annoying.”

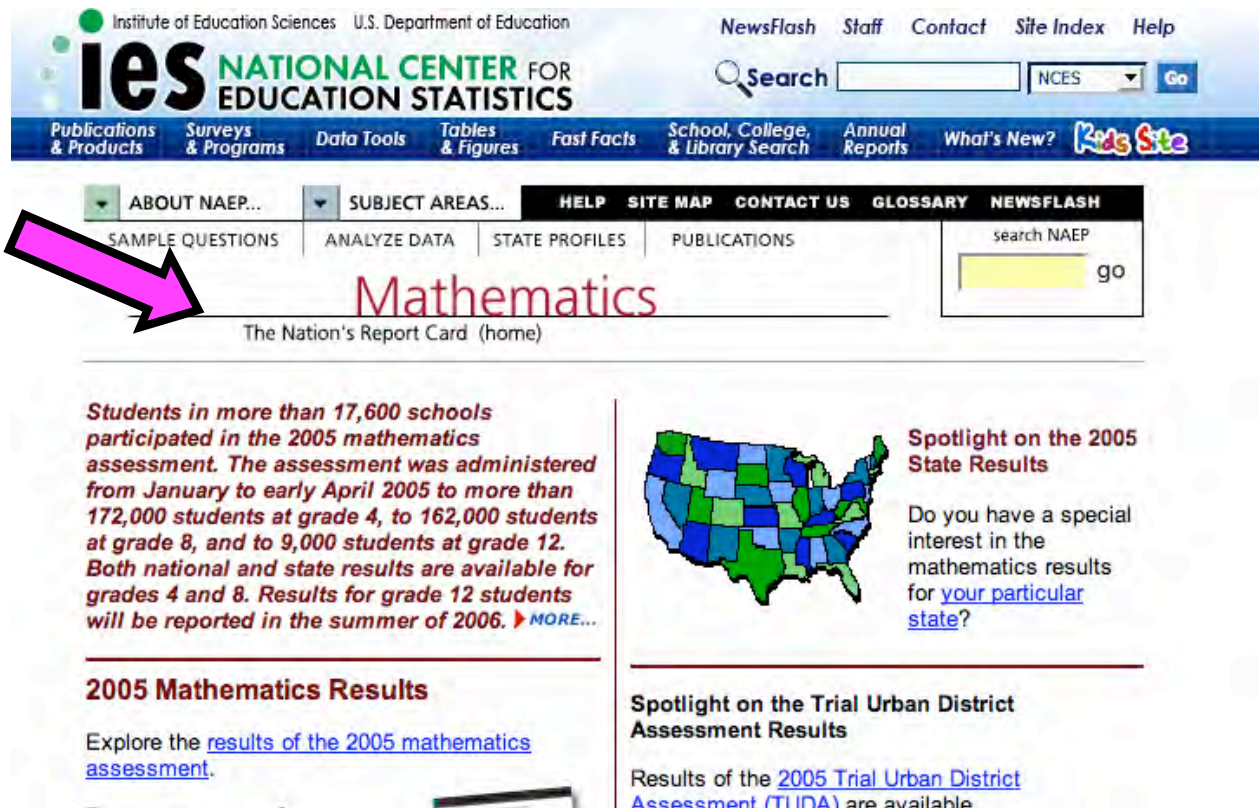
Multiple search bars. On multiple occasions, when participants wanted to search out some term or NAEP specific information, they saw Search and used that (it searches all of 1) NCES, 2) Products, or 3) Tables by way of a drop-down menu). There is also a small Search NAEP box and the same NCES search at the bottom. Most participants used the NCES search, overlooking the NAEP one, though several did use the NAEP search utility. Searching NAEP alone needs to be placed more prominently.

New and Newsworthy. Several participants commented that there were a lot of things that were marked as “New” as they looked at the NAEP Web site, and in some cases wanted to know why it couldn’t be more consolidated. At the top of the page in the IES-NCES bar, there is a NewsFlash and also What’s New? (6a), then in the black bar where the NAEP-specific part of the page begins there’s a Newsflash (6b), and about a fourth or a fifth of the page is taken up by “New and Noteworthy” (6c). While some of these links lead to different information, and some lead to the same information, there’s some duplication that participants found confusing.

Getting back to the NAEP Home page. In the course of the study, most of the participants wanted periodically to return to the NAEP Home page. Most of the individuals independently reported that this was not an easy thing to do and often used the browser’s Back button rather than clicking on the link to the home page (marked with a pink arrow below). Based on comments such as “Why is it in the middle? Society has trained us to expect a site to look a certain way... look up top and left for home page buttons” and “Links should look like links,” suggestions for improvement include redesigning the link to look more like something

clickable, making it more obvious or prominent on the page, and moving the link to the top and left.

Figure A-3. Illustration of the Link to the NAEP Home Page



Below are some examples of how other sites have positioned their “Home” link.

From www.whitehouse.gov:



From www.boston.com:



From www.firstgov.gov:



Part Two: State Profiles

As part of the semi-directed portion of the observations, all users were asked to click on the “State Profiles” link on the NAEP home page, and on the page that loaded from that link, to select any state’s results to explore and provide feedback on. In most cases, users chose their home state. The content of the State Profiles pages was in large part appreciated by participants. Users indicated that for the most part the information is accurate and a “good snapshot” of a state and its NAEP performance, though some noted that because it reflects Common Core of Data school information from the 2003–04 academic year the information is perhaps a bit old and could benefit from updating. One participant particularly liked the layout, indicating that it was consistent with how educational data of this nature is generally presented to policymakers, and other participants described the layout of the pages as “easy to read” and “not cluttered.” A least two participants remarked that they were glad to have the historical performance for both scale scores and achievement levels. Linking the graphs of performance for the selected state and the nation was also cited as an attractive feature of the State Profiles pages, because “graphs are good.”

Several comments about possible improvements or changes to the main State Profiles pages were also received from participants.

- The results on the page do not denote statistical change (within-state, from administration to administration).
- The Achievement Level results in both tables and graphs do not include the percent of kids below *Basic* (“It would be good to fully illustrate all categories”).
- When National Public is chosen as a jurisdiction from the drop down menu, no background information is provided, as in student characteristics (Number enrolled, percent in Title I schools, etc.), racial or ethnic background, and school or district characteristics.

Links to Cross-State Comparison Maps. One significant navigation issue that was raised when participants were looking at the State Profiles concerned how they accessed the cross-state comparison maps. When looking at the state profiles, many of them saw the links below:

Figure A-4. Links to the Cross-State Comparison Maps

History of NAEP Participation and Performance

Subject	Grade	Year	Scale Score		Achievement Level			Graphics
			State Avg.	[Nat. Avg.]*	Basic	Proficient	Advanced	
Mathematics (scale: 0-500)	4	1992 ⁿ	227	[219]	68	23	2	<ul style="list-style-type: none"> ● Scale Scores ● Achievement Levels ● Cross-State Comparison Maps: <ul style="list-style-type: none"> ○ Scale Scores ○ Percent at or Above Proficient
		1996 ⁿ	229	[222]	71	24	2	
		2000	233	[224]	77	31	3	
		2003	242	[234]	84	41	6	
		2005	247	[237]	91	49	8	
	8	1992 ⁿ	273	[267]	63	23	3	
		1996 ⁿ	278	[271]	68	28	5	
		2000	279	[272]	70	30	5	
		2003	287	[276]	76	38	8	
		2005	292	[278]	80	43	11	

Below is a screen capture of the page that loaded up when they clicked on either of those links.

Figure A-5. Accessing the Cross-State Comparison Maps

Institute of Education Sciences U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION STATISTICS

NewsFlash Staff Contact Site Index Help

Search [] NCES Go

Publications & Products Surveys & Programs Data Tools Tables & Figures Fast Facts School, College, & Library Search Annual Reports What's New? Kids Site

ABOUT NAEP... SUBJECT AREAS... HELP SITE MAP CONTACT US GLOSSARY NEWSFLASH

SAMPLE QUESTIONS ANALYZE DATA STATE PROFILES PUBLICATIONS

search NAEP [] go

State Profiles

The Nation's Report Card (home)

Cross-State Comparisons, Average Scale Scores: Massachusetts

Cross-state comparisons are available in two formats below. Select jurisdiction... ▾

Click on a symbol in the chart below to view a Scalable Vector Graphics ([requires SVG viewer](#)) comparison map.

For more information about SVG graphics and instructions on how to cut and paste SVG images please see our [help page](#).

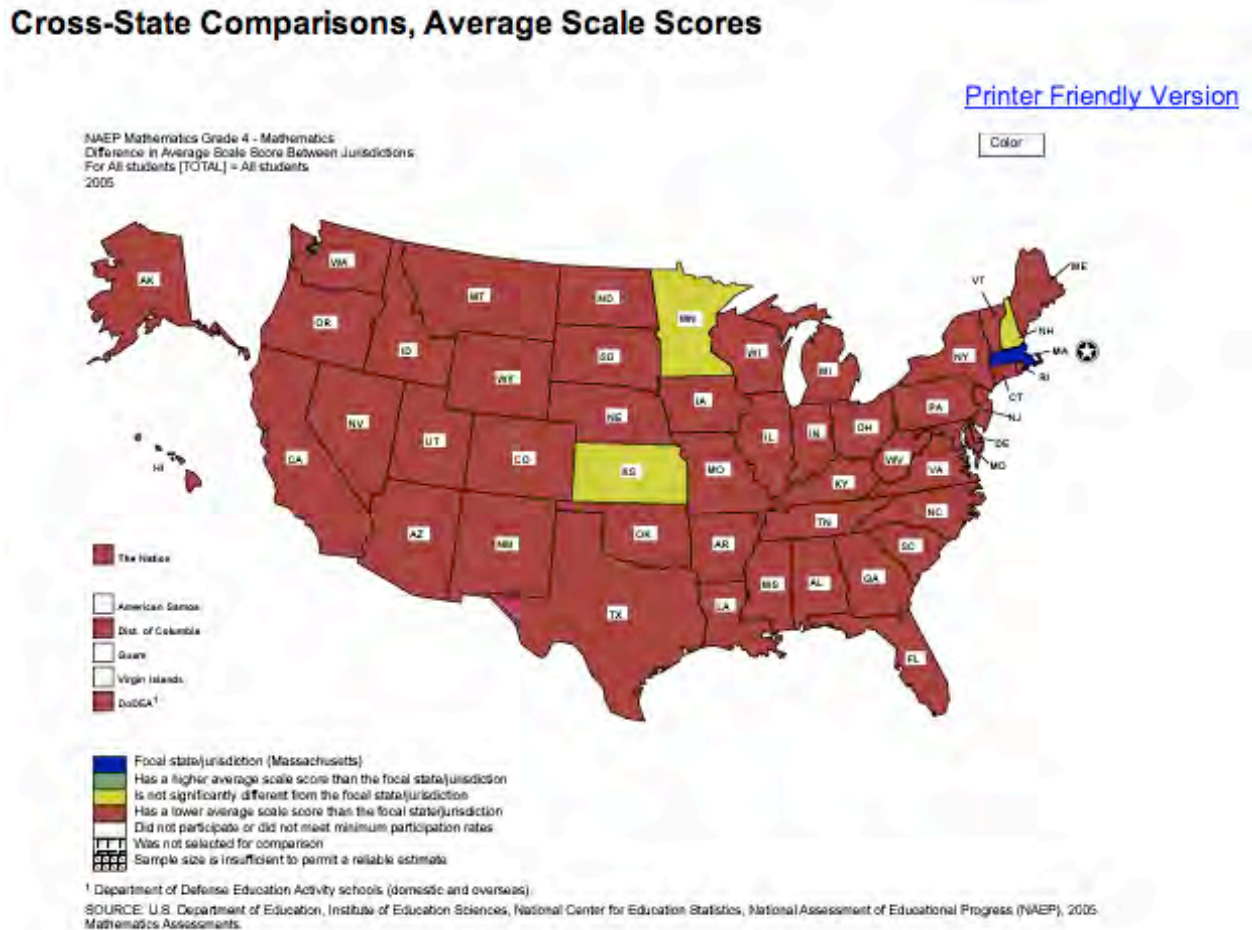
- ✓ participated
- ✗ did not participate or did not meet minimum participation requirements

Subject	Grade	Year								
		1990 ⁿ	1992 ⁿ	1994 ⁿ	1996 ⁿ	1998	2000	2002	2003	2005
Mathematics	4		✓		✓		✓		✓	✓
	8	✗	✓		✓		✓		✓	✓
Reading	4		✓	✓		✓		✓	✓	✓
	8					✓		✓	✓	✓
Science	4						✓			✓
	8				✓		✓			✓
Writing	4							✓		
	8					✓		✓		

ⁿ Accommodations were not permitted for this assessment

Each of the checkmarks in that table are links to the clickable cross-state comparison maps, although all of the participants who reached this page were unanimous in their comments that this table was not what they had expected to see (they expected cross-state maps). In addition, all participants noted that the checks in the table do not look like links, and “You wouldn’t know that unless you moused over the table” despite the instructions. This was a source of considerable frustration for most participants who viewed this, and several reiterated that on Web pages, links should look like links. This is a capture of a cross-state comparison map (what they expected to see):

Figure A-6. Cross-State Comparison Maps



It should be noted that the overwhelming feeling across participants was that these maps are among the most effective displays that NAEP uses to communicate between-state differences, and users liked these maps immensely. The main problem identified by study participants concerning their use was the difficulty in accessing them (“If you click on a link that says clickable maps, that’s what you want, and it’s frustrating when those aren’t there”). One suggestion from a participant was a small light bulb icon with “Tips for Using this Graph.”

NAEP Data Explorer

The version of the NAEP Data Explorer (NDE) currently available on the NAEP Web site was released in fall 2005, coinciding with the release of the 2005 Grade 4 and 8 Math and Reading results. Among the participants in this study, this updated version of the NDE was unfamiliar to most. Of the several who did report knowing about it, one commented, “I didn’t realize it was this good.” One expressed some frustration at not being aware of the NDE’s existence and a perceived lack of publicity about some of NAEP’s interactive tools, citing her role as an assessment person in a large urban district.

In this portion of the site usage observations, participants were directed to the data explorer welcome page (<http://nces.ed.gov/nationsreportcard/nde/>) and given the instruction to briefly familiarize themselves with the tool and take a few minutes to run a few analyses of their choosing. All users selected “Quick Start” after scanning the welcome page, and none opted to read either the Quick Start or Advanced introductions or clicked on the tutorial or help links on the welcome page, though later in using the tool several commented about the helpfulness of what one user termed the “info dots” (clickable blue circles with a white letter “i” leading to information or help).

Participants were permitted to choose the demographic variables, jurisdictions, and types of results they wanted to explore. A sampling of the analyses done by several participants is listed below.

- Grade 8, Reading, Colorado, Parents’ Education, ELL status
- Grade 8, Math, West region, Students with Disabilities
- Grade 8, Math, Texas, Ethnicity
- Grade 8, Reading, San Diego district, ELL status
- Grade 8, Math, Indiana, Parents’ Education Level
- Grade 4, Reading, Wyoming, All Students
- Grade 4, Math, Missouri and DC, Gender
- Grade 4, Reading, Houston, Nation, and Louisiana, Race/ethnicity, All years

The participants reported generally quite favorable impressions of the NDE. They found it to be “powerful,” “intuitive and great,” and “neat.” They also liked the flexibility provided by the tool to “let you choose what you want to look at” and to “build your own data”. Features such as being able to export to Excel, making graphs, and being told that the NDE was processing while it was gathering the requested data were also much appreciated. One participant wished that a NDE-type tool was available for individual state *NCLB* assessment results.

Several suggestions for improvement were also made in the course of the users’ experience with the NDE. One disconnect that was seen for many of the users was that they did not understand why some jurisdictions and variables “grayed out” when they chose a specific grade or grade and subject area combination.³² For some users with a high degree of familiarity with NAEP’s different national and state samples, this may not be an issue, but across the range of experience with NAEP seen in this group of users, this was a source of confusion. Furthermore, in selecting criteria, including demographic variables and jurisdictions for analysis, one participant expressed a preference for a greater level of flexibility than currently permitted. Specifically, this individual wanted all jurisdictions and variables to appear in both Box 3 and Box 4, rather than feeling forced into a jurisdiction in Box 3 and a variable in Box 4. For this participant, if that flexibility were allowed, “[It] would open the possibility of answering more questions with data, if allow people to define their own cross-tabulations.” For several other users, crosstabs was attempted but reported being unable to figure it out and were observed abandoning those analyses; another user had the same feeling about testing for statistical significance. One additional suggestion received was to devise a way to graph two variables.

³² For example, when a user selects Grade 12 Civics, the only jurisdiction options are National and National Public, because results for individual states are not available for analysis in this tool.

Minor Content Suggestions for the NAEP Web site

During these observations, a number of the participants offered smaller suggestions for additional or clarifying content for the NAEP Web site, based on things that came up for them during the study. These comments and suggestions are given below.

- In looking at the data in a state profile, one user was curious as to what year it was from, because it seemed slightly out of date based on that person's experience. This individual did find the footnote with info (Common Core of Data, 03-04), but then wanted to know what the Common Core of Data was, and suggested that users might appreciate a glossary-style link defining the Common Core of Data.
- Two users noted that at least some of the Parent info on the NAEP site is in Spanish, and urged NAEP to consider offering more resources and materials in Spanish as well as other languages.
- One user asked if any data or information was available on migrant and homeless students as a defined population.
- One suggestion from a user concerned the labeling on some graphs: when the data is described as 2005 results, are the results reflective of the 2004–05 or 2005–06 calendar year?
- Concerning the state profiles, when a user chose Florida from the drop-down menu, the number of school districts was defined in a footnote as “Local school districts only (type 1, 2).” The user was unable to find out what that meant.

State Profiles: Selected Participant Observations

Also, some TUDA-specific comments:

- One participant was particularly familiar with TUDAs and wanted more information about those, such as a map with TUDAs listed that people could click on and get information similar to what is available for states via the State Profile page.
- Overall, several participants found it difficult to find information about TUDAs. It's not under Special Studies, and given the difficulty finding the “right” search box, this was a point of frustration for some participants.

**Appendix B: Do-It-Yourself NAEP Data Analysis on the Web:
Evaluating the Usability of the NAEP Data Explorer**

April L. Zenisky
Center for Educational Assessment
University of Massachusetts Amherst

Feb. 2, 2007

Introduction

In NAEP, reporting scores to stakeholders typically has involved the preparation of summary document with information summed across groups of examinees, often for the Nation and the states, and by a number of different demographic breakdowns. These reports are quite professional and informative at the level appropriate for most users of NAEP, and are the result of considerable collaboration between the National Center for Education Statistics, many of the NAEP Alliance contractors, and the National Assessment Governing Board that sets policy guidelines regarding reporting efforts. At the same time, such intact documents are not intended to answer every possible question about student performance that some stakeholders may have with respect to NAEP, and so in the past several years a concerted effort has focused on providing users of NAEP information with a variety of sources of information about NAEP. One approach in particular involves the NAEP Web site (<http://nces.ed.gov/nationsreportcard/>) and drawing on the interactive nature of communications via the World Wide Web. In addition to the familiar, static technical reports of student performance, NAEP has developed the NAEP Data Explorer (NDE), which allows visitors to the Web site to explore decades' worth of NAEP data from the comfort of the offices or homes, at the click of a mouse. The current online NDE built on an earlier data tool available via the NAEP Web site, the NAEP Data Tool (NDT). The NDE was released to the public at the time of the Mathematics and Reading 2005 release, in October 2005.

A recent review of NAEP Web site usage statistics found that use of the NDE was growing considerably from its release in October 2005 (Appendix C of this chapter). However, in keeping with the long tradition of research on NAEP reporting efforts (see, for example, reports by Levine, Rathbun, Selden, and Davis (1998), Hambleton and Slater (1995), Jaeger (1992, 2003), Koretz and Deibert (1993), Linn (1998), Linn and Dunbar (1992), Wainer (2000b, 1997, 1996), Wainer, Hambleton, and Meara (1999), De Mello (2004), Ogilvy Public Relations Worldwide (2004), the National Research Council (2001), Simmons and Mwalimu (2000), and Hambleton (2002)), the NDE should be evaluated periodically for its functionality and contribution to the larger reporting efforts of NAEP. The purpose of this study, therefore, is to learn more about user impressions and the overall usability of the NAEP Data Explorer among different constituent groups for NAEP.

Methodology

As part of efforts to evaluate this aspect of the NAEP Web site, two Utility study activities were undertaken. First, participants in the Web site user observations that took place during the Council of Chief State School Officers' conference were asked their impressions of the NDE (these participants were state and district education personnel). In total, nine participants were involved in this activity. This study took place June 26 and 27, 2006, in San Francisco, Calif., during the Council of Chief State School Officers' Conference on Large-Scale Assessment, and involved individuals from state and district education officials and policymakers (job titles included superintendent of measurement, principal consultant on student achievement, testing and evaluation director, program manager for assessment services, and chief accountability officer). The participants included seven females and two males, and five persons reported having been involved in education for at least 20 years while three indicated 11 to 20 years and one with 5 to 10 years' worth of experience. Four identified themselves as "very familiar with NAEP" while five were "somewhat familiar", and as a group worked with NAEP data from once or twice a week (1 person) and once a month (3 persons) to once or twice a year (3 persons) to never (2 persons). Ways that these participants worked with NAEP data and information included making NAEP-state comparisons, facilitating NAEP administration in state or district schools, and item reviews.

During the meeting, participants were asked to start with the entry page for the NDE (Figure B-1, <http://nces.ed.gov/nationsreportcard/nde/>), and to familiarize themselves with the options and information there as they chose. Then participants were asked to carry out several

analyses based their own interests. Throughout the observations, participants were asked to employ a think-aloud protocol to provide the researcher with additional information about their perceptions of the tool. Because this evaluation of the NDE took place within the context of a larger study of the usability of the NAEP Web site, these participants spent a total of about 10–15 minutes engaged in the NDE.

Figure B-1. NAEP Data Explorer Start Page

Welcome to the NAEP Data Explorer! [Get help](#)

Do you have questions about what the nation's students know and can do? With the NAEP Data Explorer (NDE) you can create statistical tables and graphics to help you find answers. Explore the results of decades of assessment of students' academic performance, as well as information about factors that may be related to their learning.

Choose a version...

Quick Start – Provides convenient access to data about student performance in terms of NAEP's scale scores and achievement levels for major reporting group variables. Enables you to format basic tables and graphics.

[Go to Quick Start](#) [Read the Quick Start introduction](#)

Advanced – Provides full access to student groups' scale scores and achievement-level performance for any NAEP variable; allows additional flexibility in generating and formatting data tables and graphics.

[Go to Advanced](#) [Read the Advanced introduction](#)

Accessible version of the NAEP Data Explorer.

First time user? Can't remember where to begin? [View the tutorial.](#)

NOTE: Data for the [2004 Long-Term Trend assessment](#) are only available in HTML format. Data for the [1997 Arts assessment](#), [1996 Long-Term Trend assessment](#), and the [1999 Long-Term Trend assessment](#) are only available in PDF format.

What next? [Click the "Quick Start" or "Advanced" button to begin. . . or read one of the introductions.](#)

[About NAEP Data Explorer](#) [Important Legal Information](#)

A second, targeted study of the NDE took place at the annual meeting of the Northeastern Educational Research Association conference, held in Kerhonkson, N.Y. (October, 2006), involving five participants. This study of the NDE was targeted in the sense that the population of interest in this case was post-doctorate educational researchers, in order to learn more about the usability of the NDE among users with advanced training in educational statistics and data manipulation. This group consisted of three men and two women, and all but one reported more than five years' worth of experience working in education. Four of the five reported being somewhat familiar with NAEP while one was very familiar, and the frequency of working with NAEP data or information ranged from rarely or never (three participants) to a few times a year (one participant) and a few times a month (one participant).

The method for these observations was similar to the approach used in the earlier observations, but with several noteworthy differences. In addition to focused use of the tool with respect to running analyses under Quick Start mode, these participants were also explicitly requested to explore several advanced features of the NDE, including the option to evaluate statistical significance, to create graphs based on the results, to export results to Excel, to carry out regression analyses, and to use Advanced mode. These aspects of the NDE were not explored in great depth by the earlier set of observations but were determined to be elements of

the tool that were critical to the main goal of assessing the overall functionality of the NDE for carrying out analyses of NAEP data as permitted. Participants in this directed study spent 45 minutes to an hour carrying out analyses and discussing perceptions of the NDE.

The results of both of these sets of observations, described in brief below, provide considerable insight into the user experience of the NDE and likewise identify several important recommendations for improvement. The results of the observations that took place at CCSSO are reported first, followed by an overview of results from the second portion of the study. The remainder of the paper provides a summary and discussion of findings.

CCSSO Observations

Among the participants in this part of the NDE study, this updated version of the NDE was unfamiliar to most. Of the several who did report knowing about it, one commented, “[I] didn’t realize it was this good.” One expressed some frustration at not being aware of the NDE’s existence and a perceived lack of publicity about some of NAEP’s interactive tools, citing her role as an assessment person in a large urban district.

In this portion of the site usage observations, participants were directed to the data NDE welcome page (Figure 1), and given the instruction to briefly familiarize themselves with the tool and take a few minutes to run a few analyses of their choosing. All users selected “Quick Start” after scanning the welcome page, and none opted to read either the Quick Start or Advanced introductions or clicked on the tutorial or help links on the welcome page, though later in using the tool several commented about the helpfulness of what one user termed the “info dots” (clickable blue circles with a white letter “i” leading to information or help).

Participants were permitted to choose the demographic variables, jurisdictions, and types of results they wanted to explore. A sampling of the analyses done by several participants is listed below.

- Grade 8, Reading, Colorado, Parents’ Education, ELL status
- Grade 8, Math, West region, Students with Disabilities
- Grade 8, Math, Texas, Ethnicity
- Grade 8, Reading, San Diego district, ELL status
- Grade 8, Math, Indiana, Parents’ Education Level
- Grade 4, Reading, Wyoming, All Students
- Grade 4, Math, Missouri and DC, Gender
- Grade 4, Reading, Houston, Nation, and Louisiana, Race/ethnicity, All years

The participants reported generally quite favorable impressions of the NDE. They found it to be “powerful,” “intuitive and great,” and “neat.” They also liked the flexibility provided by the tool to “let you choose what you want to look at” (Figure B-2) and to “build your own data.” Features such as being able to export to Excel, making graphs and being told that the NDE was processing while it was gathering the requested data were also much appreciated. One participant wished that an NDE-type tool was available for individual state *NCLB* assessment results.

Figure B-2. Analysis Selection Page, Quick Start Mode

[About NAEP Data Explorer](#) [Important Legal Information](#)

Several suggestions for improvement were also made in the course of the users’ experience with the NDE. One disconnect that was seen for many of the users was that they did not understand why some jurisdictions and variables “grayed out” when they chose a specific grade or grade and subject area combination. As shown in Figure B-3, for example, when a user selects Grade 12 Civics, the only jurisdiction options are National and National Public, because results for individual states are not available for analysis in this tool. Furthermore, some variables likewise cannot be selected. For some users with a high degree of familiarity with NAEP’s different national and state samples, this may not be an issue, but across the range of experience with NAEP seen in this group of users, this was a source of confusion.

Figure B-3. Analysis Selection Page, Quick Start Mode with Options “Grayed Out”

[About NAEP Data Explorer](#) [Important Legal Information](#)

Furthermore, in selecting criteria, including demographic variables and jurisdictions for analysis, one participant expressed a preference for a greater level of flexibility than is currently permitted. Specifically, this individual wanted all jurisdictions and variables to appear in both Box 3 and Box 4 (see Figure B-2 above), rather than feeling forced into a jurisdiction in Box 3 and a variable in Box 4. For this participant, if that flexibility were allowed, that would “open the possibility of answering more questions with data, if allow people to define their own cross-tabulations.” For several other users, crosstabs was attempted but reported being unable to figure it out and were observed abandoning those analyses; another user had the same feeling about testing for statistical significance. One additional suggestion received was to devise a way to graph two variables.

NERA Observations

Among this group of participants, the NDE was largely unfamiliar (three had no prior use while the other two reported only having used it once or twice). As before, these individuals were asked to access the data NDE welcome page, and then told to briefly familiarize themselves with the tool and take a few minutes to run a few analyses of their choosing, followed by completion of a directed task.

Below are summaries of the user experience and possible changes suggested by participants, arranged by the following topics:

- Overall Impressions
- Opening Page and Usage Agreement
- Selection of Variables
- Data Analysis and Appearance of Results
- Help Links, Statistical Significance
- Graphing
- Exporting Results to Excel
- Advanced Mode
- Regression

Each point is classified as being a *Comment* (no suggestion for improvement), addressing the *Appearance* of elements of the NDE, relating to the content of the *Information* displayed, or concerning the *Functionality* of the NDE. When appropriate, suggested improvements are included.

- *Overall impressions*

Comment: Users were unanimous in expressing that the NDE is quite powerful, and as a tool was liked and appreciated by all users. They were impressed by the amount of data available and like the flexibility the tool gave users.

Comment: These individuals felt that the NDE was probably most useful for people with a specific question to research, as in the directed tasks portion of the study.

Comment: The inclusion of the “NCES processing screen” was cited by all as a useful addition. All users reported having been on Web sites where sometimes it can be hard to tell is a page request has timed out or a connection has gone bad or timed out, and the processing screen was reassuring to all participants.

The participants in this study were very positive in their perceptions of the NDE during and after participating in the study. Broadly speaking, these users indicated that this was a largely well-designed tool that provided visitors to the site with specific data-based questions about NAEP performance an opportunity to step outside of the bounds of a testing program’s customary paper reports and answer those questions independently. All emphasized that they perceived that the NDE would likely be most appreciated by those with specific research questions (versus random exploration of results), largely due to the sheer quantity of data and data analysis options available.

- *Opening page and usage agreement*

Comment: Most were explicit in reporting they did not read the usage agreement.

Appearance: Most users self-reported that they cursorily skimmed the opening page and immediately clicked on *Quick Start* to start. As one user pointed out, “The rationale for just jumping in is that the Quick Start and Advanced buttons are bigger and way more eye-catching.”

- *Suggestion: Make the link to the tutorial more prominent.*

- *Suggestion: Reduce the amount of text on this page.*

Information: Only one accessed one or two screens’ worth of narrated tutorial, but abandoned it after few minutes (described it as confusing and slow-paced).

- *Suggestion: Add a simple, prominently linked FAQ document.*

Information: Perhaps because they had all self-reported as having skimmed the introductory pages, all users expressed initial confusion as to what made Advanced mode advanced, and Quick Start mode more basic.

- *Suggestion: Users recommended the creation of a comparison link that specified the features of each so that users could make an informed decision on mode choice.*

Consistent with the previous set of observations, these users exhibited the tendency to “dive right in” and start running data. As shown in Figure B-1, there is a link to the tutorial, but

it is only text, appears on the right side of the Web page, and as noted by one user, is not nearly as eye-catching as the more prominent blue and green buttons that (respectively) lead to Quick Start and Advanced modes. The icon of the individual with a cane is not a clickable link. One user remarked, “Oh, they have something here about handicapped access—that’s good” (the user did not pursue the link, however). From the participants, suggestions for improvement of this section include some minor redesign of the portal page, with increased attention paid to the placement and design of a tutorial link and the information in the tutorial. One suggestion regarding the tutorial was to employ a two-pronged approach: keep the current, narrated version and create a small, FAQ-type reference document as well. Users did appreciate the “What’s Next?” bar and link at the bottom, finding it a good prompt of how to proceed.

All but one participant commented that he or she did not read the usage agreement, and the one who self-reported that he skimmed it said that the most salient information was the knowledge that the tool would time out after 20 minutes of inactivity. All saw the need for the agreement, however.

- *Selection of variables*

Appearance: Users found the layout of the variable selection box somewhat unintuitive.

- *Suggestion: Minor redesign of the variable selection page to more clearly delineate the sequence of choices.*

Information: Users did not understand why some variables were (states or jurisdictions, background variables) grayed out in the selection process.

- *Suggestion: Data availability for some grade and subject area combinations should be made more explicit.*

Information: Users wanted to know more about the background variable questions, indicating that if they were presenting any of these kinds of results at a conference or in an article, people would ask.

- *Suggestion: Hyperlink text and answer choices for background variables.*

Functionality: Users wanted to be able to compare across grades.

- *Suggestion: This was mentioned as a potential functionality change, and perhaps could be included as an option for Advanced mode.*

As shown in Figure B-2 above, users of the NDE in Quick Start mode have a number of choices to make about the analyses they will carry out, including grade, subject, jurisdiction, and variable. Users can select only one grade and subject, which then determines what jurisdictions and variables are available for analysis. For example, Grade 4 Civics is only available at the National and National Public level, while Grade 8 Mathematics results can be obtained for all listed jurisdictions. The previously discussed issue of some options in the Jurisdiction(s) and Variable(s) selection boxes being “grayed out” depending on grade or subject selection emerged as an equal source of frustration for these users as it was for the Web site observation participants at CCSSO. The suggestion was received that somewhere on the selection page it should be briefly explained why not all grade and subject combinations permit all analyses, particularly with respect to restrictions by Jurisdiction(s).

As to the finding that the sequence of choices was somewhat unintuitive for some participants, the difficulty was that they looked across first to see boxes 1, 3, and 4. In several cases, users chose a grade, a jurisdiction, a variable, and a year (completely overlooking subject) and then when they clicked “Go to Results” were prompted to enter a subject. One user also wondered why the tool did permit direct comparisons across grades, noting that that was often a reporting interest and many results that appear in NAEP reporting documents are displayed cross-year, including Long-Term trend.

The last substantive issue that was brought up relative to the analysis selection page involved the Variable(s) field. In addition to the “graying out” difficulty, several participants commented on wanting to know more about the specific background questions used. For example, one individual saw “School Location (2005)” and “School Location (9 categories) (2005)” but had no knowledge as to how those two variables differed without running an

analysis and then finding out which might have been something of interest. Being researchers, these individuals noted that if they were to publish or present results using these data, an editor or reviewer or discussant would want to know what the background questions were, but this information was not readily present in the data selection page. The suggestion was made to link the question text to the information in the selection box, perhaps as a pop-up or mouse-over.

- *Data analysis and appearance of results*

Information: One user was looking for clarification on the range of the NAEP score scale.

- *Suggestion: User requested a mouse-over or pop-up link to NAEP score information, such as a scale score range and a listing of the achievement levels.*

Information: Multiple users requested the inclusion of sample size data on multiple occasions, to the extent possible. They indicated that this was something that would be expected in presentations or publications using these data.

- *Suggestion: Provide sample sizes when possible.*

Functionality: Users did not want to have to click a radio button to access other results (percentages, achievement levels (discrete and cumulative), etc.).

- *Suggestion: Allow users to request all results at once or provide a checklist letting them choose all, some, or just one.*

Functionality: When users clicked on the radio button to change result type, the page reloaded and they “lost” whatever they had been looking at previously, meaning the new results appeared on a refreshed page and were NOT appended to the bottom of previous results

- *Suggestion: if possible append new results to existing analyses (as in SPSS).*

Functionality: One user noted that the headers in the table were in blue (typical color for links) and wanted to sort tables by those headers but that wasn't an option.

- *Suggestion: Make links appear as links, and things that are not links should be distinct and not link-like.*

Participants were quite interested in the displays of results obtained through the NDE. Initial impressions were generally positive at the quantity of information available, although participants expressed some preference to be able to see more data at once, rather than clicking the radio button to switch the results view (Figures B-4 and B-5).

Figure B-4. NDE Results Page, Average Scale Score

The screenshot shows the 'View Results' page for the NDE. At the top, there are navigation buttons: 'Switch to Advanced Mode', 'View Quick Start intro', 'Select Criteria', and 'Go to Results'. The main heading is 'View Results' with sub-links for 'Printer-friendly' and 'Save HTML / Export To Excel'. A 'Get help' button is also present. Below the heading, there is a paragraph explaining the results and a 'TIPS' icon. The main content area contains several filter sections:

- 'I want to see results by:' with radio buttons for:
 - average scale score
 - average scale score with percentages
 - average scale score with standard deviation
 - percentages only
 - achievement level (cumulative)
 - achievement level (discrete)
 - percentiles
- 'I want results in a:' with radio buttons for:
 - table
 - graphic (SYG)
- 'Show me:' with radio buttons for:
 - totals only
 - cross-tabulation
- 'Are differences statistically significant?' with a 'Find out' button.

Below the filters, there is a table titled 'Average scale scores for mathematics, grade 4, All students (TOTAL): By jurisdiction, 1990, 1992, 1996, 2000, 2003 and 2005'. The table has five columns: 'All students', 'Year', 'Jurisdictions', 'Average Scale Score', and 'Standard Error'. The data is as follows:

All students	Year	Jurisdictions	Average Scale Score	Standard Error
All students	1990 ¹	National	213	(0.9)
	1992 ¹	National	220	(0.7)
	1996	National	224	(1.0)
	2000	National	226	(0.9)
	2003	National	235	(0.2)
	2005	National	238	(0.1)

Footnote: ¹ Accommodations were not permitted for this assessment.
 NOTE: The NAEP Mathematics scale ranges from 0 to 500. Observed differences are not necessarily statistically significant.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1990, 1992, 1996, 2000, 2003 and 2005 Mathematics Assessments.

At the bottom of the page, there are navigation buttons: 'Switch to Advanced Mode', 'View Quick Start intro', 'Select Criteria', and 'Go to Results'. There are also links for 'About NAEP Data Explorer' and 'Important Legal Information'.

Figure B-5. NDE Results Page, Achievement Levels (Discrete)

Switch to Advanced Mode [View Quick Start intro](#) [Select Criteria](#) [Go to Results](#)

View Results

Printer-friendly [Save HTML / Export To Excel](#) [Get help](#)

Here are the results you've selected. From this point, you can view your results as a graphic, check whether differences in the results are statistically significant, view different performance measures, and create cross-tabulations using two variables. Other options are available in the [Advanced mode](#) of the NAEP Data Explorer.

I want to see results by:

- average scale score [i](#)
- average scale score with percentages [i](#)
- average scale score with standard deviation [i](#)
- percentages only [i](#)
- achievement level (cumulative) [i](#)
- achievement level (discrete) [i](#)
- percentiles [i](#)

I want results in a table graphic ([SVG](#))

Show me totals only cross-tabulation

Are differences statistically significant? [Find out](#)

Percentages of students at or above each achievement level for mathematics, grade 4, All students [TOTAL]: By jurisdiction, 1990, 1992, 1996, 2000, 2003 and 2005

All students	Year	Jurisdictions	Below Basic	Standard Error	At or above Basic	Standard Error	At or above Proficient	Standard Error	At Advanced	Standard Error
All students	1990	National	50	(1.4)	50	(1.4)	13	(1.2)	1	(0.4)
	1992	National	41	(1.0)	59	(1.0)	18	(1.0)	2	(0.3)
	1996	National	37	(1.3)	63	(1.3)	21	(1.1)	2	(0.3)
	2000	National	35	(1.3)	65	(1.3)	24	(1.0)	3	(0.3)
	2003	National	23	(0.3)	77	(0.3)	32	(0.3)	4	(0.1)
	2005	National	20	(0.2)	80	(0.2)	36	(0.2)	5	(0.1)

[†] Accommodations were not permitted for this assessment.
 NOTE: Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1990, 1992, 1996, 2000, 2003 and 2005 Mathematics Assessments.

Switch to Advanced Mode [View Quick Start intro](#) [Select Criteria](#) [Go to Results](#)

[About NAEP Data Explorer](#) [Important Legal Information](#)

Part of the frustration with these radio button links is that users referred to having “lost” previous results as they chose to look at additional results. Among the researchers, a common refrain was that they wanted a view that was similar to what they were familiar with in the SPSS or SAS statistical packages, as opposed to this format that is likely constrained by what can be done via the Web.

The color of the headers for the data tables was likewise raised as a difficulty for one user who perceived those labels (in blue) as ways to sort the data tables, pointing out that in Web usage, blue is typically a color that is indicative of a link that does something, and there are other phrases on the page that are similarly colored and are active hyperlinks. The user did note that these perceived header links were not underlined, but expressed a preference that they be made more distinct to eliminate or reduce the confusion.

The final suggestion obtained concerned the use of sample sizes. All of these participants were familiar with NAEP and the statistics involved in computing the scale scores and achievement level percentages, but noted that if they were to use these data, questions about sample size would be raised and it would be necessary to report something. The inclusion of standard errors was regarded as useful.

- *Help links*

Comment: The blue-circle “i” for help in defining terms was appreciated.

Appearance: At first the function of the TIPS icon was not clear but, later, users realized it was a clickable icon positioned as an arrow pointing at the directions on different NDE pages.

- *Suggestion: Consider reformatting or reorienting the label of the Tips button to make it clearer that this is a Help function.*

The available Help links were among the least utilized feature of the NDE, generally. Users did like the blue circles with the letter “i” as links to additional information when those were included. Some redesign of the appearance of the clickable TIPS icon was requested by participants: one indicated that because the word TIPS was positioned on an angle, visually that person overlooked it.

- *Statistical significance*

Comment: The users liked and wanted to see this information from the start.

Information: Users had many questions about the test(s) used here. They wanted to know explicitly how the test was built, what was the statistic type used was, and what was the alpha level. Two users explained what they were looking for as the output as they would expect to see with SAS or SPSS.

- *Suggestion: Additional, simplified documentation explaining the statistical significance tests implemented.*

Information: Users requested information to help interpreting significance test results, such as effect sizes.

- *Suggestion: Report Cohen’s d and provide text categorizing the effects as small, medium, or large. Help text for interpreting effect sizes should also be provided.*

Functionality: When they clicked on “check statistical significance” users expected that to be a seamless process, not a popup with a series of decisions to be made. Users perceived this process to be redundant to the process of variable selection.

- *Suggestion: Users wanted the significance results embedded in the regular display of results, perhaps appended to the end of previous results.*

The statistical significance feature was important to all participants. Before several users noticed the option to compute significance, they had remarked on how they wished that information was provided, so the inclusion of this was appreciated by this population. The action of carrying out significance testing was, however, perceived by participants as slightly tedious for several reasons. First, to access the significance test results users clicked the labeled icon on the results page, which opened a new browser window. Obtaining the results involved selecting a jurisdiction, year, variable, statistic type, and display option from the tree menus shown on the left in Figure B-6.

Figure B-6. Significance Testing Window from the NDE

Check Statistical Significance Get help

Make selections in the left-hand area below. You will see a preview of your statistical table on the right. The symbol points to selections that you must make in order to perform a statistical test. When you have made all the necessary selections this symbol disappears, and you can click on the "Compute" button.

Selections Reset

- Jurisdiction**
 - National
- Year**
 - Select All
 - 1990¹
 - 1992¹
 - 1996
 - 2000
 - 2003
 - 2005
- Variable**
 - All students**
 - All students
- Statistic Type**
 - Below Basic
 - At or above Basic
 - At or above Proficient
 - At Advanced
- Display Options**
 - Show Detail
 - Show Graph

Table Preview Compute

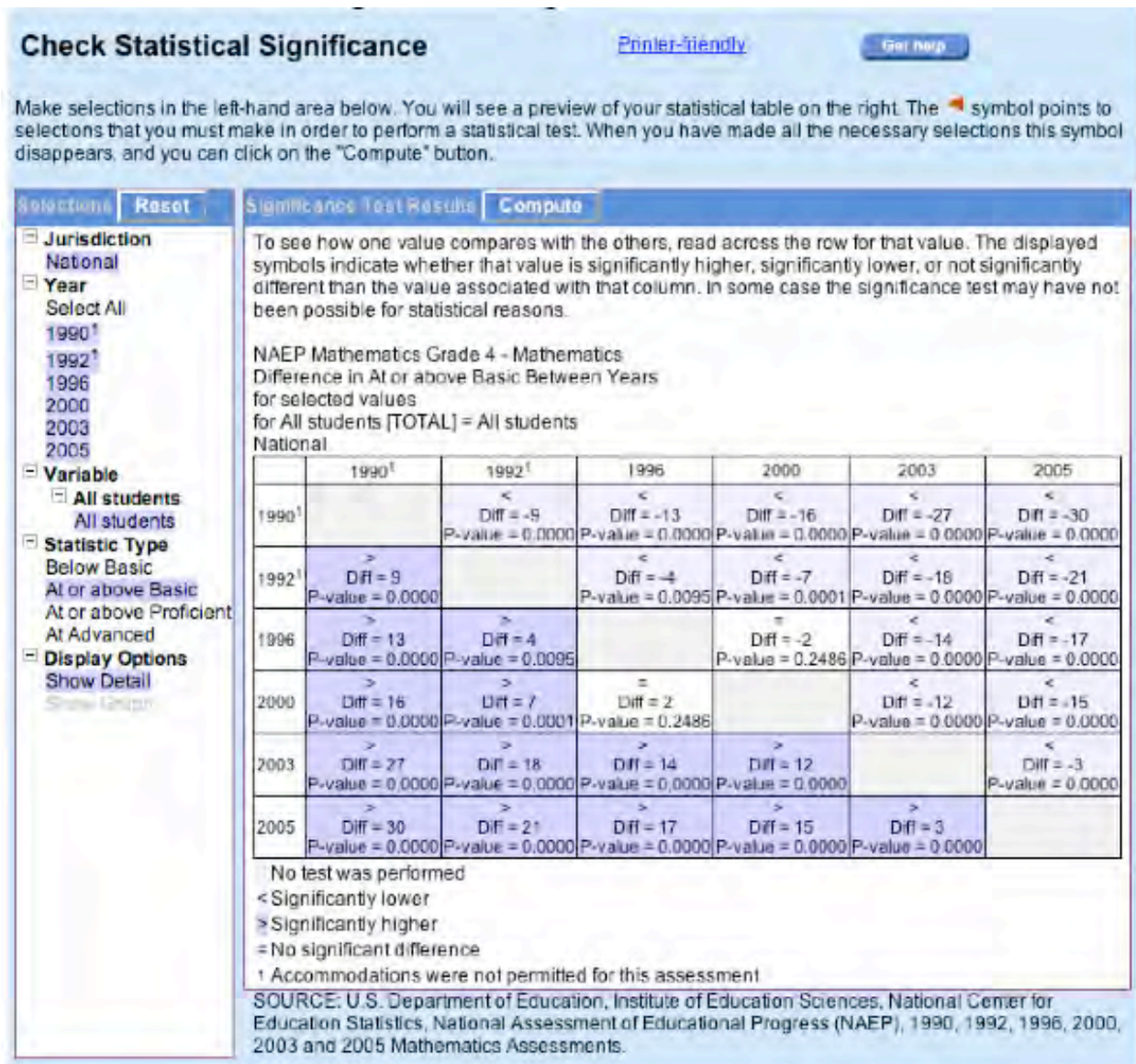
You have selected more than two years. In the printed reports, trend effect are tested only two years at a time. In your selection three or more years are simultaneously tested. Subsequently, your results will be more conservative than the printed reports. Select exactly two years, if you want your results to correspond to the printed reports.

NAEP Mathematics Grade 4 - Mathematics
Difference in Your choice of Statistics Between Years
for selected values
for All students = All students
National

	1990 ¹	1992 ¹	1996	2000	2003	2005
1990 ¹						
1992 ¹						
1996						
2000						
2003						
2005						

Users indicated that this process was perceived as extraneous given that they had already selected all of these options within the main NDE selection page, although they recognized that this selection was nested within the previously selected choices (meaning that this process allowed for significance testing on one or all combinations of selected variables). The display of the significance test results is given in Figure B-7.

Figure B-7. NDE Significance Test Results



In considering the information displayed in Figure B-7, participants again reiterated that they liked the availability of the information, but they wanted the layout to appear as it would have with analyses run in SPSS and SAS (which they described as what they were accustomed to). In addition, they requested the inclusion of effect size data and information to aid users in interpreting effect sizes.

- **Graphing**

Appearance: Users did not understand the difference between a full graph and a scroll window.

- *Suggestion: Provide simple help text to explain the difference.*

Functionality: One user's computer timed out in making a graph.

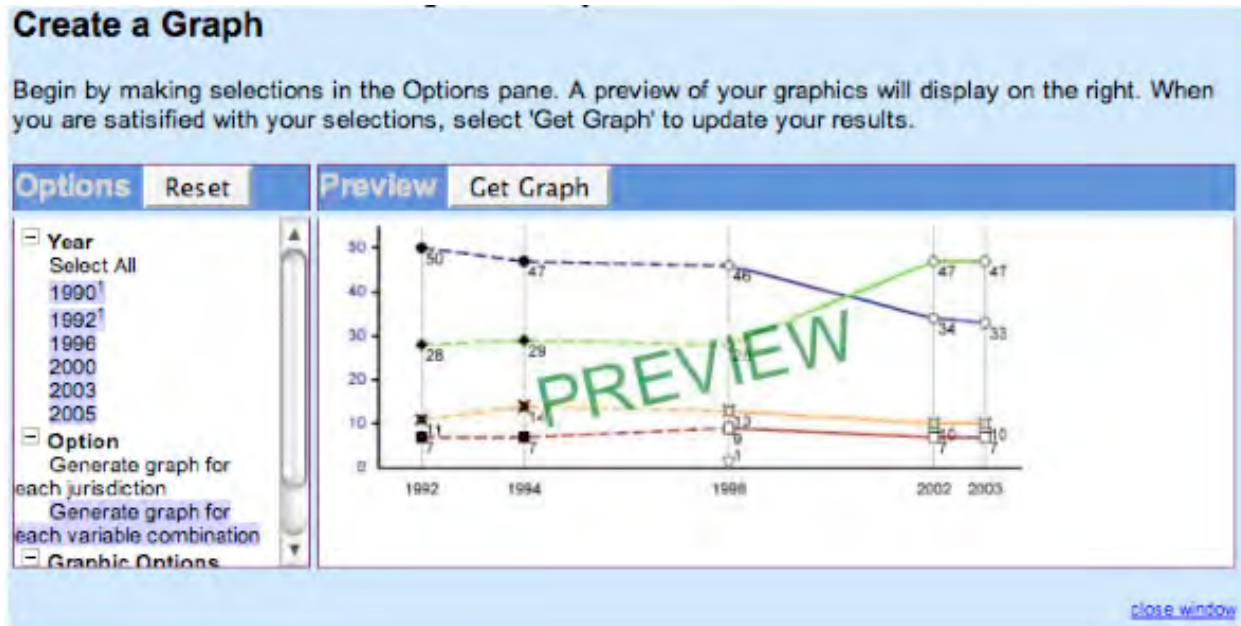
- *No specific suggestions offered*

Functionality: When they clicked on "create graph" users expected that to be a seamless process, not a popup with a series of decisions to be made. Users perceived this process to be redundant to the process of variable selection.

- *Suggestion: Users wanted the graphs embedded in the regular display of results, perhaps appended to the end of previous results.*

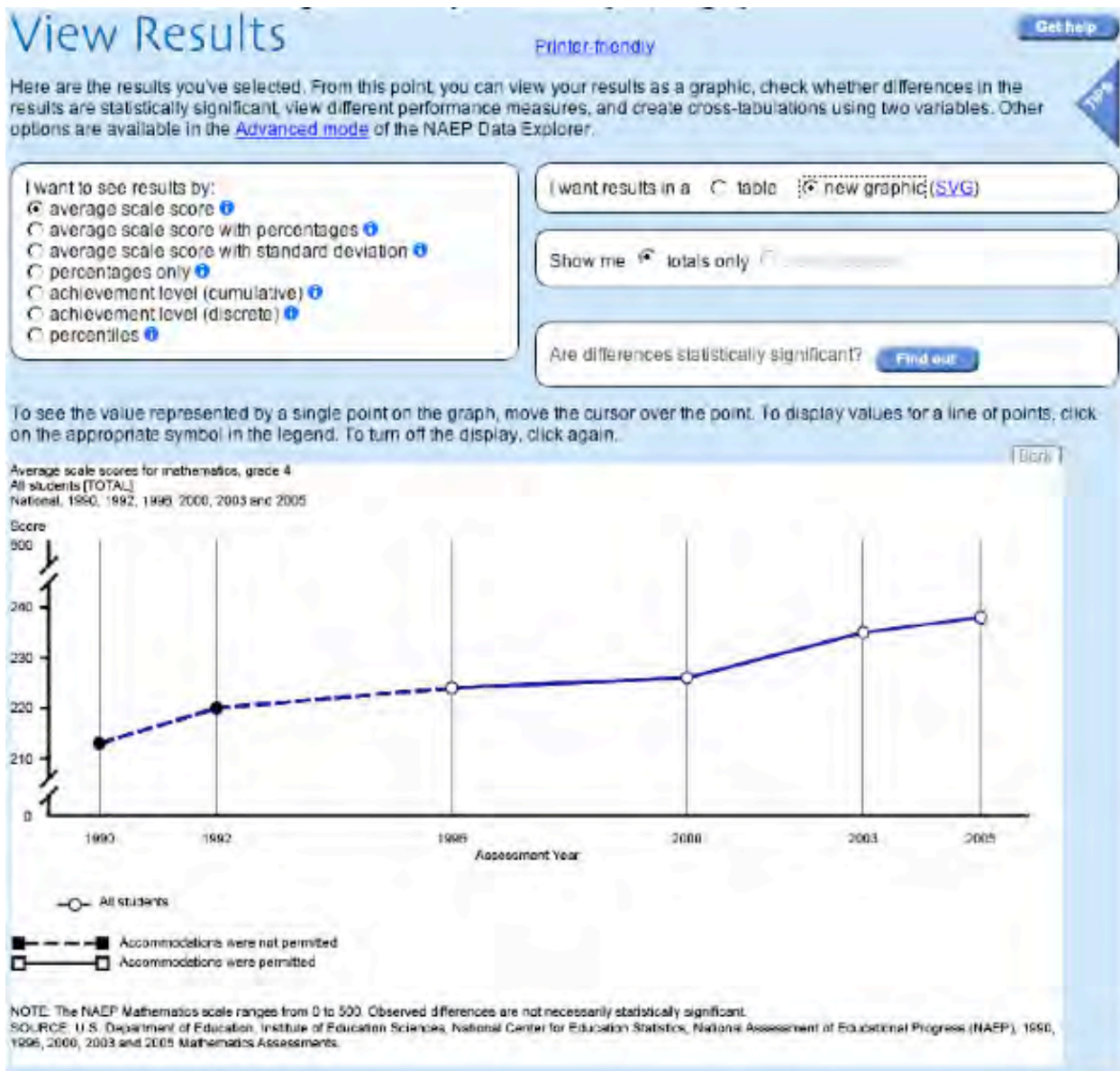
As is the case is carrying out significance testing, obtaining graphs of results involves the clicking on a link that opened a window external to the main NDE analysis window.

Figure B-8. Graph Creation Window



Under Graphic Options, users can choose to create the graph as a full graph or in a scroll window, but nowhere are these options explained. Users were confused by those options, and usually selected full graph by default (Figure B-9).

Figure B-9. Sample NDE Graph (full graph mode)



- *Exporting Results to Excel*
Functionality: No user had success with this, but all concurred it was an important and useful feature when it worked.
 - *Suggestion: More testing of this feature is needed, including with respect to compatibility across various operating system and browser combinations, and when detected, as necessary, a pop-up notifying of plug-ins needed should be implemented.*

Users typically were able to complete the steps required to download the Excel file, but the downloaded file was blank for all users. Trials by the researcher both before and after the observations were successful, but users indicated that this might be a source of frustration and negativity for some users on occasions when it did not work.

- *Advanced Mode*

Comment: The “formatting table” page contained many options, and users liked the flexibility. The option to report by subscale was particularly interesting to participants.

Comment: Users appreciated the legend in Advanced mode clarifying symbols used to indicate the ability to collapse categories and to reduce data shown.

Information: Users did not understand the counts of variables at the bottom of the page in Advanced mode.

- *Suggestion: The addition of an explanatory link or pop-up with clarifying information would be helpful.*

Information: Users appreciated the option to close or collapse categories in Advanced mode but could not undo the changes they made if they wanted to.

- *Suggestion: Add how-to text in a mouse-over or pop-up.*

Participants in this study were asked to carry out an analysis in Advanced mode and provide feedback on the ease of use, appearance, and overall functionality of this mode. Users were at first uncertain of the differences between the two modes (Quick Start and Advanced) but with use identified the features present in Advanced that are not available in Quick Start. Given the familiarity with Quick Start that most users developed, however, the navigation and general use of Advanced was regarded as satisfactory (Figure B-10).

Figure B-10. Analysis Selection Page, Advanced Mode

Switch to Quickstart Mode View Advanced Info Select Criteria Choose Year(s) Format Table Go to Results

Start over Get help

Select one or more options from each of the categories below. For more information about a category, click on the information symbol (i) next to its name.

Show options available for all assessments the latest assessments only

Grade: *i*

Grade 4
 Grade 8
 Grade 12

Subject: *i*

Civics
 Geography
 U.S. History
 Mathematics
 Reading
 Science
 Writing

Jurisdiction(s): *i*

National
 National Public
 State/Jurisdiction
 Urban District
 Region

Variable(s): *i*

Major Reporting Groups
 All Students (Overall Results)
 Gender
 Natl School Lunch Prog eligibility (3 categories)
 Public or nonpublic school (5 categories)
 Public or nonpublic school (7 categories) (2002+)
 Race/ethnicity used in NAEP reports after 2001
 Region of the country (2003 and later)
 School identified as charter (National Public)
 School location (2005)
 School location (9 categories) (2005)
 Student classified as having a disability
 Student is English Language Learner (2 categories)
 Student is English Language Learner (3 categories)

Student Factors
 Demographics

Show: all variables (1170)
 selected variables (1)
 search results

What next? Click "Choose Year(s)" to refine your criteria, or "Go to Results"

Switch to Quickstart Mode View Advanced Info Select Criteria Choose Year(s) Format Table Go to Results

[About NAEP Data Explorer](#) [Important Legal Information](#)

In Figure B-11 is shown the formatting table options in Advanced mode. Users here have the option to report by subscale using the drop-down menu, and can choose to include all cumulative achievement levels, or, by clicking on the red X's, can eliminate columns.

Figure B-11. Format Table, Advanced Mode

Selected criteria: Mathematics, Grade 4, National, All students, 1990¹, 1992¹, 1996, 2000, 2003, 2005

On this screen, you may format your results by selecting the performance measures you are interested in, choosing a subscale (if available), and specifying the order of the variables. You may also remove any performance measure categories.

I want to see results by:

- average scale score
- average scale score with percentages
- average scale score with standard deviation
- percentages only
- achievement level (cumulative)
- achievement level (discrete)
- percentiles

Subscale: **Mathematics**

Show me: totals only cross-tabulation

Show long titles Include standard error in parenthesis

Show empty cells Number precision:

Preview Window: This table will contain 6 rows.

All students	Year(s)	Jurisdiction(s)	<input checked="" type="checkbox"/> Below Basic	Standard Error	<input checked="" type="checkbox"/> At or above Basic	Standard Error	<input checked="" type="checkbox"/> At or above Proficient	Standard Error	<input checked="" type="checkbox"/> Advanced	Standard Error
All students	1990 ¹	National	nn	(n.n)	nn	(n.n)	nn	(n.n)	nn	(n.n)
	1992 ¹	National	nn	(n.n)	nn	(n.n)	nn	(n.n)	nn	(n.n)
	1996	National	nn	(n.n)	nn	(n.n)	nn	(n.n)	nn	(n.n)
	2000	National	nn	(n.n)	nn	(n.n)	nn	(n.n)	nn	(n.n)
	2003	National	nn	(n.n)	nn	(n.n)	nn	(n.n)	nn	(n.n)
	2005	National	nn	(n.n)	nn	(n.n)	nn	(n.n)	nn	(n.n)

Legend: Close column

What next? Click "Go to Results" to display your results.

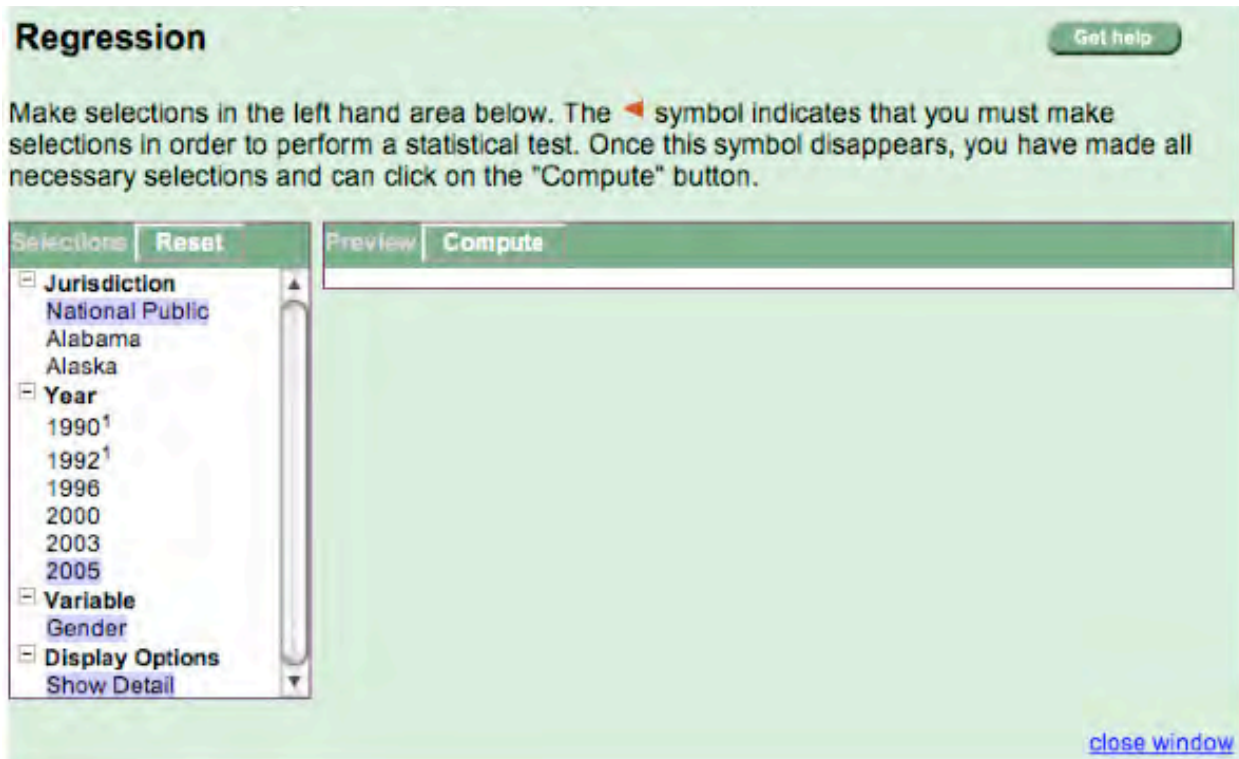
- **Regression**
 Comment: This was uniformly described as a great feature of the NDE.
 Appearance: Users wanted to be able to rearrange the tables. One user specifically mentioned that how the regression coefficients were ordered in the display of results was not typical, indicating that it was preferable to put the regular regression coefficients first, then the standardized ones.
 - *Suggestion: Reformat regression results tables.*
 Information: Two users commented that though they knew the tool is carrying out contrast coding to do the analysis, they wanted the analyses being done to be very explicit, more than currently in the footnote.
 - *Suggestion: Provide a direct hyperlink to further explanatory text.*

Functionality: When they clicked on “regression” users expected that to be a seamless process, not a pop-up with a series of decisions to be made. Users perceived this process to be redundant to the process of variable selection.

- *Suggestion: Users wanted the regression results embedded in the regular display of results, perhaps appended to the end of previous results.*

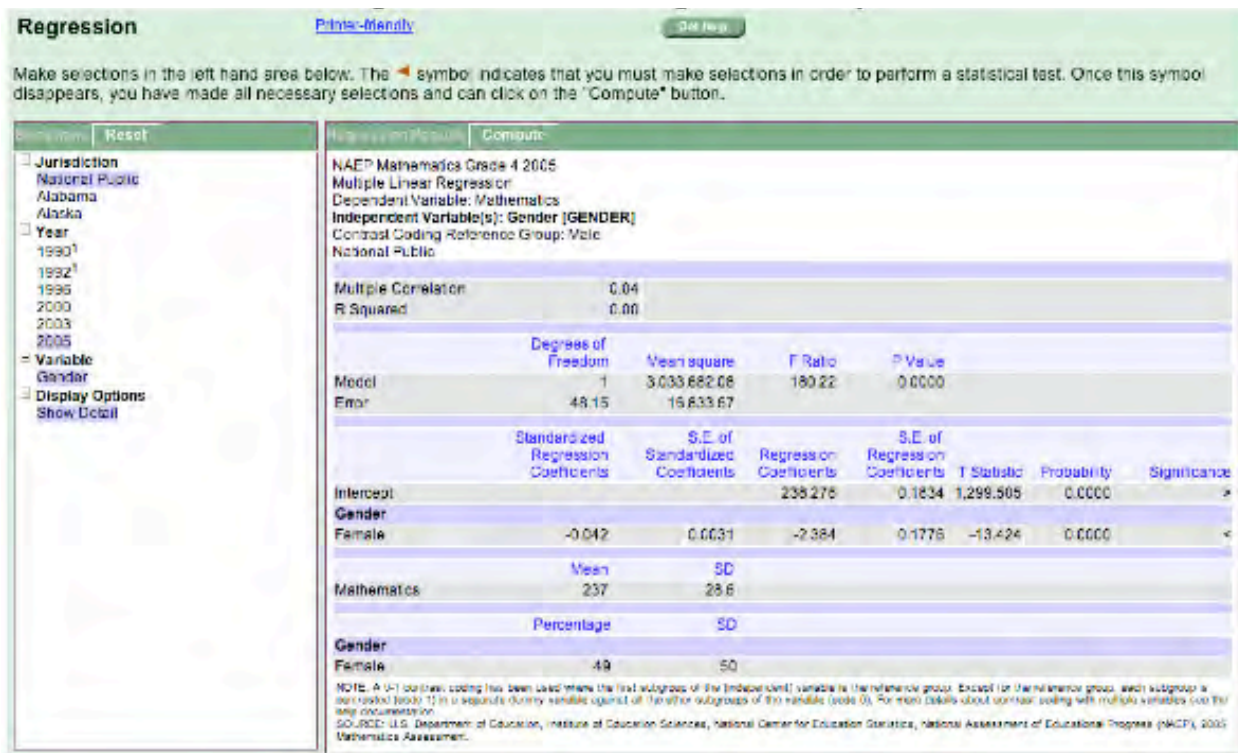
The option to carry out regression analyses within the structure of the NDE’s Advanced mode was cited as an important positive feature of the NDE (Figure B-12).

Figure B-12. Regression Option Screen, Advanced Mode



Most of the comments received about the Regression functionality were minor, including the suggestion to revise the order of the display of results and to add explanatory text to the display about how the analysis is being carried out (Figure B-13). Depending on the sophistication of the user, the current footnote may not be enough to ensure understanding of the data analysis.

Figure B-13. Results of Regression Analysis



Conclusions and Next Steps

The NAEP Data Explorer clearly offers visitors to the NAEP Web site a tremendous opportunity to run analyses and answer data questions that in many cases are idiosyncratic to the user and might not otherwise be easily answered through other means of NAEP score reporting. In addition, it permits users to customize results, obtain graphics, and test for statistical significance as needed. This is a unique resource among educational testing programs and represents another way in which NAEP is on the cutting edge of score reporting practices today.

The feedback from participants in the two sets of user observations reported here are remarkably consistent. Both groups reported a high level of overall satisfaction with the tool, finding it useful and somewhat simple to learn to use. As with any Web site or data analysis tool, there was a learning curve, but for the most part users did not encounter significant navigation or logical difficulties in carrying out analyses. When there were difficulties, these were identified as concerning the NDE’s *Appearance*, the *Information* displayed, and the NDE’s *Functionality*.

As to *Appearance*, users appreciated the “What’s Next?” bar on the Selection page. In some cases, users suggested minor redesigns of links and clickable icons. Similarly, certain aspects of the appearance of the variable selection screen and some result displays were found to be confusing to some users, and these should be followed up on with additional user groups to further determine if edits are warranted.

The participants here were most impressed with the volume of *Information* available for analyses and reported in the results screens of the NDE. When suggestions concerning the content of the pages within the tool were made, many of these involved additional contextual information (the range of the NAEP score scale, the text of the background variable questions, sample sizes). Other information feedback received has potential to impact on the users’ perceptions of the functionality of the NDE: when users saw some jurisdictions and variables

“grayed out” without explanation, they thought they perhaps had done something wrong, and in some cases wanted to start over. More, prominently placed explanation is needed, in this case. Users also wanted to know more about the analyses performed and the tests carried out, commenting that if they were to try and use these results in a professional setting (a technical report or a conference presentation) they would need to understand the statistics in full.

Concerning *Functionality*, users identified several possible modifications to the current operating infrastructure of the NDE. Comparing student performance across grades was one such use. In addition, the opening of a new browser window and the sequence of decisions involved in making graphs, testing for statistical significance, and carrying out regression analyses were considered somewhat cumbersome by participants. They wanted the process to be somewhat simplified and to occur within the main browser window. On a related point, the users requested that new results within a session be appended to the bottom of previous results. This is perhaps due to their familiarity with how data analysis software packages such as SPSS and SAS work: as users carry out new analyses output is organized sequentially within one output window (unless the user chooses to close an output window and open a new one), so this may be a different way of managing results for users.

The current findings are limited by the small samples but do provide considerable insight into the usability of the NDE. To the extent that all participants in the research were largely unfamiliar (i.e., not “expert”) in their use of the data explorer, this research is particularly informative with respect to learning about the user experience of beginning users. Another direction for future research involves the use of additional participants with different a) levels of statistical familiarity and b) needs for data manipulation tools such as the NDE, including additional policymakers. In particular, future research might consider identifying more veteran or frequent users of the NDE to obtain their feedback, due to the difference in perspective that such users would bring.

This page intentionally left blank

References

- De Mello, V. B. (2004). *NAEP state analysis project. Task 2.2. State Profile and Report Enhancement: Recommendations on State Web Profiles* (Contract ED-01-CO-0026/0019). Washington, D.C.: American Institutes for Research.
- Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scale and reports more understandable? In R. W. Lissitz and W. D. Schafer (eds.), *Assessment in educational reform* (pp. 192–205). Boston, Mass.: Allyn and Bacon.
- Hambleton, R. K., and Slater, S. C. (1995). Using performance standards to report national and state assessment data: Are the reports understandable and how can they be improved? In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments*, pp. 325–343.
- Jaeger, R. M. (1992). General issues in reporting of the NAEP trial state assessment results. In R. Glaser and R. Linn (eds.), *Assessing student achievement in the states* (pp. 107–109). Stanford, Calif.: National Academy of Education.
- Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress (Working Paper 2003-11)*. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.
- Koretz, D., and Deibert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991* (MR-385-NCES). Santa Monica, Calif.: RAND.
- Levine, R., Rathbun, A., Selden, R., and Davis, A. (1998). *NAEP's constituents: What do they want? Report of the National Assessment of Educational Progress Constituents Survey and Focus Groups*. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, NCES 98-521.
- Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, 11, 23–47.
- Linn, R. L. and Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement*, 29, 177–194.
- National Research Council. (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting*. P. J. DeVito and J. A. Koenig (eds.). Washington, D.C.: National Academy Press.
- Ogilvy Public Relations Worldwide. (2004). *NAEP: Reporting initial results. Analysis and recommendations for improvement*. Washington, D.C.: Author.
- Simmons, C., and Mwalimu, M. (2000). What NAEP's publics have to say. In M. L. Bourque and S. Byrd (eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 184–219). Washington, D.C.: National Assessment Governing Board.
- Wainer, H. (1996). Using trilinear plots for NAEP data. *Journal of Educational Measurement*, 33, 41–55.

- Wainer, H. (1997). Improving tabular displays: with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, 22, 1–30.
- Wainer, H. (2000b). Cholera, rocket ships, and Tom’s veggies: Contemporary and historical ideas toward the effective communication of school performance. *Evaluation and Research in Education*, 14, 148–180.
- Wainer, H., Hambleton, R.K., and Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335.

Appendix C: NAEP Web Site Usage: March 2005 to February 2006

April L. Zenisky and Polly Parker

*Center for Educational Assessment
University of Massachusetts Amherst*

Introduction

The National Assessment of Educational Progress (NAEP) uses a variety of strategies to report results to interested audiences, and increasingly, the information and resources on the NAEP sites on the World Wide Web (WWW) are becoming a main means by which interested audiences can access documents explaining the test results and also explore the data themselves. A tremendous amount of information about the NAEP testing program, including multiple years' worth of results for a number of content areas including core subjects such as Mathematics and Reading, is currently available on the main NAEP Web site (<http://nces.ed.gov/nationsreportcard/>). In addition, a recent initiative in NAEP reporting has involved the creation and support of a temporary *Initial Release Site*, denoted by the URL www.nationsreportcard.gov. This site becomes active when there is a major release of NAEP data, such as the 2005 Math and Reading Grade 4 and 8 results, the Trial Urban District Assessment (TUDA) results, and the 2005 Science Grade 4 and 8 results. It made "live" for a period of a few weeks subsequent to the release, and after that specified time period users who try to access that site are redirected to the main NAEP site. Between these two Web sites, users have the opportunity to access a considerable amount of information on NAEP, from executive summaries of results and interactive tools to more general information about the NAEP program and the content of the assessments.

In this paper, we analyze data obtained from the National Center for Education Statistics (NCES) regarding use of the NAEP Web site for a one-year period from March of 2005 to February of 2006, as collected by the government contractor Webtrends. This snapshot of site use is intended to briefly characterize the users of the NAEP Web site and explore how visitors to the NAEP site interact with the resources there, to the extent allowed by the data that can be collected. Ultimately, we hope to use this report to help develop understanding of the visitors to and users of the NAEP Web site as part of the larger evaluation of NAEP's score reporting efforts.

Brief Overview of NAEP on the Web

The NAEP testing program in 2005 involved 50 states and other jurisdictions, and thousands of students. Fourth-, eighth-, and twelfth-graders participated in the national assessment of reading, math, and science, fourth- and eighth-graders were also tested to provide state-level results in all three content areas, and 11 cities were involved in the Trial Urban District Assessment. As "The Nation's Report Card," NAEP is "the only nationally representative and continuing assessment of what America's students know and can do in various subject areas" (NCES, n.d.). Consumers interested in NAEP data can get access to a tremendous amount of information on the NAEP Web site. These resources fall into four main categories: programmatic Web pages, static data-oriented Web pages, interactive media tools, and downloadable PDFs of paper-based NAEP reports that have been released over the years. For the purposes of categorizing the resources on the site, we define these categories as follows.

- *Interactive media tools* are defined by a high degree of user choice in generating what results or analysis are called up to be displayed on a page: we refer here to the use of multimedia and clickable data resources which, for example, might allow users to manipulate the format (tables or graphs), information (scale scores, proficiency levels, percentiles), and type of results displayed (national, state, subgroups, gaps, etc.).
- In contrast, *static data-oriented Web pages* provide assessment findings in structured tables, charts, or text formats that Web sites users cannot manipulate.
- A number of pages on the NAEP site contain links to numerous *downloadable PDFs*, which package information in easy-to-print formats for user review, often in traditional technical report-style layouts with tables of contents.
- *Programmatic Web pages* are text-based resources accessible by branching off the main NAEP page which are explanatory in nature and do not contain assessment data or results.

Method

Data for this analysis were collected from March 1, 2006, to Feb.28, 2006. A Web site usage tracking firm (Webtrends) has been used by many government agencies, including the U.S. Department of Education, to obtain anonymous Web analytic data on the public's use of various government Web sites, subject to federal privacy regulations associated with data collection. Staff from the National Center for Educational Statistics obtained the Webtrends data files for the one-year period from March 2005 to February 2006 for the pages encompassed by the NAEP web site (<http://nces.ed.gov/nationsreportcard/>) and the initial release site (<http://www.nationsreportcard.gov>), and provided the information to UMass for analysis.

The focus in this analysis was largely on global usage statistics, such as the number of page views³³ and visits³⁴ to the NAEP site.³⁵ Other statistics of interest include the use of specific tools on the NAEP site and the most popular pages on the site.

Results

Operating systems and browsers. During February 2006, data were collected to identify the operating systems (also referred to as “platform”) and Internet browsers used by visitors to the National Center for Education Statistics site (NCES) by number of visits and views (this information was not available for the NAEP site separately). These results for the NCES site are shown in Table C-1. The top three most often used platforms were versions of the Windows operating system, including Windows XP (75.14 percent), Windows 2000 (9.25 percent) and Windows 98 (4.67 percent), while the Macintosh operating system was a close fourth, used by 4.61 percent of visitors. The top three most often used browsers among these users included Microsoft Internet Explorer (82.10 percent), Mozilla Firefox (9.45 percent) and Safari (2.41 percent). The percentage of usage is based upon the total number of visits. In both cases, the number one platform or browser used is higher than all the rest creating a significant drop from the number one to the number two.

³³ Generally defined as a request to load a single page of a Web site. (Source: <http://www.webtrends.com/Resources/WebAnalyticsGlossary.aspx#p>)

³⁴ A visit is an interaction an individual or unique visitor has with a Web site over a specified period of time or activity. (Source: <http://www.webtrends.com/Resources/WebAnalyticsGlossary.aspx#v>)

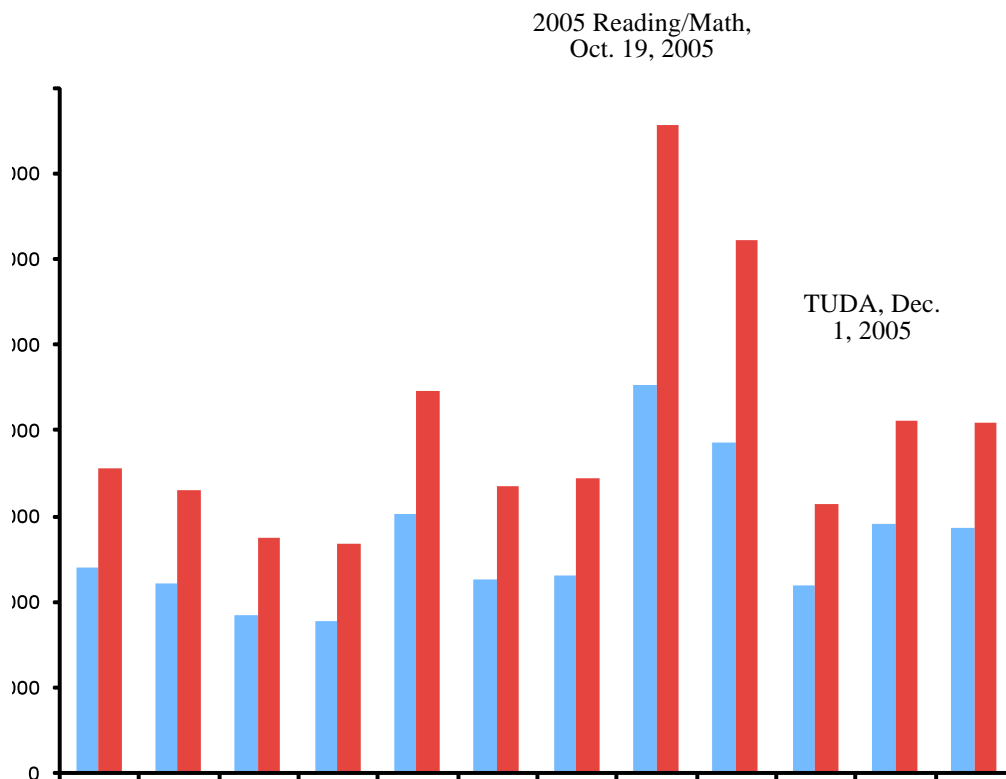
³⁵ Note that because a web site user may return and reload a page multiple times within a visit, the count of page views may be higher than page visits.

Table C-1. Frequency of visits and views to the NCES Web site by platforms and browser use

		Visits	%	Views
Platforms	Windows XP	1,285,703	75.14	13,699,677
	Windows 2000	158,291	9.25	1,676,178
	Windows 98	79,913	4.67	949,286
	Macintosh PowerPC	78,797	4.61	1,254,676
	Others	65,309	3.82	442,802
	Windows ME	21,750	1.27	244,525
	Windows 2003	6,796	0.40	75,113
	Linux	5,135	0.30	28,328
	Windows NT	4,600	0.27	40,872
	Windows 95	2,394	0.14	24,707
	Windows Win32s	1,560	0.09	900,296
	Macintosh	294	0.02	3,458
	SunOS	254	0.01	1,809
	FreeBSD	210	0.00	4,611
	NetBSD	14	0.00	87
	Macintosh 68K	6	0.00	108
	Hewlett Packard Unix (HP9000)	5	0.00	30
	OpenBSD	3	0.00	23
	Windows 3.x	2	0.00	5
OS/2	1	0.00	26	
Browsers	Microsoft Internet Explorer	1,404,772	82.10	16,655,473
	Mozilla	161,680	9.45	1,374,945
	Safari	42,145	2.46	628,246
	Others	28,172	1.65	110,625
	Opera	18,142	1.06	42,616
	Netscape	17,886	1.05	198,244
	Other Netscape Compatible	11,265	0.66	57,992
	OCP HRS 1.0	1,730	0.10	9,864
	AppleSyndication/49	1,363	0.08	3,150
	Moozilla	1,280	0.07	3,734
	RssReader/1.0.88.0	1,092	0.06	1,103
	Jakarta Commons- HttpClient/3.0-rc4	886	0.05	2,543
	Avant Browser	831	0.05	915
	Konqueror	800	0.05	6,177
	FeedFetcher-Google	622	0.04	921
	Linkscan/11.5 Unix	442	0.03	1,664
	Xenu Link Sleuth 1.2g	421	0.02	2,151
	NewsGatorOnline/2.0	404	0.02	404
	Ia_archiver	346	0.02	18,216
	Onfolio/2.02	328	0.02	1,868

Visits to the NAEP site by month. The next area of results analyzed involved overall visits to the NAEP site. These results may be observed in Figure C-1. Visits to the NAEP home page from March 2005 through February 2006 show the highest level of usage (as defined by the number of visits) during the months of November and October of 2005. This can be explained by the Oct. 19, 2005, release of the 2005 Mathematics and Reading results for grade 4 and 8, as well as the release of the Trial Urban District Assessment (TUDA) results in late November. The views throughout those visits were also proportionately increased. In this snapshot of the data, the results for March through June show that there were between 18,000 and 24,000 visits in each of those months, although the lowest hit rate for the NAEP Web site in the one-year period under consideration was observed during the months of May and June 2005, followed by a sharp increase during the month of July due to the release of the Long Term Trend data (*NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance*, July 14, 2005). The number of visits and views observed in August and September 2005 then returns to about the same levels of usage seen in March and April until October and November. The results for December are again consistent with previous months such as March, April, August, and September, with a subsequent slight increase observed for January and February 2006 (to about the level seen in July 2005).

Figure C-1. Frequency of Page Views and Visits to the NAEP Home page, March 2005—February 2006



The peak month for visits to and views of to the NAEP home page was October 2005, followed by November 2005: this coincides with the release of the 2005 Math and Reading results for Grades 4 and 8 as well as TUDA. The month of June 2005 showed the lowest number of visits or views. The overall percentage of usage for the NAEP home page in relation to other pages on the site for the months of March through June hovered between 40–44 percent. In months July through December 2005, the overall percentage usage of the NAEP home page increased between 46–56 percent. Even though the number of visits during January and February 2006 were higher than some of those during the July through December range, the percentage of usage lowered a bit to 42–44 percent. December 2005 indicates increased activity prior to January and February 2006 for entries to the NAEP pages at 49.95 percent, but is still slightly lower than November 2005 which is 52.66 percent, and even lower than October 2005 at 55.19 percent. The highest month for percentage of usage for the NAEP home page was July at 55.94 percent. The NAEP home page was consistently ranked 11th at the NCES entry page throughout the months between December and February.

By way of context for the overall level of traffic on the NAEP Web site, an executive summary of NCES Web Usage for April 2005 showed 12,988,965 page views and 1,230,955 visits for all of NCES in that month: with 33,089 page views and 22,229 visits for NAEP that month, NAEP accounted for a very small fraction of the NCES traffic (0.003 percent and 0.018 percent, for page views and visits, respectively).

Table C-2 presents the most commonly accessed pages within the NAEP site across the twelve months considered in this report, culled from the top ten pages for each month. It may be clearly noted that the NAEP home page received far greater traffic than any of the other pages within the site. The results are interesting in that the most frequently viewed Web pages through the NAEP Web site are the NAEP home page (NAEP—The Nation’s Report Card—National Assessment of Educational Progress; <http://nces.ed.gov/nationsreportcard/>), and the page which serves as a portal with links to individual state profiles (NAEP—State Profiles. Educational Assessments by State, Student demographics: <http://nces.ed.gov/nationsreportcard/states/>). A breakdown of this information by month is provided in Appendix C-1.

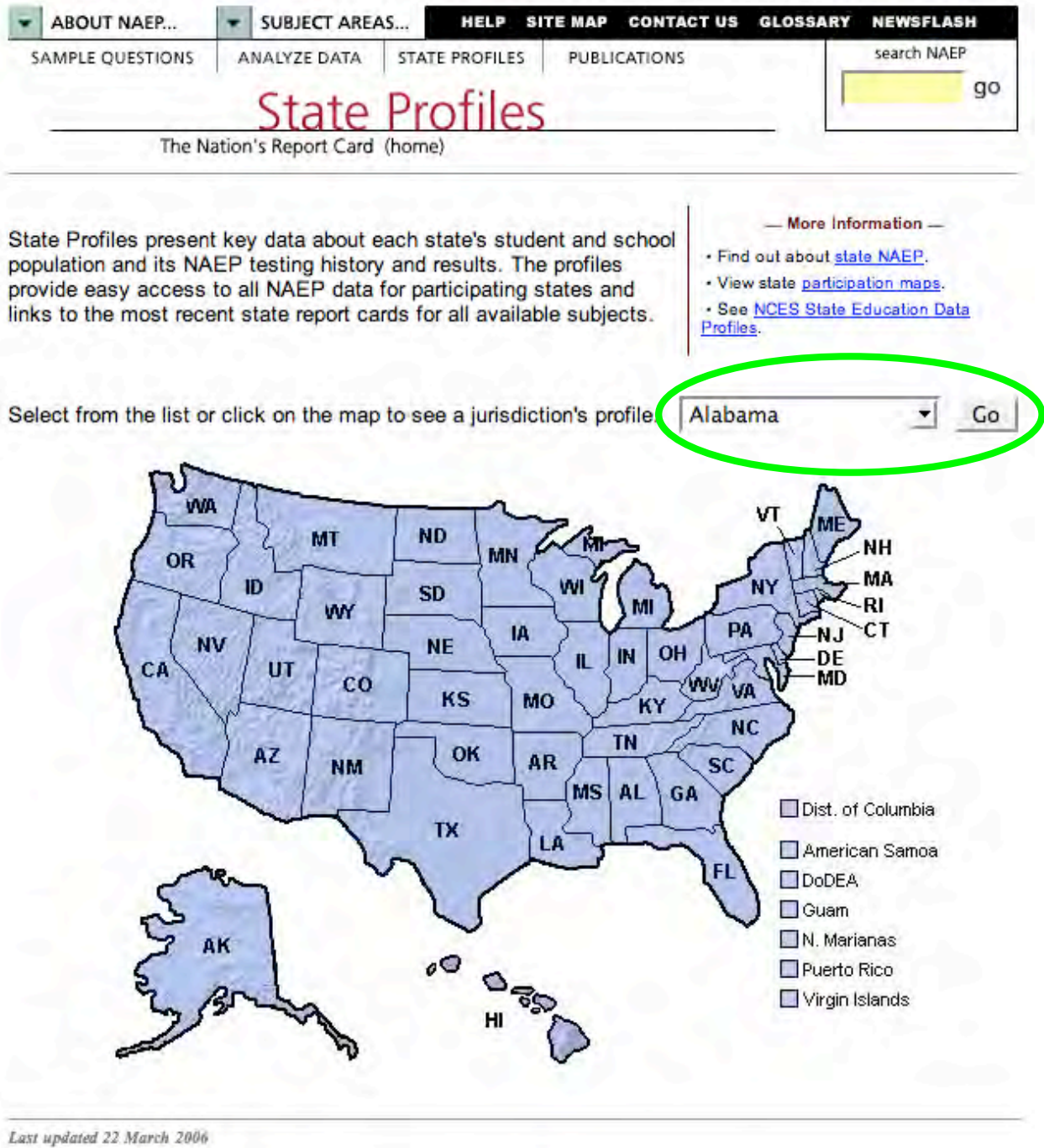
Table C-2. Most popular pages on the NAEP site (March 2005 to Feb. 2006)

Page	Visits	Views
The Nation's Report Card—National Assessment of Educational Progress—NAEP http://nces.ed.gov/nationsreportcard/	322,541	486,936
NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	94,190	125,588
NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	91,640	232,200
NAEP NQT v2.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp	63,536	278,670
NAEP—Released NAEP Questions for math, reading, science, writing, more http://nces.ed.gov/nationsreportcard/itmrls/	59,156	99,977
NAEP NQT v2.0—Question http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	39,678	443,314
NAEP—Mathematics. Scheduled NAEP reading assessments, past results, trends http://nces.ed.gov/nationsreportcard/mathematics/	38,595	52,355
Script for Initial Release Site http://nces.ed.gov/nationsreportcard/nrc/scripts/irscripts.vbs	33,380	0
NAEP Question Tool 3.0 http://nces.ed.gov/nationsreportcard/itmrls/	29,141	39,307
SVG Browser check http://nces.ed.gov/nationsreportcard/nrc/scripts/svgcheck.vbs	28,435	0
NAEP—Reading. Scheduled NAEP reading assessments, past results, trends http://nces.ed.gov/nationsreportcard/reading/	27,458	38,187
NAEP—2004 results [Long Term Trend] http://nces.ed.gov/nationsreportcard/ltr/results2004/	19,694	24,700
What Is NAEP? http://nces.ed.gov/nationsreportcard/about/	19,483	23,212
NAEP Data Explorer [Main Page] http://nces.ed.gov/nationsreportcard/nde/	13,953	17,645
NAEP Reading Mathematics 2005—Reading: Summary http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/s0002.a sp	8,051	9,981
NAEP Reading Mathematics 2005—Mathematics: Summary http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/s0017.a sp	7,871	9,559
NAEP - The Nation's Report Card: 2005 Reading and Mathematics http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/	6,262	7,346

The Question Tool was accessed quite frequently, as was the NAEP Data Tool (prior to October) and the NAEP Data Explorer (October and forward). Overall, it seems significant to note that the state results and state profile access page were among the most popular pages on the NAEP site throughout the year.

Also, though it may appear that the Alabama state profile page in many months is a very commonly accessed page, this seems to be a result of how the tool works and how usage data are collected and organized, rather than a disproportionate interest in NAEP performance among Yellowhammer State residents. When users access a state profile using the interface shown in Figure C-2, they use a drop-down menu to select their state or jurisdiction of interest, but the Web architecture behind the access page shown in Figure C-2 works in such way that although the Web usage statistics may make it appear that Alabama's profile is accessed quite often, in actuality that URL represents all state profiles accessed in a given month. It does not seem possible to break out results for individual states.

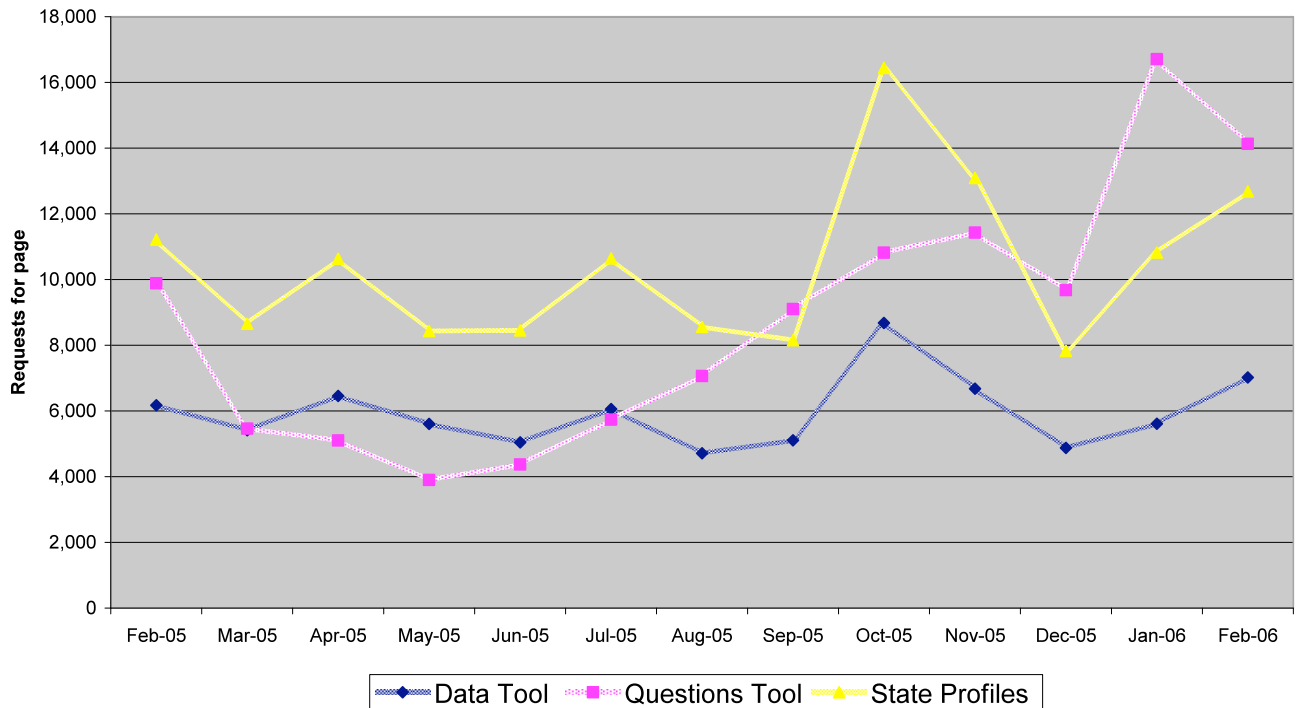
Figure C-2. Screen capture of the State Profile access page, with Alabama and Go button highlighted



The following chart (Figure C-3) shows the NAEP Web site tool usage frequency from March 2005 through February 2006. Three tools have been plotted: the NAEP Data Tool (reintroduced as the NAEP Data Explorer in October 2005), the NAEP Question Tool, and the state profiles. From February to September 2005, the State Profiles received the most attention of the three from visitors, averaging about 9,000 uses (as entry pages) per month. Use of the Data Tool was consistent at about 5,000 to 6,000 uses (as entry pages) per month, and the

Question Tool exhibited high use (10,000 uses as entry pages) in February 2005, followed by a decline to a low of 4,000 uses in July 2005 and then a steady increase in use. Beginning in October 2005, these patterns generally continued: the State Profiles are most used, while the Question Tool continues to grow in popularity and use of the NAEP Data Tool or Explorer remains relatively steady, by month.

Figure C-3. Use of the NAEP Data Tool, the Question Tool, and the State Profiles³⁶



Entry pages. Also of interest are statistics identifying the first page that users of the NAEP Web site encounter during a visit to the site. Table C-3 contains information that ranks pages on the NAEP Web site based on how many times each page served as the first page a visitor viewed when he or she accessed the NAEP site across the year’s worth of data. In this case, a visitor might have bookmarked a particular page that they use often for easy access, have typed a particular URL into their browser’s address bar, or have come to the NAEP site via a search engine such as Google, Yahoo, or MSN Search, among many others. When people visit the NAEP Web site, the most common of these entry pages is the NAEP home page. Other common first pages are the NAEP Questions Tool v3.0, and the NAEP State Profiles page. A breakdown of this data by month is provided in Appendix C-2.

³⁶ Graphic reprinted from Web site usage documents provided by NCES, March 2006.

Table C-3. Most common entry pages for the NAEP site

Page	Visits
NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	211,054
NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	26,132
NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	16,831
NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	5,149
NAEP—Released Questions for math, reading, science, writing and more. http://nces.ed.gov/nationsreportcard/ITMRLS/	4,059
Object Moved http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	3,308
NAEP High School Transcripts—How is Grade Point Calculated? http://nces.ed.gov/nationsreportcard/hsts/howgpa.asp	3,092
NAEP NQT v2.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp	1,532
NAEP NQT v2.0—Question http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	1,242
NAEP Reading Subject Area http://nces.ed.gov/nationsreportcard/reading	840

Initial release site usage. Beginning Oct. 19, 2006, the initial release site for NAEP results was live at the Web address <http://www.nationsreportcard.gov>. While the contents of that site change depending on what results are released, some reflection on commonly viewed and visited pages is helpful for evaluating what information users are interested in. The most commonly accessed pages on this Initial Release Site for October 2005 through February 2006 are listed in Table C-4, and a breakdown of this information by month is in Appendix C-3.

Table C-4. Most commonly loaded pages on <http://www.nationsreportcard.gov>, Oct. 2005–Feb. 2006

Page	Visits	Views
[Redirect page to current Initial Release site home page] http://nationsreportcard.gov/	42,029	48,454
The Nation's Report Card http://nationsreportcard.gov/reading_math_2005/	40,743	51,724
NAEP Reading Mathematics 2005—Reading: State Results: State Achievement Levels http://nationsreportcard.gov/reading_math_2005/s0006.asp	22,551	65,461
NAEP Reading Mathematics 2005—Mathematics: State Results: State Achievement http://nationsreportcard.gov/reading_math_2005/s0021.asp	14,209	39,739
NAEP - TUDA Reading Mathematics 2005 http://nationsreportcard.gov/tuda_reading_mathematics_2005/	9,136	16,225
NAEP Reading Mathematics 2005—Reading: National Results: Average Scale Score http://nationsreportcard.gov/reading_math_2005/s0003.asp	8,695	13,510
NAEP Reading Mathematics 2005—Reading: Summary http://nationsreportcard.gov/reading_math_2005/s0002.asp	8,132	10,362
NAEP Reading Mathematics 2005 - Mathematics: National Results: Average Scale http://nationsreportcard.gov/reading_math_2005/s0018.asp	6,696	10,174
NAEP Reading Mathematics 2005—Mathematics: Summary http://nationsreportcard.gov/reading_math_2005/s0017.asp	4,923	5,988
NAEP Reading Mathematics 2005—Reading: Student Group Results: Race/ Ethnicity http://nationsreportcard.gov/reading_math_2005/s0011.asp	3,426	6,116
NAEP Reading Mathematics 2005—Downloads and Tools http://nationsreportcard.gov/reading_math_2005/s0046.asp	3,741	4,581
NAEP—TUDA Reading Mathematics 2005: Reading Results by Race/ Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0006.asp	2,622	6,638
NAEP—TUDA Reading Mathematics 2005: Reading District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0012.asp	2,305	6,466
NAEP—TUDA Reading Mathematics 2005: Reading Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0003.asp	2,238	3,309
NAEP—TUDA Reading Mathematics 2005: Reading Summary http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0002.asp	1,928	2,376
NAEP—TUDA Reading Mathematics 2005: Mathematics Results by Race/ Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0019.asp	1,833	4,500
NAEP—TUDA Reading Mathematics 2005: Reading Scale Score Trends http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0004.asp	1,400	2,091

Continues next page

Table C-4. Most commonly loaded pages on <http://www.nationsreportcard.gov>, Oct. 2005–Feb. 2006 (Continued)

Page	Visits	Views
NAEP—TUDA Reading Mathematics 2005: Mathematics Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0016.asp	1,383	2,008
TUDA Reading Mathematics 2005: Mathematics District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0025.asp	816	2,049
Document Moved http://nationsreportcard.gov/reading_math_2005/s0007.asp	369	1,187
NAEP—TUDA Reading Mathematics 2005: Mathematics Sample Questions http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0026.asp	130	151
NAEP—TUDA Reading Mathematics 2005: Mathematics Results By English Language http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0024.asp	164	245
NAEP—TUDA Reading Mathematics 2005: Reading Results by English Language Learners http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0011.asp	136	249

Overall, the primary pages accessed on the Initial Release site include the main page the 2005 Reading and Math results, as well as the achievement level results for both content areas. On the TUDA side, the main TUDA page garnered the most traffic, followed by various pages of Reading results.

A closer look at the data for the Initial Release Site from the months of October 2005 through February 2006, as shown in the monthly breakdown in Appendix C of this chapter, reveals the following results. Beginning in October 2005, the number of visits/views spiked on or about Oct. 19, which corresponds to the date of the release of the national and state math and reading results to the NCR site. Also for this month, the data showed interests in subgroups performances for race or ethnicity and gender. There was also considerable interest for viewing sample questions in both reading and mathematics. During the month of November, the Initial Release Site was active for the 2005 Math and Reading results through Nov. 13, and the results of particular interest were the state-level results in both Math and Reading and the average scale score results in both subjects as well.

During the month of December, the TUDA math and reading results appeared on the Initial Release Site beginning Dec. 1. Specific pages of interest for both reading and mathematics include results by race or ethnicity (third and eighth most popular for the month, respectively), district comparisons, and scale score trends. In January and February 2006, the number of visits to the Initial Release Site decreased significantly. The most often visited pages were the TUDA math and reading district comparisons, and visitors seemed to be interested in specific subgroup results such as English language learners (February). Again, it is important to note that both of these months had far fewer visits than the previous months.

Conclusions

Overall, this brief review of Web site usage statistics from the NAEP and NCES sites suggests several interesting findings with implications for future studies of the NAEP Web site. First, given that just over 89 percent of visitors to the NCES Web site are using a computer running a Microsoft Windows XP, 2000, or 98 operating system and are using Microsoft Internet Explorer as their Web browser, future research observing user behavior should consider those characteristics of the user population. At the same time, about 4.5 percent of visitors use a Mac, and over 9 percent surf the Web using the Mozilla Firefox browser. This is also an important issue for NCES and NAEP as they maintain and improve on the functionality of the NAEP site and the tools found there, so that users of alternative (non-Microsoft) products have a satisfactory user experience and can access information as needed.

The good news for NAEP is that during these 12 months the home page was viewed nearly half a million times, and visits numbered well over 300,000. In that time, there were several major data releases, and some new interactive tools were released or updated. Overall user traffic on the NAEP site increases in months when there is a release, as might be expected. At the same time, during the one-year period of data examined here, there seems to be a relatively consistent level of interest among users in the Web tools of the Question Tool, the Data Tool or Data Explorer, and the State Profiles. The NAEP Data Explorer was only available for five of the twelve months considered here, and in that time the main NDE page was viewed more than 17,000 times.

It is important to note that in several of the areas of Web traffic considered here, users exhibited a consistently high level of interest in the state profiles. By this data, individual state profiles were accessed over 230,000 times between March 2005 and February 2006. This suggests that future studies of the NAEP Web site should evaluate the extent to which the state profile pages are user-friendly and meet the information needs of visitors.

As noted previously, the Web usage statistics reported here indicate considerable interest in the interactive online tools available on the NAEP site (the Question Tool, the Data Tool Data Explorer, and the State Profiles), with growth in the use of the Question Tool. Among the Initial Release Site results for the five months for which it was active in the year considered here, state results were again among the pages most accessed, as were student group results. For the TUDA results (December, January, and February), district comparisons and results by race and ethnicity were popular among visitors to the site.

This review of several dimensions of Web usage statistics for the NAEP Web site is illuminating in several respects, as it provides a broad summary of the kinds of pages and information that visitors to the *The Nation's Report Card* Web site seek out. At the same time, in reviewing data collected about visitors to any Web site, caution must be taken not to over-interpret results, particularly with respect to the one-year snapshot of use presented here. First, this Web site, like most, is an evolving entity that is constantly maintained and updated. Between March 2005 and February 2006, several new features were added and older ones were completely revamped, and several major assessment results were released, so that there is a something of an ebb and flow to the counts of visits and views to the site month to month. At best, reviewing data from a single year provides a general pattern to use. In addition, as the NAEP site is a U.S. government site, there are data collection limitations that require a high level of anonymity and aggregation of the results.

Even given such limitations, this review of NAEP site usage absolutely identifies a number of key directions for further research on use of the NAEP site. For example, given the high volume of use of the State Profiles and the Question Tool, these represent aspects of the site that users might be observed using with respect to both navigational ease and content. The high level of interest in the results for the states and for student subgroups likewise suggests areas for further study. What information on those pages are visitors focusing on? In addition, efforts might turn to the Initial Release Site and obtaining user impressions of that page. Lastly, use of the interactive NAEP Data Explorer tool seems to be growing, and it will certainly be informative to learn more about how users work with the tool and for what purposes. Ultimately,

developing an understanding of what pages and information are of interest to the aggregate of visitors to the NAEP site (as provided in this report) has much practical value for future utility study activities, such as observations of individual users navigating the Web site and focus groups convened to discuss aspects of the site.

References

National Center for Education Statistics. (n.d.). *NAEP—Overview*. Retrieved Jan. 18, 2006, from <http://nces.ed.gov/nationsreportcard/about/>.

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)

Month	Page	Visits	Views
March 2005	The Nation's Report Card—National Assessment of Educational Progress—NAEP http://nces.ed.gov/nationsreportcard/	24,124	35,647
	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	7,570	10,171
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	7,469	20,046
	NAEP NQT v2.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp	5,926	27,244
	NAEP NQT v2.0—Question http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	5,112	45,933
	NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/	4,724	6,325
	NAEP—Released NAEP Questions for math, reading, science, writing, more. http://nces.ed.gov/nationsreportcard/itmrls/	4,582	6,322
	NAEP Reading Subject Area http://nces.ed.gov/nationsreportcard/reading/	3,827	5,463
	NAEP Mathematics Subject Area http://nces.ed.gov/nationsreportcard/mathematics/	3,585	4,788
	What Is NAEP? http://nces.ed.gov/nationsreportcard/about/	3,209	3,789
April 2005	The Nation's Report Card—National Assessment of Educational Progress—NAEP http://nces.ed.gov/nationsreportcard/	22,229	33,089
	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	7,655	10,634
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	7,495	20,341
	NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/	4,789	6,455
	NAEP NQT v2.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp	4,416	19,578
	NAEP NQT v2.0—Question http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	4,215	35,757
	NAEP Reading Subject Area http://nces.ed.gov/nationsreportcard/reading/	3,644	5,247
	NAEP—Released NAEP Questions for math, reading, science, writing, more. http://nces.ed.gov/nationsreportcard/itmrls/	3,640	5,099
	NAEP Mathematics Subject Area http://nces.ed.gov/nationsreportcard/mathematics/	3,367	4,555

Continues next page

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)
(Continued)

Month	Page	Visits	Views	
April 2005	What Is NAEP? http://nces.ed.gov/nationsreportcard/about/	3,006	3,580	
May 2005	The Nation's Report Card—National Assessment of Educational Progress—NAEP http://nces.ed.gov/nationsreportcard/	18,507	27,457	
	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	6,143	8,434	
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	6,070	16,162	
	NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/	4,065	5,606	
	NAEP NQT v2.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp	3,162	14,619	
	NAEP NQT v2.0—Question http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	3,013	26,932	
	NAEP Reading Subject Area http://nces.ed.gov/nationsreportcard/reading/	2,865	3,939	
	NAEP—Released NAEP Questions for math, reading, science, writing, more. http://nces.ed.gov/nationsreportcard/itmrls/	2,852	3,898	
	NAEP Mathematics Subject Area http://nces.ed.gov/nationsreportcard/mathematics/	2,640	3,463	
	NAEP—Overview http://nces.ed.gov/nationsreportcard/about/	2,514	2,975	
	June 2005	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	17,861	26,853
		NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	6,050	8,451
NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp		5,912	16,050	
NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/		3,682	5,045	
NAEP NQT 3.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp		3,266	15,754	
[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp		3,136	28,314	
NAEP NQT 3.0 http://nces.ed.gov/nationsreportcard/itmrls/		3,017	4,372	
NAEP Reading Subject Area http://nces.ed.gov/nationsreportcard/reading/		2,950	4,131	

Continues next page

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)
(Continued)

June 2005	NAEP—Overview http://nces.ed.gov/nationsreportcard/about/	2,710	3,288
	NAEP Mathematics Subject Area http://nces.ed.gov/nationsreportcard/mathematics/	2,669	3,646
July 2005	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	30,359	44,658
	NAEP—2004 Results http://nces.ed.gov/nationsreportcard/ltt/results2004/	8,747	11,306
	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	7,901	10,644
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	7,679	19,695
	NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/	4,471	6,060
	NAEP NQT 3.0 http://nces.ed.gov/nationsreportcard/itmrls/	3,511	5,737
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/itmrls/NQT_ItemDisplay.asp	3,383	37,219
	NAEP—Overview http://nces.ed.gov/nationsreportcard/about/	3,243	3,788
	NAEP—Reading. Scheduled NAEP reading assessments, past results, trends http://nces.ed.gov/nationsreportcard/reading/	3,133	4,256
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/ITMRLS/NQT_Search.asp	3,074	14,554
	Aug. 2005	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	22,576
NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/		6,444	8,559
NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp		6,271	15,213
NAEP—2004 Results http://nces.ed.gov/nationsreportcard/ltt/results2004/		5,713	6,988
NAEP NQT 3.0 http://nces.ed.gov/nationsreportcard/itmrls/		4,026	7,062
[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/ITMRLS/NQT_Search.asp		3,660	16,730
[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/itmrls/NQT_ItemDisplay.asp		3,635	38,316
NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/		3,512	4,712
NAEP—Mathematics. Scheduled NAEP math assessments, past results, trends http://nces.ed.gov/nationsreportcard/mathematics/		2,977	3,897

Continues next page

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)
(Continued)

Aug. 2005	NAEP—Reading. Scheduled NAEP reading assessments, past results, trends http://nces.ed.gov/nationsreportcard/reading/	2,753	3,823
Sept. 2005	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	23,121	34,440
	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	6,275	8,152
	NAEP— State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	6,191	15,670
	NAEP—2004 Results http://nces.ed.gov/nationsreportcard/ltt/results2004/	5,234	6,406
	NAEP NQT 3.0 http://nces.ed.gov/nationsreportcard/itmrls/	4,914	9,099
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/ITMRLS/NQT_Search.asp	4,842	22,375
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/itmrls/NQT_ItemDisplay.asp	4,736	48,933
	NAEP Data Tool v3.0—Introduction http://nces.ed.gov/nationsreportcard/naepdata/	3,898	5,104
	NAEP—Mathematics. Scheduled NAEP math assessments, past results, trends http://nces.ed.gov/nationsreportcard/mathematics/	3,607	4,934
	NAEP—Reading. Scheduled NAEP reading assessments, past results, trends http://nces.ed.gov/nationsreportcard/reading/	3,479	4,821
	Oct. 2005	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	45,431
NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/		12,817	16,442
NAEP— State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp		12,182	30,489
NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls/		6,448	10,816
NAEP—Mathematics. Scheduled NAEP math assessments, past results, trends http://nces.ed.gov/nationsreportcard/mathematics/		6,224	8,269
NAEP—Scheduled Reading Assessments, Past Results, Trends, Methods http://nces.ed.gov/nationsreportcard/reading/		4,807	6,507

Continues next page

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)
(Continued)

Oct. 2005	NAEP—Overview http://nces.ed.gov/nationsreportcard/about/	4,801	5,792
	NAEP Data Explorer http://nces.ed.gov/nationsreportcard/nde/	4,472	5,751
	NAEP Data Explorer http://nces.ed.gov/nationsreportcard/nde/criteria.asp	4,012	15,578
	NAEP Data Explorer http://nces.ed.gov/nationsreportcard/nde/viewresults.asp	3,668	14,146
	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/ [Browser check page, related to redirect to Initial Release site] http://nces.ed.gov/nationsreportcard/nrc/scripts/irscripts.vbs [Browser check page, related to Initial Release and SVG Viewer check] http://nces.ed.gov/nationsreportcard/nrc/scripts/svgcheck.vbs	38,641	62,141
Nov. 2005	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	15,400	0
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	15,348	0
	NAEP Reading Mathematics 2005—Reading: Summary http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/s0002.asp	9,937	13,092
	NAEP Reading Mathematics 2005—Mathematics: Summary http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/s0017.asp	9,875	25,839
	NAEP NQT v3.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/searchresults.asp	8,051	9,981
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls/	7,871	9,559
	NAEP—The Nation's Report Card: 2005 Reading and Mathematics http://nces.ed.gov/nationsreportcard/nrc/reading_math_2005/	6,679	33,676
		6,483	11,427
		6,262	7,346
Dec. 2005	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	21,906	31,468
	NAEP— State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	6,025	7,825
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	5,644	13,416
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls/	5,326	9,678

Continues next page

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)
(Continued)

Dec. 2005	NAEP NQT v3.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/searchresults.asp	4,888	23,901
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/itmrls/itemdisplay.asp	4,298	61,086
	[Browser check page, related to redirect to Initial Release site] http://nces.ed.gov/nationsreportcard/nrc/scripts/irscripts.vbs	4,232	0
	[Browser check page, related to Initial Release and SVG Viewer check] http://nces.ed.gov/nationsreportcard/nrc/scripts/svgcheck.vbs	4,220	0
	NAEP Data Explorer http://nces.ed.gov/nationsreportcard/nde/	3,890	4,874
	NAEP—Mathematics. Scheduled NAEP math assessments, past results, trends http://nces.ed.gov/nationsreportcard/mathematics/	3,619	4,954
	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	29,089	41,156
Jan. 2006	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls/	9,173	16,709
	NAEP NQT v3.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/searchresults.asp	9,028	41,091
	[Browser check page, related to redirect to Initial Release site] http://nces.ed.gov/nationsreportcard/nrc/scripts/irscripts.vbs	8,895	0
	[Browser check page, related to Initial Release and SVG Viewer check] http://nces.ed.gov/nationsreportcard/nrc/scripts/svgcheck.vbs	8,867	0
	NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/	8,144	10,821
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	8,095	19,044
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/ITMRLS/itemdisplay.asp	7,553	94,593
	NAEP NQT v3.0—Search Options page http://nces.ed.gov/nationsreportcard/itmrls/startsearch.asp	6,255	8,743
	NAEP—Mathematics. Scheduled NAEP math assessments, past results, trends http://nces.ed.gov/nationsreportcard/mathematics/	4,956	7,173
	Feb. 2006	NAEP—The Nation's Report Card—National Assessment of Educational Progress http://nces.ed.gov/nationsreportcard/	28,697
NAEP—State Profiles. Educational Assessments by State. Student demographics. http://nces.ed.gov/nationsreportcard/states/		9,229	12,363

Continues next page

Appendix C-1. Most popular pages on the NAEP site, by month (March 2005 to February 2006)
(Continued)

Feb. 2006	NAEP— State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	8,757	20,235
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls/	8,201	14,130
	NAEP NQT v3.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/searchresults.asp	8,147	38,332
	[Unknown page, related to Question Tool] http://nces.ed.gov/nationsreportcard/ITMRLS/itemdisplay.asp	7,116	91,764
	NAEP NQT v3.0— Search Option page http://nces.ed.gov/nationsreportcard/itmrls/startsearch.asp	5,635	8,028
	NAEP Data Explorer http://nces.ed.gov/nationsreportcard/nde/	5,591	7,020
	NAEP—Mathematics. Scheduled NAEP math assessments, past results, trends, me http://nces.ed.gov/nationsreportcard/mathematics/	4,951	6,676
	[Browser check page, related to redirect to Initial Release site] http://nces.ed.gov/nationsreportcard/nrc/scripts/irscripts.vbs	4,853	0

Appendix C-2. Most common entry pages for the NAEP site by month

Month	Page	Visits	%
March 2005	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	15,346	42.17
	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	2,017	5.54
	NAEP—Released Questions for math, reading, science, writing and more. http://nces.ed.gov/nationsreportcard/ITMRLS/	1,715	4.71
	NAEP NQT v2.0—Search Results http://nces.ed.gov/nationsreportcard/itmrls/search.asp	1,532	4.21
	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	13,475	40.33
April 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	2,155	6.45
	NAEP—Released Questions for math, reading, science, writing and more. http://nces.ed.gov/nationsreportcard/ITMRLS/	1,431	4.28
	NAEP NQT v2.0—Question http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	1,242	3.72
	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	11,144	41.8
May 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	1,686	6.32
	NAEP- The Nation’s Report Card http://nces.ed.gov/nationsreportcard	1,098	4.12
	NAEP—Released Questions for math, reading, science, writing and more. http://nces.ed.gov/nationsreportcard/ITMRLS/	913	3.42
	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	11,056	43.92
June 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	1,696	6.74
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	864	3.43
	NAEP Reading Subject Area http://nces.ed.gov/nationsreportcard/reading	840	3.34
	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	18,300	55.94
July 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	2,209	6.75
	Object Moved http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	1,108	3.39

Continues next page

Appendix C-2. Most common entry pages for the NAEP site by month (Continued)

Month	Page	Visits	%
July 2005	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	881	2.69
	NAEP—The Nation's Report Card http://nces.ed.gov/nationsreportcard	13,757	47.47
Aug. 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	1,982	6.84
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	1,419	4.90
	Object Moved http://nces.ed.gov/nationsreportcard/itmrls/qtab.asp	1,018	3.51
	NAEP—The Nation's Report Card http://nces.ed.gov/nationsreportcard	15,543	45.97
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	1,852	5.48
Sept. 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	1,618	4.79
	Object Moved http://nces.ed.gov/itmrls/qtab.asp	1,182	3.5
	NAEP—The Nation's Report Card http://nces.ed.gov/nationsreportcard	29,874	55.19
	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	3,924	7.25
Oct. 2005	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	2,487	4.59
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	1,305	2.41
	NAEP—The Nation's Report Card http://nces.ed.gov/nationsreportcard	25,741	52.66
	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	2,476	5.07
Nov. 2005	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	2,404	4.92
	NAEP High School Transcripts—How is Grade Point Calculated? http://nces.ed.gov/nationsreportcard/hsts/howgpa.asp	1,736	3.55
	NAEP—The Nation's Report Card http://nces.ed.gov/nationsreportcard	15,095	49.95
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	1,790	5.82
Dec. 2005	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	1,460	4.83
	NAEP High School Transcripts—How is Grade Point Calculated? http://nces.ed.gov/nationsreportcard/hsts/howgpa.asp	1,356	4.49

Continues next page

Appendix C-2. Most common entry pages for the NAEP site by month (Continued)

Month	Page	Visits	%
Jan. 2006	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	21,374	44.26
	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	3,166	6.56
	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	2,242	4.64
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	1,477	3.06
	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard	20,349	42.51
Feb. 2006	NAEP Questions Tool v3.0 http://nces.ed.gov/nationsreportcard/itmrls	2,849	5.95
	NAEP—State Profiles and Student Demographics http://nces.ed.gov/nationsreportcard/states/	2,667	5.57
	NAEP—State Profile: AL http://nces.ed.gov/nationsreportcard/states/profile.asp	1,486	3.10
	NAEP—The Nation’s Report Card http://nces.ed.gov/nationsreportcard		

Appendix C-3. Most commonly accessed pages on the www.nationsreportcard.gov site by month

Month	Page	Visits	Views
Oct. 2005	The Nation's Report Card http://nationsreportcard.gov/reading_math_2005/	35,937	45,755
	[Redirect page to current Initial Release site home page] http://nationsreportcard.gov/	29,465	34,054
	NAEP Reading Mathematics 2005—Reading: State Results: State Achievement Levels http://nationsreportcard.gov/reading_math_2005/s0006.asp	20,848	60,999
	NAEP Reading Mathematics 2005—Mathematics: State Results: State Achievement http://nationsreportcard.gov/reading_math_2005/s0021.asp	13,133	36,901
	NAEP Reading Mathematics 2005—Reading: National Results: Average Scale Score http://nationsreportcard.gov/reading_math_2005/s0003.asp	7,801	12,117
	NAEP Reading Mathematics 2005—Reading: Summary http://nationsreportcard.gov/reading_math_2005/s0002.asp	7,288	9,267
	NAEP Reading Mathematics 2005—Mathematics: National Results: Average Scale http://nationsreportcard.gov/reading_math_2005/s0018.asp	5,652	8,788
	NAEP Reading Mathematics 2005—Mathematics: Summary http://nationsreportcard.gov/reading_math_2005/s0017.asp	4,411	5,343
	NAEP Reading Mathematics 2005—Reading: Student Group Results: Race and Ethnicity http://nationsreportcard.gov/reading_math_2005/s0011.asp	3,426	6,116
	NAEP Reading Mathematics 2005—Downloads and Tools http://nationsreportcard.gov/reading_math_2005/s0046.asp	3,353	4,127
Nov. 2005	[Redirect page to current Initial Release site home page] http://nationsreportcard.gov/	5,246	5,956
	The Nation's Report Card http://nationsreportcard.gov/reading_math_2005/	4,806	5,969
	NAEP Reading Mathematics 2005—Reading: State Results: State Achievement Level http://nationsreportcard.gov/reading_math_2005/s0006.asp	1,703	4,462
	NAEP Reading Mathematics 2005—Mathematics: State Results: State Achievement http://nationsreportcard.gov/reading_math_2005/s0021.asp	1,076	2,838
	NAEP Reading Mathematics 2005—Mathematics: National Results: Average Scale Score http://nationsreportcard.gov/reading_math_2005/s0018.asp	1,044	1,386
	NAEP Reading Mathematics 2005—Reading: National Results: Average Scale Score http://nationsreportcard.gov/reading_math_2005/s0003.asp	894	1,393

Continues next page

Appendix C-3. Most commonly accesses pages on the www.nationsreportcard.gov site by month
(Continued)

Month	Page	Visits	Views
Nov. 2005	NAEP Reading Mathematics 2005—Reading: Summary http://nationsreportcard.gov/reading_math_2005/s0002.asp	844	1,095
	NAEP Reading Mathematics 2005—Mathematics: Summary http://nationsreportcard.gov/reading_math_2005/s0017.asp	512	645
	NAEP Reading Mathematics 2005—Downloads and Tools http://nationsreportcard.gov/reading_math_2005/s0046.asp	388	454
	Document Moved http://nationsreportcard.gov/reading_math_2005/s0007.asp	369	1,187
Dec. 2005	NAEP—TUDA Reading Mathematics 2005 http://nationsreportcard.gov/tuda_reading_mathematics_2005/	8,430	15,281
	[Redirect page to current Initial Release site home page] http://nationsreportcard.gov/	3,952	4,678
	NAEP—TUDA Reading Mathematics 2005: Reading Results by Race/ Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0006.asp	1,979	5,425
	NAEP—TUDA Reading Mathematics 2005: Reading Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0003.asp	1,892	2,859
	NAEP—TUDA Reading Mathematics 2005: Reading Summary http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0002.asp	1,798	2,226
	NAEP—TUDA Reading Mathematics 2005: Reading District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0012.asp	1,489	4,402
	NAEP—TUDA Reading Mathematics 2005: Reading Scale Score Trends http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0004.asp	1,400	2,091
	NAEP—TUDA Reading Mathematics 2005: Mathematics Results by Race and Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0019.asp	1,211	3,283
	NAEP—TUDA Reading Mathematics 2005: Mathematics Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0016.asp	1,083	1,612

Continues next page

Appendix C-3. Most commonly accessed pages on the www.nationsreportcard.gov site by month
(Continued)

Month	Page	Visits	Views
Dec. 2005	TUDA Reading Mathematics 2005: Mathematics District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0025.asp	1,030	3,292
	[Redirect page to current Initial Release site home page] http://nationsreportcard.gov/	1,701	1,896
Jan. 2006	NAEP—TUDA Reading Mathematics 2005: Reading District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0012.asp	410	1,102
	TUDA Reading Mathematics 2005: Mathematics District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0025.asp	406	1,070
	NAEP—TUDA Reading Mathematics 2005 http://nationsreportcard.gov/tuda_reading_mathematics_2005/	395	536
	NAEP—TUDA Reading Mathematics 2005: Reading Results by Race and Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0006.asp	316	549
	NAEP—TUDA Reading Mathematics 2005: Mathematics Results by Race Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0019.asp	306	506
	NAEP—TUDA Reading Mathematics 2005: Reading Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0003.asp	186	234
	NAEP—TUDA Reading Mathematics 2005: Mathematics Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0016.asp	147	191
	NAEP—TUDA Reading Mathematics 2005: Reading Summary http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0002.asp	130	150
	NAEP—TUDA Reading Mathematics 2005: Mathematics Sample Questions http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0026.asp	130	151

Continues next page

Appendix C-3. Most commonly accessed pages on the www.nationsreportcard.gov site by month
(Continued)

Month	Page	Visits	Views
Feb. 2006	[Redirect page to current Initial Release site home page] http://nationsreportcard.gov/	1,665	1,870
	TUDA Reading Mathematics 2005: Mathematics District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0025.asp	410	979
	NAEP—TUDA Reading Mathematics 2005: Reading District Comparisons http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0012.asp	406	962
	NAEP—TUDA Reading Mathematics 2005: Reading Results by Race and Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0006.asp	327	664
	NAEP—TUDA Reading Mathematics 2005: Mathematics Results by Race and Ethnicity http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0019.asp	316	711
	NAEP—TUDA Reading Mathematics 2005 http://nationsreportcard.gov/tuda_reading_mathematics_2005/	311	408
	NAEP—TUDA Reading Mathematics 2005: Mathematics Results By English Language http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0024.asp	164	245
	NAEP—TUDA Reading Mathematics 2005: Reading Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0003.asp	160	216
	NAEP—TUDA Reading Mathematics 2005: Mathematics Overall Results http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0016.asp	153	205
	NAEP—TUDA Reading Mathematics 2005: Reading Results by English Language Learners http://nationsreportcard.gov/tuda_reading_mathematics_2005/t0011.asp	136	249

**Appendix D: NAEP Web-Based Score Reporting Evaluation:
Review of Web Site Usage and Usability Research Methodologies**

April L. Zenisky and Stephen Jirka
University of Massachusetts Amherst

Feb. 3, 2006

Introduction

Given the importance of test scores as indicators of performance for high-stakes accountability purposes, appropriate and effective dissemination of the results of student testing is an important activity for many states and their test development contractors. With the passage of the *No Child Left Behind Act*, annual state testing cycles provide a wealth of information about students' performance on tests, and often the results are released disaggregated by schools, districts, and major reporting groups. In addition, test results are used to track students' performance over time. While we note that there are many potential intended and unintended uses of these data, the sheer volume of score information produced has resulted in progressively more and more of the statistics being made available to educators, the media, policymakers, and the public via the World Wide Web.

For the National Assessment of Educational Progress (NAEP), the Internet is a primary means by which interested audiences can access test results. The NAEP testing program in 2005 involved 50 states and other jurisdictions, and approximately 320,000 fourth-graders and 303,000 eighth-graders in reading and mathematics. As the "Nation's Report Card," NAEP is "the only nationally representative and continuing assessment of what America's students know and can do in various subject areas" (NCES, n.d.). Much information about the NAEP testing program, including multiple years' worth of results for a number of content areas including core subjects such as Mathematics and Reading, is currently available on the NCES-hosted NAEP Web site (<http://nces.ed.gov/nationsreportcard/>).

Jaeger (1998) identified multiple different audiences for NAEP results (including the federal executive branch, congressional staff members, the state executive branch, state legislatures, district-level administrators and professional staff, school principals and teachers, the general public, members of the press, and educational research personnel). While the Web is not the only means by which NAEP results are being disseminated to these audiences, the increasingly key role of the NAEP Web site as a source for quick access to information about NAEP is unsurprising in today's world. NAEP's presence on the Internet is evolving and expanding steadily, as evidenced by the creation of an Initial Release Web site for special events such as the 2005 fourth and eighth grade Mathematics and Reading results release as well as ongoing efforts with respect to developing Web-based data analysis tools such as the NAEP Data Explorer (NDE).

The utility of score reports and score reporting methods used for NAEP has been identified as one of the priorities in the congressionally mandated evaluation of NAEP. The evaluation of NAEP score reporting must include examination of possible Web reporting methodologies that may provide insight into the user experience on the main NAEP Web site. Does the Web site present what its users need or want to know? Are the users taking advantage of all the information being presented? What else do the users wish the Web site contained? How can these questions best be answered? From the disciplines of library and information science to marketing to economics to psychology and computers, many researchers are focusing on factors that affect the experience of visitors to different sites and their behavior when navigating a site. One key area of interest is *how* visitors use a Web site to conduct research. In this paper, we review different research methods for evaluating Web sites. As we plan for studies with different NAEP audiences, this methodological review will inform the design of our studies. First, we review the NAEP Web site and the kinds of resources visitors to the site encounter. Subsequently, we identify multiple research techniques for evaluating Web sites.

Brief Review of NAEP's Current Electronic Resources

From the main NAEP Web site (<http://nces.ed.gov/nationsreportcard/>), consumers interested in NAEP data can get access to a tremendous amount of information. These resources fall into four main categories: programmatic Web pages, static data-oriented Web pages, interactive and media tools, and downloadable PDFs of paper-based NAEP reports that have

been released over the years. For the purposes of categorizing the resources on the site, we define these categories as follows.

- *Interactive and media tools* are defined by a high degree of user choice in generating what results or analysis are called up to be displayed on a page: we refer here to the use of multimedia and clickable data resources which, for example, might allow users to manipulate the format (tables or graphs), information (scale scores, proficiency levels, percentiles), and type (national, state, subgroups, gaps, etc.) of results displayed.
- In contrast, *static data-oriented Web pages* provide assessment findings in structured tables, charts, and/or text formats that Web site users cannot manipulate.
- A number of pages on the NAEP site contains links to numerous *downloadable PDFs*, which package information in easy-to-print formats for user review, often in traditional technical report-style layouts with tables of contents.
- *Programmatic Web pages* are text-based resources accessible by branching off the main NAEP page which are explanatory in nature and do not contain assessment data or results.

Example of items in each of these four categories currently found on the NAEP Web site are shown in Figure D-1.

Figure D-1. Selected Web resources on <http://nces.ed.gov/nationsreportcard/> by category

<p><i>Programmatic Web Pages</i></p> <ul style="list-style-type: none"> • Overview • Inclusion of Special Needs Students • FAQs About NAEP (State NAEP) • NAEP Assessment Schedules • Site Map (in brief and detailed) • Glossary (pop-up) 	<p><i>Static Data-Oriented Web Pages</i></p> <ul style="list-style-type: none"> • State Profiles • Long-Term Trend Key Findings slides • Long-Term Trend Summary Data Tables • Exclusion Rates
<p><i>Downloadable PDFs</i></p> <ul style="list-style-type: none"> • 150+ Report Cards and other reports Arts, Civics, Geography, Mathematics, Reading, Science, U.S. History, Writing (1990–2005) • Trial Urban District Snapshot Reports • 20 Technical/Methodological papers • NAEP frameworks documents for assessed subject areas 	<p><i>Interactive and Media Tools</i></p> <ul style="list-style-type: none"> • NAEP Data Explorer • NAEP Question Tool • Item Maps (linked to Question Tool) • NAEP Newsflash e-mail alert sign-up • Archived results release Webcasts • Cross State Comparison Maps

Since these are the kinds of Web site features that must be evaluated, our discussion of the Web evaluation methodologies in the next section will include strategies that are in varying degrees appropriate to evaluating the utility of each of those types of resources.

Research Techniques for Evaluating Web sites

Studies evaluating the usability and utility of public Web sites generally look at two primary dimensions of the user experience: first, what do users *do* on a site, and second, how do they *feel* about the experience? The purpose of this short document is to present a number of options that might be used to conduct a study on the utilization and usability of the NAEP Web site. In this section, a description of each methodology is presented. The options will be listed according to the level of perceived complexity and involvement on the part of users of the Web site and the researchers, from least to most.

Browser Tracking Software: The first option entails purchasing commercial software that tracks the movements of Web site users and analyzes a log file of online behavior while a visitor is on a Web site. Tracking software for the most part entails the least engagement with visitors,

as this method for evaluating usage often is used in a way that is largely invisible to visitors. However, the trade-off is information that is generally very broad in nature. Common Web usage statistics reported include daily and monthly summaries of the number of page views, the number of unique hosts visiting a site, the most popular pages on a site, the percent of different kinds of browsers visiting a site, the duration of visits, and the time spent on different pages. It is important to note, however, that privacy laws limit more in-depth data collection.

A quick search on the Web reveals there are many products available and many companies that specialize in this service. For example, www.wusage.com provides a log file that tracks where users are coming from, if they are repeat or unique visitors, how often they visit the site, and what specific pages within the site they are accessing. Products such as this one can offer basic information on how the Web site is used. This option would probably have the least type of involvement from staff, other than setting up the software and possible consultation with the vendor. Cost can be high, depending on the set up, and this would take some amount of time and effort to set up the software. Subjects per se would not have to be recruited, so there would not be any costs for this aspect. A window for how long Web site usage would be tracked would need to be decided.

Online or Paper Survey: A second option to consider in this research study might be to administer a survey, either paper or online, to a sample of the population using the Web site. Surveys have been done on studies of online Web courses (Zaphiris and Zacharia, 2001) and in other educational settings (Dix and Anderson, 2000), as well as evaluations of Web sites (Zhang and von Dran, 2001). There are several questionnaires that have been used in Web site usability studies, including the SUS (System Usability Scale), the QUIS (Questionnaire for User Interface Satisfaction), the CSUQ (Computer System Usability Questionnaire), and another utilized by Microsoft as part of reaction studies to products (Tullis and Stetson, 2004). Questions from uses such as these can be adapted for evaluation of the NAEP Web site. In the NAEP-specific context surveys have been used for state assessment directors, curriculum directors, school superintendents, chief state school officers, education association staff, state board of education chairs, governors and education policy aides, and state legislature education committee staff by Levine, et al. (1998). With surveys, a large number of participants can be solicited, and the information gathered is user self-reports generally related to user preferences, attitudes, and experiences.

Research staff would have to spend some setup time before the survey is ready to be sent out, for development, perhaps a brief field test, and final review. However, the possibility of later needing a specific piece of information that was not included in the initial survey may cause problems, so this option lacks some flexibility. The option of follow-up questions for clarification based on the respondents' answers is not an option here as well. If a paper survey were used, time to prepare and distribute these must be considered, along with the additional costs necessary. Administering the survey online using a free Web site such as www.surveymonkey.com will help defray these costs and has the benefit of being easier to gather of the data and the possibility for participation by more subjects.

One-on-one interviews: By meeting with different users of the NAEP Web site individually, it is possible to talk with people in-depth to discuss attitudes, experiences, and self-reports of behavior. This method allows for the use of both closed and open-ended questions, and can be a highly rich source of data because the one-on-one format allows a trained interviewer to follow up on user comments to gather more information in depth. Interviews can be face-to-face, by telephone, or through electronic means such as videoconferencing or instant messaging. One-on-one interviews with NAEP stakeholders looking at paper-based reporting methods have been conducted by Ogilvy Public Relations Worldwide (2004).

One-on-one contextual observations: This method involves observing users in the context of their homes or place of business while going about normal, everyday activities (e.g., Blackwell, et al., 2005). The researcher is interested in collecting information about how a site is used in practice and the task analysis information collected can be used to inform design of a site (placing more commonly used links or information more prominently on a page, and also to identify tasks and activities that might benefit from closer examination using other methods). An

additional benefit of this approach is the capability to observe the kind of technology users have at their disposal (screen resolution, hardware and software standards) which can also be critical in information the design of a site. Ultimately, by observing users in naturalistic use of a site, researchers can gain a great deal of information about how users read screens, scroll through options, and make browsing choices that may be informative for redesigning sites.

One-on-one “Treasure Hunt” observations: In this method the participant is assigned to explore the functionality of a site to find, retrieve or download specific pieces of information or perform a function or transaction (English, et al., 2001). The treasure hunt is also sometimes referred to as a cognitive walk-through. Government Micro Resources, Inc. (2004) used the treasure hunt approach to evaluate an early prototype of the NAEP Data Explorer (then called the NAEP Data Analyst). NAEP users, for example, might be directed to find specific results or programmatic information such as the definition of terms or data about participation. Other common treasure hunt activities might include the exploration of an interactive functionality, such as the NDE or the Question Tool, or require participants to download an application or document. The use of the observed treasure hunt was used successfully with respect to the development of the FedStats Web site (<http://www.fedstats.gov>; Ceaparu, 2003). The task can be varied in difficulty by setting the “treasure” as something linked to on the home page, or choosing the target to be a result or finding that is more buried in the site or which requires the user to more actively interact with a Web tool to obtain. Throughout the observation process, each step the user takes is scrutinized, and places where the interface serves to “roadblock” a user from completing the task are taken as indicators of a usability problem.

Think aloud or Delphi method: A variation on the contextual and treasure hunt one-on-one observations described above is to have a smaller sample of subjects sit with a researcher and “think aloud” or vocalize their thought process as they use the Web site to accomplish a set of common tasks. This narration provides insight into what users are thinking as they make choices. Researchers could have a series of prepared tasks or questions and the users could then respond to them, all the while vocalizing their thoughts. How the user responds to problems while attempting to perform the tasks can be captured as an important piece of data. This technique has been used numerous times in analyzing Web site usage (Beaton, et al., 1998; Brower, et al., 2002; Yin, et al., 2002; Benbunan-Fich, 2001) and usage of a statistical Web site in particular (Ceaparu, 2003). One of the advantages of the think aloud approach is that it gathers the information from the users in real time while they are completing a real task. Additionally, it avoids any issues of the user forgetting or trying to justify their actions in retrospect (Benbunan-Fich, 2001). Recording the session either visually or audibly (as well as the keystrokes) might ensure that a more accurate record is kept of the session. This method also allows for more immediate feedback for the researcher and immense flexibility for follow-up questions based on the subjects’ responses. It has been reported that fewer subjects, some say as few as five, are needed to gather enough information (Benbunan-Fich, 2001).

A drawback of this method would be the logistics of gathering the sample of subjects and the time required for examining each individual one. Will the researcher go to the subjects’ location or will the subjects be brought to one place? Other criticisms of the think aloud method are the lack of realism in the tasks and the interference that the subjects having to talk while doing the tasks might affect the tasks. It may take them longer to perform the tasks, or they might be self-conscious. In the end, the responses from fewer subjects may or may not become an issue for the variety of opinions and the amount of information gained. A carefully thought out sampling plan should address these issues.

Eye-tracking: This is an interesting approach to data gathering that provides unique information about users’ interaction with a Web site (Duchowski, 2003; Russell, 2005b). In this method, a computer monitor is set up with an integrated eye-tracking system that can detect and collect participants’ eye-gaze data while they are on a Web site. Users’ eyes can be tracked nearly continuously (in one study by Russell (2005a) readings were taken on an average of every 20 milliseconds), and the collected data is examined in terms of fixations recorded within areas of interest. Examples of areas of interest can include menus on a page, graphics, and banners atop of a page. Results of eye-tracking studies are often quantified with respect to the order in

which different areas of a page receive their first fixation, the number of fixations recorded in each area of interest, and the cumulative “dwell time” of fixations recorded in each area of interest.

Focus Groups: Another option would be the use of a focus group or groups. Depending on how the focus group is implemented, this strategy may require the use of the most amount of resources from participants. This entails gathering a small group of subjects, typically 8–12 so as to give everyone a chance to participate, and the use of a trained moderator to facilitate a discussion (Morgan and Stinson, 1997). These focus groups can be conducted by assembling the participants in person or virtually by way of videoconferencing or WebEx (which can minimize the effects of group dynamics). The focus group method has been successfully utilized as an aspect of a study of how consumers gather and evaluate information from medical Web sites (Eysenbach and Kohler, 2002), and would provide self-report data from participants with respect to feedback, initial reactions to a design, and discuss their preferences. Studies using focus groups composed of NAEP stakeholders include Levine, et al. (1998) and Government Micro Resources, Inc. (2005). To a limited extent focus groups can be used for some usability testing, but primarily opinions would be collected. Focus groups can be useful for raising issues that may not come out during interviews.

The facilitators would be able to have a set of prepared questions that can be given to all members to fill out and then the facilitator would be able to look at these answers and then ask a series of feedback or follow up questions. During these focus group meetings, a subset of the participants might be asked to do a think-aloud session to gather even more information. Transcripts are often later analyzed for patterns of responses and coded. Cost and time involvement would be the limiting factors for this option, and the budget of the project would have to be taken into consideration, but another possibility would be the use of virtual focus groups. Again, there may be issues of a limited sample, but if care is taken in the selection, then this issue will be minimized.

In Table D-1, below, is a summary of the Web usability evaluation methodologies identified in this report. Brief descriptions of how each method works and of the primary information gathered are provided.

Table D-1. Summary of Web evaluation methods

Method	How it Works	Primary Information Gathered
Tracking Software	When users visit a Web site, a program on the server records information about the visit.	Statistical data about Web site traffic, including page views, duration, domain, and referring hosts
Online or Paper Survey	Users are e-mailed a URL or sent a document with a series of questions inquiring about their perceptions of and behaviors during use of a site.	Self-reports of user attitudes, preferences, and experiences
One-on-one interviews	Users are called or visited by a trained researcher who follows a question protocol to find out about perceptions of and behaviors during use of a site.	Self-reports of user attitudes, preferences, and experiences
One-on-one contextual observations	Users are assigned a treasure hunt (a specific sequence of tasks) to complete on a Web site, and are observed while in progress.	Observations of users in naturalistic setting to obtain information about authentic use of online resources
One-on-one 'Treasure Hunt' observations	Users are assigned a treasure hunt (a specific sequence of tasks) to complete on a Web site, and are observed while in progress.	Observations of users in controlled setting to obtain information about usability of online resources
Think aloud or Delphi protocol	Users are given a specific sequence of tasks to complete on a Web site, and are asked to provide a running narrative of their experience and perceptions while in progress.	Observations of users in naturalistic or controlled setting to obtain information about how users resolve problems as well as perceptions of usability
Eye-tracking	A computer monitor is set up with an integrated eye-tracking system used to detect and collect participant eye-gaze data during testing.	Statistical data on visual fixation and attention to areas of interest on pages within a site
Focus Groups	Small groups of users are brought together either virtually or in person to stimulate discussion about aspects of the user experience and perceptions of the site.	Self-reports of user attitudes, preferences, and experiences

Summary

Evaluating aspects of the utility of the NAEP Web site requires multiple approaches, as this evaluation encompasses several different dimensions of reporting NAEP results via the Web. Some methods described here involve observation and analysis of visitor behavior in natural, nonexperimental conditions, which can help quantify which features are of most interest to visitors and to further understanding of what normal visitors to the site focus on. Other strategies are designed to learn more about how visitors interact with and use specific tools to accomplish specific tasks on the site. Still others focus broadly on self-reports of the visitor experience. In mapping out the current agenda in the Utility study for capturing an accurate picture of the efficacy of NAEP's Web reporting activities, many of the Web research techniques presented here can be carried out efficiently and expediently to provide rich sources of data about the utility of current NAEP score reporting efforts.

References

- Beaton, A., Nicholson, S., Halliday, N., and Thomas, K. (1998). *HCI Lecture 5-Think-aloud protocols*. Retrieved July 26, 2003, from <http://staff.psy.gla.ac.uk/~steve/HCI/cscln/trail1/Lecture5.html>.
- Benbunan-Fich, R. (2001). Using protocol analysis to evaluate the usability of a commercial web site. *Information and Management*, 39, 151–163.
- Blackwell, A., Jones, R., Milic-Frayling, N. and Rodden, K. (2005). Combining logging with interviews to investigate web browser usage in the workplace. Presented at workshop on *Usage analysis, combining logging and qualitative methods*. CHI 2005. Retrieved Jan. 18, 2006, from <http://www.usage.nl/docs/chi2005-usageworkshop.pdf>.
- Brower, G., Raphael, A., and Missimer, C. (2000). *Site evaluation: Think-aloud protocol—getting your users to talk while they work*. Retrieved July 26, 2003, from www.washington.edu/webguides/design.class/thinkaloud.html.
- Ceaparu, I. (2003). Finding Governmental Statistical Data on the Web: A Case Study of FedStats. *IT and Society* 1(3), 1–17.
- Dix, K., and Anderson, J. (2000). Distance no longer a barrier: Using the Internet as a survey tool in educational research. *International Education Journal*. Retrieved Oct. 8, 2005, from <http://ehlt.flinders.edu.au/education/iej/articles/v1n2/DIX/begin.HTM>.
- Duchowski, A. T. (2003). *Eye tracking methodology: Theory and practice*. London: Springer.
- English, J., Hearst, M., Sinha, R., Swearington, K. and Yee, P. (2001). Examining the usability of web site search. Unpublished manuscript. Retrieved Jan. 18, 2006, from <http://bailando.sims.berkeley.edu/papers/epicurious-study.pdf>.
- Eysenbach G, and Köhler C. (2002). How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ (British Medical Journal)*, 324, 573–577.
- Government Micro Resources, Inc. (2005). *NAEP ambassadors and the initial release site usability study*. Manassas, Va.: Author.
- Government Micro Resources, Inc. (2004). *NAEP data analysts usability study*. Manassas, Va.: Author.
- Jaeger, R. (1998). *Reporting the results of the National Assessment of Educational Progress (NVS NAEP Validity Studies)*. Washington, D.C.: American Institutes for Research.
- Levine, R., Rathbun, A., Selden, R., and Davis, A. (1998). *NAEP's constituents: What do they want? Report of the National Assessment of Educational Progress Constituents Survey and Focus Groups*. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, NCES 98-521.
- Morgan, D., and Stinson, L. (1997). *What are focus groups?* Alexandria, Va.: Section on Survey Research Methods, American Statistical Association
- National Center for Education Statistics. (n.d.). *NAEP—Overview*. Retrieved Jan. 18, 2006, from <http://nces.ed.gov/nationsreportcard/about/>.

- Ogilvy Public Relations Worldwide (2004). *NAEP: Reporting initial results—Analysis and recommendations for improvement*. Washington, D.C.: Author.
- Russell, M. (2005a). Using eye-tracking data to understand first impressions of a web site. *Usability News*, 7(1). Retrieved Jan. 18, 2006, from http://psychology.wichita.edu/surl/usabilitynews/71/eye_tracking.html.
- Russell, M. (2005b). Hotspots and hyperlinks: Using eye-tracking to supplement usability testing. *Usability News*, 7(2). Retrieved Jan. 18, 2006, from <http://psychology.wichita.edu/surl/usabilitynews/72/eyetracking.htm>.
- Tullis, T. S., and Stetson, J. N. (2004). *A Comparison of Questionnaires for Assessing Web site Usability*. Usability Professionals Association (UPA) 2004 Conference, Minneapolis, Minn., June 7–11, 2004.
- Yin, Y., Ayala, C., and Shavelson, R. (2002). Hands On or Minds On: Cognitive Activities in Performance Assessments: An Empirical Study on Think Aloud. PowerPoint presentation. Retrieved July 26, 2003, from www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/Conference%20paper%20PDF/AE_RA2002_Melody.pdf.
- Zaphiris, P., and Zacharia, G. (2001). User-Centered Evaluation of an On-Line Modern Greek Language Course. In *Proceedings of WebNet 2001 Conference*, October 23–27. Orlando, Fla.
- Zhang, P., and von Dran, G. (2001). User expectations and ranks of quality factors in different web site domains. *International Journal of Electronic Commerce (IJEC)*, 6(3), 9–34.

**Appendix E: Do Mathematics Educators Use and Understand NAEP Score Reports?
Evaluating the Utility of Selected NAEP Data Displays**

April L. Zenisky, Ronald K. Hambleton, and Zachary R. Smith
Center for Educational Assessment
University of Massachusetts Amherst

Abstract

No matter what methods for disseminating scores and results from large-scale educational assessments are chosen, they have great implications for the potential of test data to be used to help schools, districts, and states make data-based decisions about instruction and student progress. These methods must be scrutinized for their effectiveness with respect to different intended audiences. With the National Assessment of Educational Progress (NAEP), the amount of data available and the multiple methods used to communicate the results appear to have the potential to be a rich source of information for some, and for others, a daunting mass of confusing numbers. As part of a larger evaluation of the utility of NAEP score reports, a focus group composed of mathematics curriculum leaders from across the United States was held to explore the extent to which different NAEP data displays have meaning and usefulness. Perhaps the most important finding is that even educators with quantitative skills experienced some difficulty with many of the common NAEP score reports. The focus group made several suggestions for revising the layout of several data displays, particularly with respect to footnotes and arrangement of keys or legends within figures. This study highlights (1) the utility of focus groups for gaining insights about the NAEP score reports, and (2) the importance of either revising the NAEP score reports to make them more user-friendly or the need for more explanatory materials for persons using the NAEP reports. Of course, a focus group of eight is not a sufficient basis for initiating major report changes, but it does suggest the need for substantially more research, and immediately, if NAEP is going to achieve the high hopes that many policymakers have for it.

Introduction

Educators are increasingly being provided with large amounts of data about how students are doing, with the expectation that the data will be used to assess student achievement and develop instructional strategies and improvement plans. However, for many educators, formal data-based decision-making is no small undertaking (Herman and Gribbons, 2001), because they are not always well-prepared to understand the data they receive or sure about how to connect that information to specific instructional strategies or activities. Indeed, Sharkey and Murnane (2003) identified three challenges curtailing wider use of test data among educators, in areas of technology, opportunity, and knowledge: (1) Educators need data systems that are user-friendly and provide data in multiple formats (e.g., tables *and* graphs) specific for their needs, (2) they need time and resources to make sense of the data, and (3) they need training to help them identify the kinds of questions that might be useful to them given the data available to them. Jaeger (1998) identified several topics for study that had particular relevance for educators as he outlined a larger research agenda on NAEP score reporting, including aspects of what to report, how to report the information, and the method by which the data is disseminated to educators as users of assessment data.

Hambleton and Slater (1996) found that both policymakers and educators had difficulty with NAEP data displays contained in the Executive Summary Reports—this was surely a disappointing result to NCES because these reports were widely distributed (about 100,000 copies of each Executive Summary Report) to policymakers and educators. For states and their testing contractors, score reporting clearly represents an important area for investment and professional development to ensure that assessment findings are part of the broader process of instructional planning, as is intended, and this reporting is part of the justification for spending large sums of money on state and national testing programs.

With respect to NAEP-specific score reporting practices, while NAEP does not have the formal instructional connection with state curricula that state assessments do, NAEP results do represent another important source of information about how students in the nation and individual states and jurisdictions are doing. Also, and importantly, as a leader among K–12 testing programs in its role as the “Nation’s Report Card,” NAEP is at the forefront of developing methods to communicate test results to a wide variety of audiences. Dissemination of NAEP results draws on a wide array of score reporting strategies, including a number of Web-based tools that are readily available for interested parties to use to access results of specific interest. If NAEP can increase clarity, meaningfulness, and use of its reports, it is likely that states would quickly follow with similar reporting practices.

Purpose of the Study

Within the framework of focusing studies of NAEP reporting along two dimensions, *use* and *understanding*,³⁷ as informed by Jaeger (1998), the purpose of this study was to explore the extent to which educators (in this case, state math curriculum leaders) were familiar with current methods of displaying NAEP mathematics results and what kinds of inferences they might make on the basis of those displays. Indeed, given that NAEP does employ multiple strategies and tools to communicate results to the different user audiences, this study is an important part of a broader evaluation of the utility of score reporting methods used by NAEP in gathering information as to how members of different audiences both use and understand NAEP findings. Guiding this study are questions as to 1) how NAEP results are displayed, particularly in electronic communications with respect to principles of good reporting (e.g., Goodman and Hambleton, 2004), 2) what are the ways in which users understand and do not understand the data presented, and 3) the development of alternative displays that may alleviate misunderstandings and misconceptions.

³⁷ *Use* centers on how different users of NAEP information access information and the uses of that information. *Understanding* entails representation and understanding of NAEP results.

Method

A focus group of mathematics educators was convened to gain information about the meaningfulness of different NAEP data displays for educators. The eight participants in this focus group, all mathematics curriculum leaders from different states attending a NAEP math item review meeting in Baltimore, were drawn from a list of ten math curriculum leaders at the Baltimore meeting provided by the National Center for Education Statistics. All agreed to participate in a two-hour focus group meeting in the evening after the conclusion of the item review discussions for the day (Dec. 6, 2005). In return for their attendance at the focus group, the participants were provided with dinner and a small honorarium of \$150, though four of the eight declined the honorarium because of requirements imposed on them by the states they worked for.

The eight participants in the focus group represented eight different states, and with respect to the Census Bureau reporting regions used by NAEP, two of the participants were from the West (Nevada and Utah), two were from the Midwest (Nebraska and Ohio), three were from the South (Georgia, Kentucky, and South Carolina), and one hailed from the Northeast (New Hampshire). There were six females and two males. All of the participants had more than 20 years of experience in the field of education. There were five state mathematics consultants in the group, along with a codirector of mathematics curriculum resource center at a state university, a state director of mathematics, and a state mathematics curriculum specialist. All had a background in mathematics.

All participants in this focus group were familiar with NAEP and had varying degrees of experience working with NAEP data, ranging from three who worked with NAEP data several times a week, to two who indicated that their work with NAEP data was a few times a month, to three who reported more rare use of NAEP data. Those who were the most familiar with NAEP also worked with its data and information the most often, meaning once or twice a week. Those who used NAEP only once every couple months were still somewhat familiar with it, or so they indicated on a brief survey we distributed at the beginning of the meeting. Only one participant did not work with NAEP data but reported being somewhat familiar with it.

When the participants had worked with the NAEP data in the past, most conducted item reviews, as well as studied and shared trends with schools and teachers. Some of them also compared the NAEP data with their state's curriculum or frameworks. Others worked with educators, districts, and other organizations to interpret and analyze the data.

At the focus group meeting, participants were provided with an overview of the project and asked to complete a brief demographic survey. Next, a series of data displays consisting of both tables and figures from several recent releases of NAEP mathematics results (National and State results, 2005; Long-Term Trend results, 2004, and Trial Urban District Assessment, 2005) were projected on a screen in color. These were also provided to participants as full-page handouts in black-and-white. We focused on mathematics displays to increase interest among participants. As each figure or table was displayed, participants were asked questions about those displays. The displays shown at this meeting were chosen as a sampling of the types of tables and figures seen throughout recent NAEP reports, including:

- line graphs,
- stacked bar charts,
- clickable state comparison maps of average scale scores and percents of students at or above achievement levels,
- tabs from the NAEP question tool with item text, student item performance, a distracter analysis,
- “pantyhose” charts,
- bar graphs, and
- item maps.

As each display was projected on the screen, participants were asked to reflect on each display for a few minutes, and then they were asked questions about the data display by one of the two meeting facilitators. Questions ranged from those that were informational in nature (“What was the average score for eighth-graders in 2005 in math?”) to opinion (“What, if anything, do you find confusing or not clear about this display?”). The focus group discussion format was appropriate for this study because this format served to stimulate some broader conversations among the participants and facilitators about the data displays, building on what was being displayed on the screen, and allowed the participants to answer some of the more difficult data interpretation questions collaboratively.

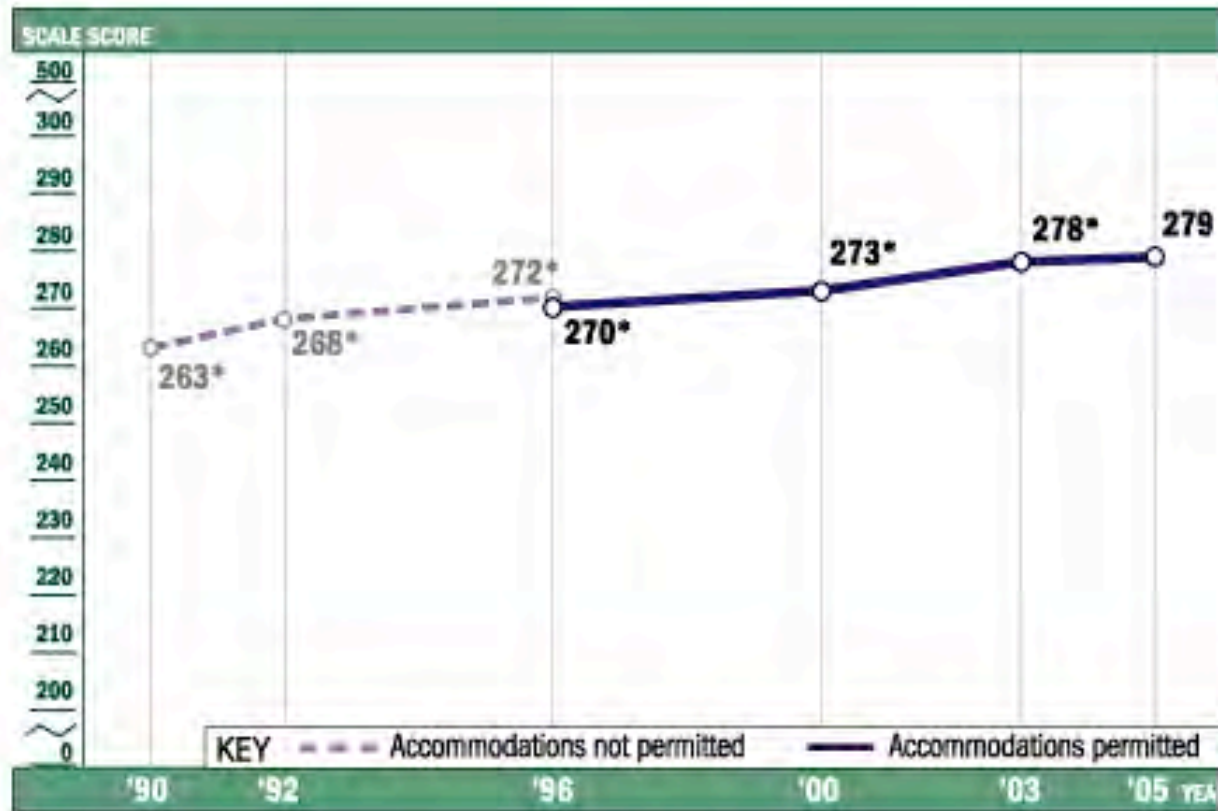
Results

Information gathered from focus group participants is presented next, with a discussion of both correct interpretations and sources of misunderstanding that were identified.

Figure E-1 presents a line graph of average mathematics scale scores showing the performance of grade 8 students nationally from 1990 to 2005. On the graph was listed the average scale score for each year that the assessment was given, and an * was placed next to the number if the difference between that year and 2005 was found to be significant in a pairwise significance test. While all of the participants in the focus group indicated that they had seen test results shown in the form of line graphs before and were generally familiar with the concept of scale scores, the range of the NAEP scale itself was not clear for this group: one participant reported having no idea of what the NAEP scale was except what could be read from this display. One participant asked what the 11-point gain observed between 1992 and 2005 really meant with respect to something other than just a gain of 11 scale score points, suggesting that that this is something that the public and educators alike would find useful for making test score changes more easily understood. The group correctly identified the average scale scores at different points in time, and could determine if differences in average scale scores were significant. Consistent with the results of Hambleton and Slater (1996), however, this group demonstrated considerably varied understanding of the correct interpretation of significant differences between two scale scores (comments made here range from “Large enough to say improvement” to “Growth, not chance. It relates to the number of people tested”). Clearly, the concept of “statistical significance” remains a mystery to some of the participants.

Figure E-1.

Average mathematics scale scores, grade 8: Various years, 1990-2005



* Significantly different from 2005.

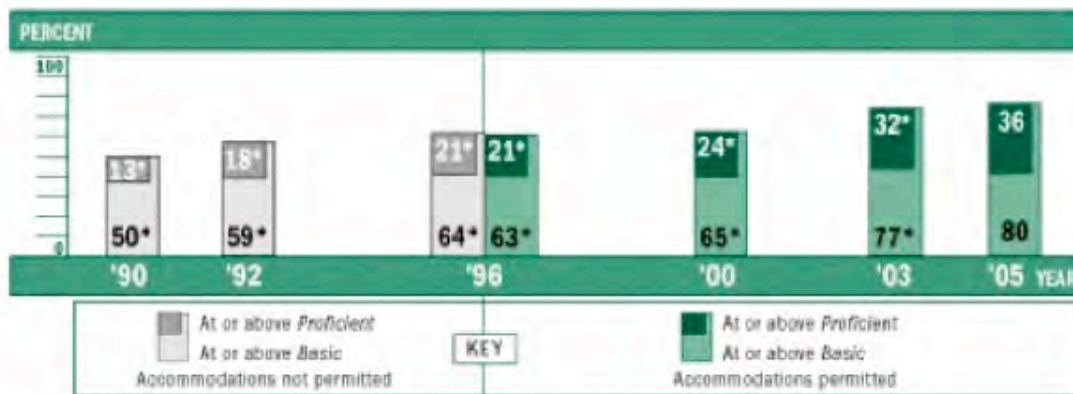
NOTE: The dashed and solid lines represent results based on administrations when accommodations were not permitted and when accommodations were permitted, respectively. View complete data with standard errors for [grade 8](#).

Figure E-2 was a stacked bar chart illustrating both the percentage of students at or above Basic and at or above Proficient in grade 4 mathematics between 1990 and 2005. Though all participants responded they were familiar with NAEP reporting using the format of “at or above,” when asked what “at or above” meant, one participant replied “those who are proficient at just a minimal level.” All participants correctly recognized that 80 percent of students in 2005 were Basic or above, and that also meant that 20 percent were below Basic and the 80 percent was composed of students in the Basic, Proficient, and Advanced categories. They could also identify significant differences in the percentages “at or above Proficient” between years.

This group raised a question about the formatting of this graph and the placement of the “Key” label below the graph. They were not sure what was meant by “Key” here, because they saw the label “Key” was on the line and this made them unsure whether the line was key to interpreting the figure or if the information displayed in the area near the word “Key” was what was important.

Figure E-2.

Percentage of students at or above *Basic* and at or above *Proficient* in Mathematics, Grade 4 Various Years, 1990-2005



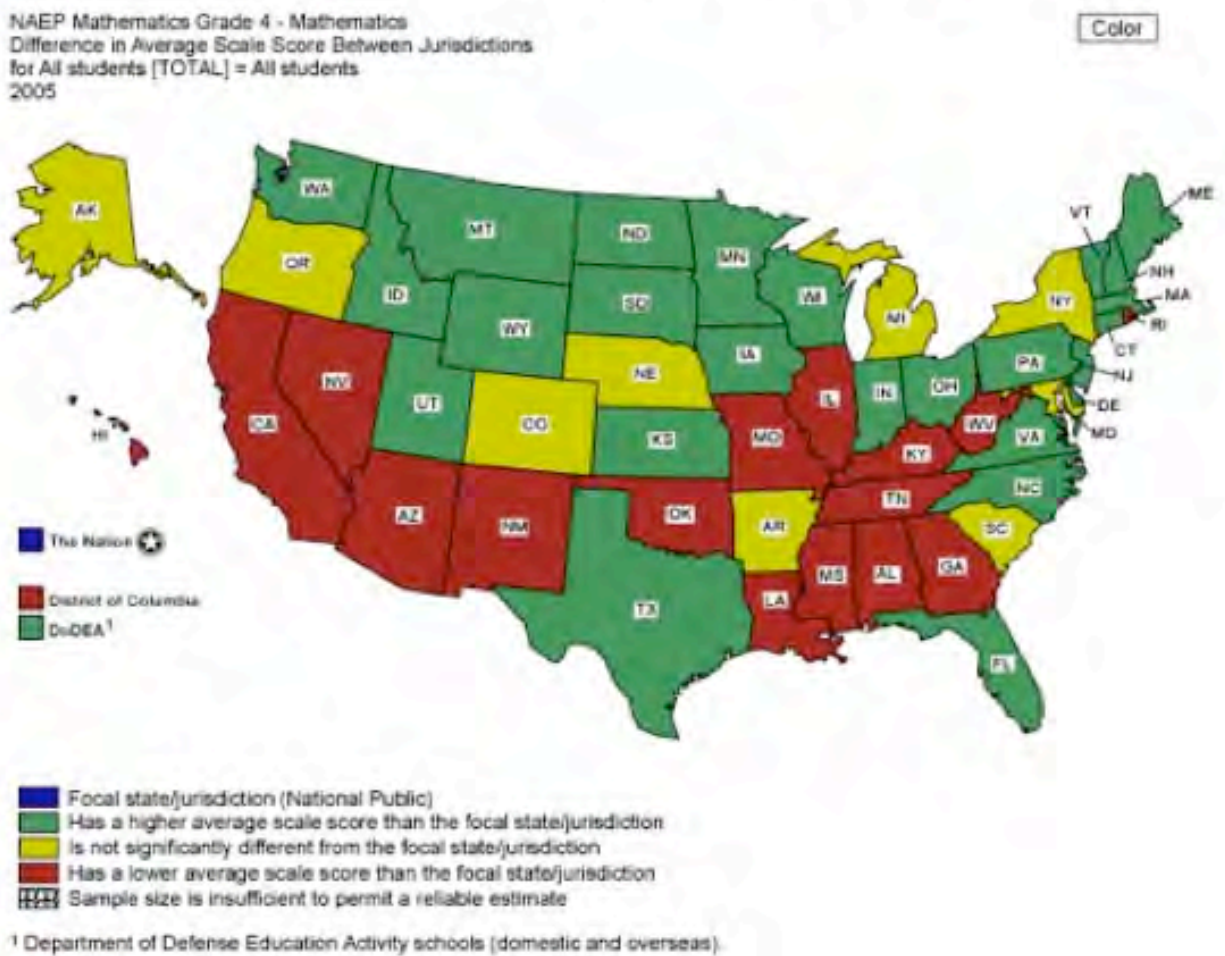
* Significantly different from 2005.

NOTE: The gray shaded boxes represent results based on administrations when [accommodations](#) were not permitted. View complete data with standard errors for [grade 4](#).

Figure E-3 was a screen capture of an interactive cross-state comparison map of average mathematics scale scores in 2005 for grade 4 public schools obtained from the NAEP Web site. In its interactive form, visitors to the NAEP Web site can choose a focal state or jurisdiction for comparing all other states or jurisdictions to. If a user wants a quick visual showing of how Massachusetts does compared to the rest of the nation, the user clicks on Massachusetts and all other states and jurisdictions become color-coded to reflect how they compare to Massachusetts with respect to either scale scores or percent at or above an achievement level (this too is a choice for each user). In Figure E-3, the focal group was chosen to be students in ‘National Public’ schools.

Figure E-3.

Cross-state comparisons of average mathematics scale scores, grade 4 public schools: 2005

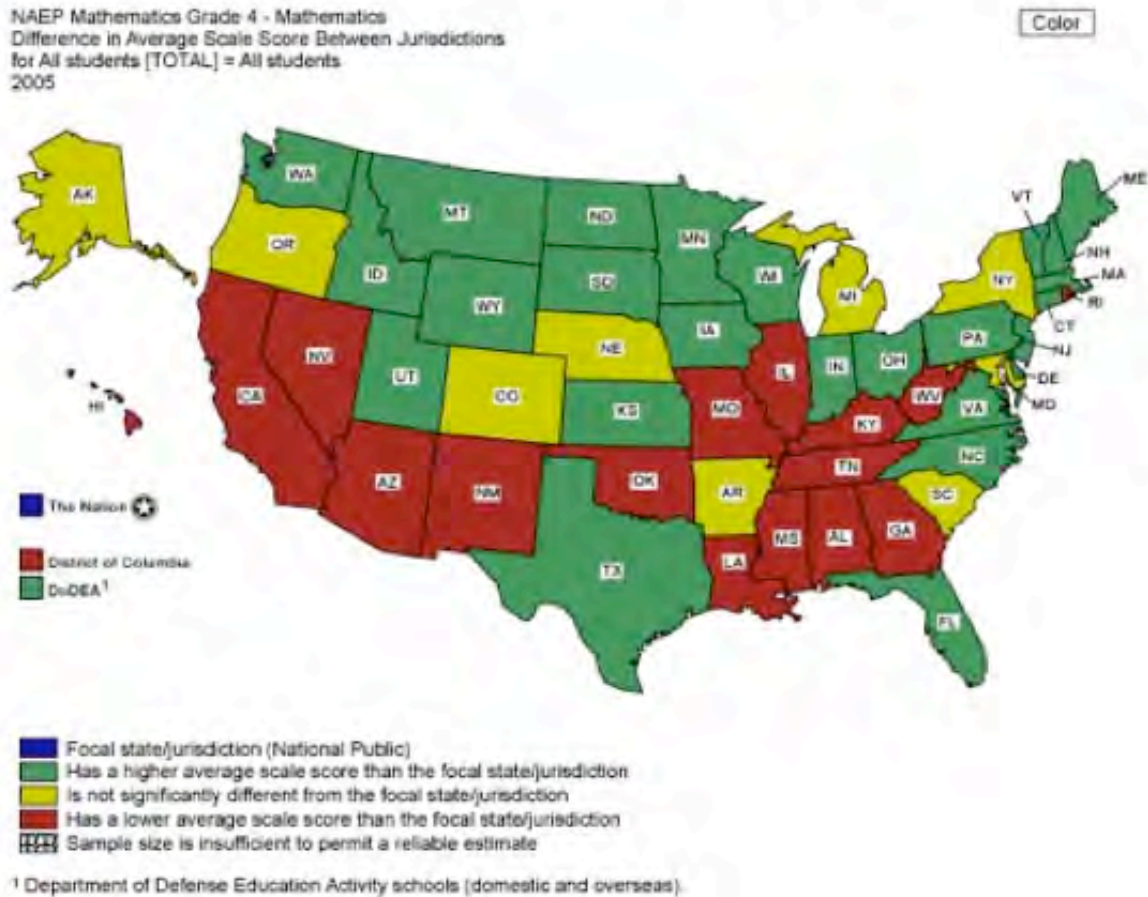


This figure posed some initial difficulty for the focus group. While about half said they had seen this particular figure before, the entire group had a very difficult time identifying which group was the focal group for comparison. This may suggest a layout or design issue, when at least one panelist knew there should be “a blue” but the participants could not find it in the display (what they were looking for was a small blue-colored square labeled “The Nation,” with a encircled star next to it). Participants in this focal group who were unfamiliar with this display seemed interested in this method of communicating results. They realized that they could access a display that would very quickly show how their home state did relative to the rest of the country in 2005.

Figure E-4 was a follow-up to Figure E-3, in that it was another screen capture of the clickable cross-state comparisons from 2005, but instead of average scale scores the data reported was the percentage of students “at or above Proficient.” Members of the group were able to answer more questions correctly about this display, given the familiarity they gained from Figure E-3, and when asked about patterns observed in the data in Figure E-4, their comments included “Lots of red in the South, Southeast,” “That big bunch in the middle looks good,” and “[Students in] California, with their tough standards, are getting it.”

Figure E-4.

Cross-state comparisons of average mathematics scale scores, grade 4 public schools: 2005



Figures E-5, E-6, and E-7 were related in that each contained information from a different tab in the NAEP Question Tool, available on the NAEP Web site. We included these displays because they would likely be of special interest to curriculum specialists. As participants were shown Figure E-5, which was the text of a multiple-choice item, they were asked how many of them had used the question tool before: Four of the eight responded affirmatively and the other four were uncertain. This group was able to look at Figure E-6 and identify how many students answered this multiple-choice item correctly, and was also very interested in the distracter analysis shown in Figure E-7. By and large, the statistics of the distracter analysis were highly familiar to these mathematics educators as well, as they were able to identify the percent of students answering the item correctly and, with a bit of prodding, knew that the average scores reported for each answer option were the average scale scores of students choosing each answer option. One area that did pose problems for some participants in interpreting results, however,

was the idea of standard errors (SEs), as one participant defined standard errors are “a sense of the error in scores,” and another said it was “a band around scores.” When asked if the SEs in Figure E-7 were large or small, one participant did have a good point, noting that because one point could make such a difference in scores being significant, the SEs here should be tiny. But most of the group had little idea about the meaning of standard errors.

Figure E-5. Example multiple choice item for Grade 4 mathematics

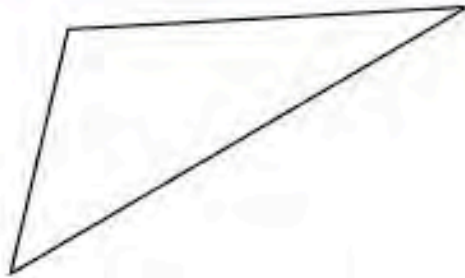
[New Search](#) [Previous Search Results](#) [Tool Help](#)
◀ Question 3 of 149 ▶ [Add Question](#)

📁 [To Print Folder: Empty](#)

Subject: **Mathematics** [[Subject Info](#)] Grade: 4 Block: 2005-4M4 No.: 4
Description: **Determine how many angles are less than 90 degrees**

[Question](#) \ [Performance Data](#) \ [Content Classification](#) \ [Scoring Guide/Key](#) \ [Student Responses](#) \ [More Data](#)

[Printable Version](#)



4. How many of the angles in this triangle are smaller than a right angle?

- A) None
- B) One
- C) Two
- D) Three

Figure E-6. Example multiple choice item results from Grade 4 mathematics

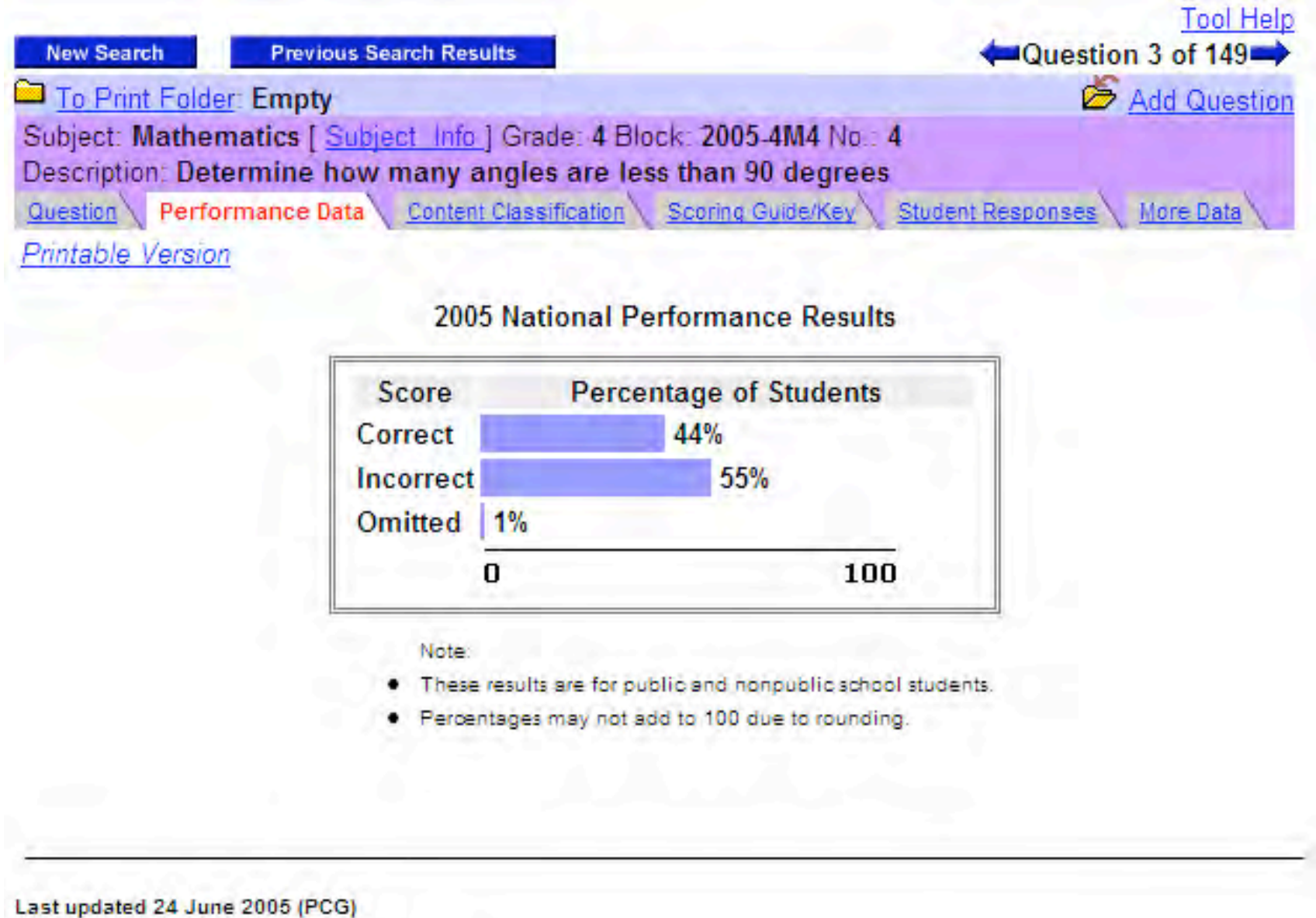


Figure E-7. Example scale score results from Grade 4 mathematics

NAEP National Mathematics Grade 4 2005 Accommodations Permitted
 Average Scale Score with Percentages (with Standard Errors in Parentheses), Mathematics
Determine how many angles are less than 90 degrees [M046701]

All students															
	A			B			C *			D			Omitted		
	Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)	
All students	234 (0.5)	13 (0.2)		233 (0.4)	26 (0.3)		249 (0.3)	44 (0.4)		221 (0.5)	15 (0.3)		228 (1.6)	1 (0.1)	

Gender															
	A			B			C *			D			Omitted		
	Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)		Avg. Score (S.E.)	Row Pct. (S.E.)	
Male	235 (0.7)	14 (0.4)		234 (0.5)	25 (0.5)		250 (0.5)	45 (0.6)		222 (0.8)	16 (0.4)		227 (2.3)	1 (0.1)	
Female	232 (0.8)	13 (0.4)		231 (0.5)	28 (0.5)		248 (0.5)	43 (0.5)		219 (0.7)	15 (0.4)		229 (2.4)	1 (0.1)	

Percentage rounds to zero.

‡ Sample size is insufficient to permit a reliable estimate.

(***) Standard error estimates cannot be accurately determined.

NOTE: The NAEP Mathematics scale ranges from 0 to 500. Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment.

Last updated 24 June 2005 (PCG)

Figures E-8 and E-9 were screen captures of another item from the NAEP Question Tool, and this item was an extended constructed-response question. In this display of the different response categories (incorrect, partial credit, correct, and omitted), the participants provided some suggestions about the layout of this table. They stated that while they wanted the information about the size of the standard errors, they found this display distracting and cumbersome in trying to compare the average scale scores for each response category. Suggestions for redesign included moving it to a second screen, dropping the below average scale scores, or graying out the information to make it less prominent. As one member of this focus group commented, “There’s just too much information.”

Appendix D: Focus of lifecycle audit

Lifecycle stages of NAEP	Developmental Goal	Validity Criteria for Meeting Goal
1. Organizational characteristics of NAEP agencies and contractors; management of NAEP Alliance (internal focus)	Maintain and improve the quality and usefulness of the assessment considering the missions of NAGB and NCES; providing leadership and oversight to the contractors who make up the NAEP alliance	<ul style="list-style-type: none"> a. Clarity of organizational roles and functions b. Clarity of the review processes c. Internal quality control procedures d. Management/communication infrastructure e. Qualifications of staff f. Documentation of security policies
2. Articulate the intended scope and uses of NAEP assessments	Develop a validity framework that is consistent with the intended uses of NAEP assessments; gather validity information to support those uses	<ul style="list-style-type: none"> a. Clarity of validity framework b. Technical reports supporting intended uses c. Articulation of unintended/inappropriate score uses and interpretations
3. Develop the Content Framework and Test Specifications	Develop a content framework and table of item specifications that reflect NAEP's intended use and scope	<ul style="list-style-type: none"> a. Evidence of alignment with intended use and scope b. Documentation of framework development and item specifications (type or format, distribution) c. Documented expert judgments that framework and item specifications are appropriate for measuring all intended groups
4. Develop Items (Test Questions) and Background Questions	Develop test items/tasks corresponding scoring guides/rubrics, and background questions, aligned with framework, table of specifications, and intended uses	<ul style="list-style-type: none"> a. Evidence of alignment with table of framework and item specifications b. Appropriate documentation of review procedures, results, and any needed improvements c. External validity evidence of link of background questions to assessing educational progress d. Documentation of security procedures for item development

Continues next page

Focus of lifecycle audit (Continued)

Lifecycle stages of NAEP	Developmental Goal	Validity Criteria for Meeting Goal
5. Create Draft Assessment, Prepare Field Design, and Conduct Field Trials	Identify and determine acceptability of new test items for assessment	<ul style="list-style-type: none"> a. Alignment of draft assessment with table of item specifications b. Documentation of all procedures (including quality control) followed in trial test assembly, sample selection, and site selection as well as post-trial review and approval procedures c. Documentation of decision rules and results for item/task review including such characteristics as item difficulty, item discrimination, differential item functioning, and/or item information along with estimation procedures d. Documentation of decision rules and results for evaluating background questions e. Documentation of security procedures for field trials and work on draft assessment.
6. Set Achievement Level Standards for the Assessment	Set performance standards for what students know and are able to do	<ul style="list-style-type: none"> a. Clear documentation for rationale and procedures used to recommend standards as well as of selection and qualification of standard-setting judges b. Relationship to internal, external, and procedural validity criteria c. Consistency/rater agreement among judges d. Documentation of security procedures for standard setting process.
7. Construct final assessment (content, design, and production)	Produce assessment instruments ready for operational administration that align with framework and table of specifications	<ul style="list-style-type: none"> a. Evidence of alignment of assessment with framework and table of specifications b. Item sampling or spiraling procedures c. Quality control requirements to ensure proper spiraling, printing, packaging, and distribution, including security of the test booklets.

Continues next page

Focus of lifecycle audit (Continued)

Lifecycle stages of NAEP	Developmental Goal	Validity Criteria for Meeting Goal
8. Sample Schools and Students	Select representative samples for reporting on relevant population groups for main, state, and Trend NAEP assessments	<ul style="list-style-type: none"> a. Population definitions and quality of frames used in multistage sampling b. Sample strata and the representation of student group samples c. Randomization procedures d. Relative bias and variability of total and group samples e. Response and participation rates f. Substitution and imputation procedures and rates h. Post-sample weighting procedures
9. Administer the Assessment	Administer the assessment to the proper students and ensure the integrity of student responses	<ul style="list-style-type: none"> a. Documented test administration procedure manuals b. Documented procedures for selection and training of administrators c. Documented quality control procedures to ensure that test administrators follow directions and documentation of their practice d. Administration security procedures e. Application of exclusion and accommodation rules
10. Score the Assessment and Prepare Final Analysis Database	Produce a database of student scores from the administered assessments for scaling and analysis	<ul style="list-style-type: none"> a. Documented procedures and results for selection and training of scorers b. Evidence of scorer reliability (for constructed response items) c. Documented quality control procedures for transport of assessments, data entry, and database preparation and delivery, including security procedures d. Data entry error rates
11. Create Scales and Links and Analyze Data	Create and use test score scales and background variables to analyze educational progress in terms of achievement status (main and state) and trends (trend) in the assessed domain for designated populations	<ul style="list-style-type: none"> a. Documentation of analysis procedures (including software, if available) b. Scale Stability across Subgroups (e.g., states, gender, race/ethnicity) c. Differential item functioning d. Test and information functions e. Score precision (standard errors)

Continues next page

Focus of lifecycle audit (Continued)

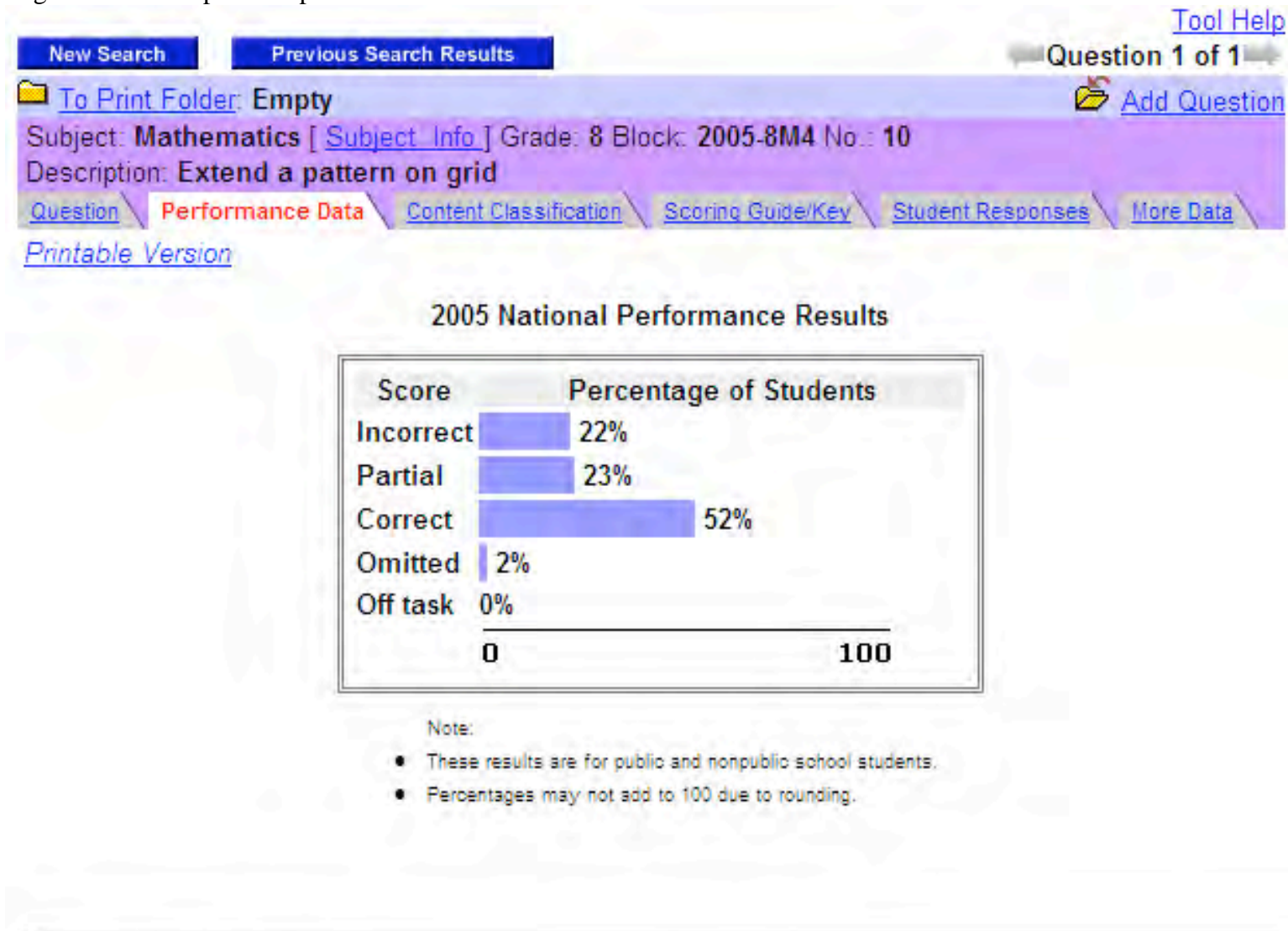
Lifecycle stages of NAEP	Developmental Goal	Validity Criteria for Meeting Goal
12. Report on Findings from the Assessment	Write, review, issue, and disseminate results from analyzing assessment scores on a timely basis	a. Timely reporting of main results (within six months) b. Timely reporting of technical analyses c. Utility of results for different audiences d. Web-evaluation —Hit-rates and skip patterns —Record of downloads —Quality of interactive tools —Responsiveness to requests e. Customer satisfaction j. Customer services (e.g., training, provision of databases)
13. Renew and Improve the Assessment (External Focus)	Maintain and improve the quality and usefulness of the assessment in the context of education change and reform and improvements in psychometrics and other testing methodologies	a. Rate of innovations in assessments b. Comparisons to other state and assessments c. Documented quality control initiatives and agencies/contractor responses to those initiatives

Appendix E: NAEP responsibilities matrix

	NCES	NAGB	ETS	Westat	AIR	PEM	HumRRO	Hager Sharp	ACT (Pacific Metrics)	GMRI	State Coordinators
1. Organizational Characteristics	X	X	X	X	X	X	X	X		X	X
2. Specify intended scope and uses of NAEP assessments	X*	X*	X		X	X					X
3. Develop assessment framework and test and background item specifications		X*									
4. Develop items and background questions	X	X*	X*		X*						
5. Create draft assessment, prepare field design and conduct field trials			X	X							
6. Set achievement level standards		X							X		
7. Construct final assessment and field design			X			X					
8. Sample schools and students			X	X*							
9. Administer the assessment				X		X					X
10. Score the assessment and prepare final analysis database			X			X					
11. Create scales and links and analyze data			X*		X		X				
12. Write, review, issue, and disseminate reports and data	X	X	X					X		X	X
13. Renew and improve the assessment	X	X	X	X	X	X	X	X			X

* Denotes primary responsibility

Figure E-8. Example multiple choice item results from Grade 8 mathematics



Last updated 24 June 2005 (PCG)

Figure E-9. Example scale score results from Grade 8 mathematics

NAEP National Mathematics Grade 8 2005 Accommodations Permitted
 Average Scale Score with Percentages (with Standard Errors in Parentheses), Mathematics
Extend a pattern on grid [M066601]

All students														
	Incorrect				Partial				Correct				Omitted	
	Avg. Score (S.E.)	Row Pct. (S.E.)			Avg. Score (S.E.)	Row Pct. (S.E.)			Avg. Score (S.E.)	Row Pct. (S.E.)			Avg. Score (S.E.)	Row Pct. (S.E.)
All students	263 (0.6)	22 (0.3)			270 (0.5)	23 (0.4)			292 (0.3)	52 (0.4)			229 (1.8)	2 (0.1)

Percentage rounds to zero.

‡ Sample size is insufficient to permit a reliable estimate.

(***) Standard error estimates cannot be accurately determined.

NOTE: The NAEP Mathematics scale ranges from 0 to 500. Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment.

Figure E-10 was a version of the “pantyhose” that has been used to allow comparisons between different reporting jurisdictions. The data in Figure E-10 was cross-district comparisons of average grade 4 mathematics score scores from the Trial Urban District Assessment (TUDA) of 2005. The participants were able to identify that Charlotte and Austin were likely to be pleased by the results reported in this figure. An arrow pointing up encased by a lightly shaded square was known by participants to indicate that the district in whose row that notation was located did better than the district listed at the top of the column, but participants also identified a problem with the key for this graph. There were three notations used in this pantyhose chart: In this key, while the blank square indicating no differences included the phrase “statistically significant,” neither the higher nor lower notation included this phrase. The group surmised that this was information they were supposed to infer, but they agreed that this represented a potential source of confusion for users of what they considered to be a useful display of data. They also inquired as to the meaning of a “large central city” and thought it would be helpful to have additional material available on why certain other districts were not included in the comparisons.

Figure E-10.

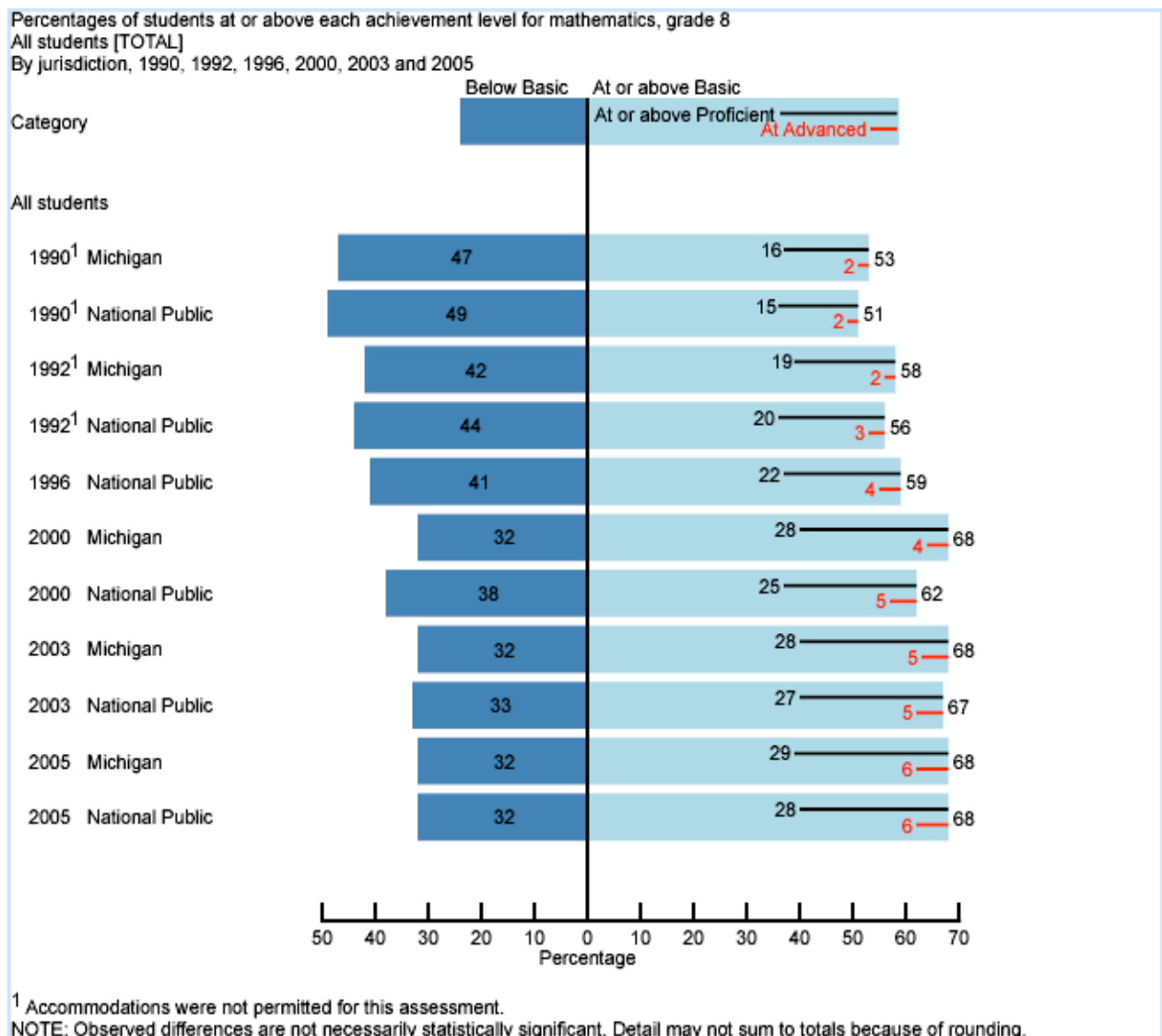
Overall cross-district comparisons of average mathematics scale scores, grade 4 public schools: 2005



Another suggestion made by a participant for future use of the panty hose chart would be to make it interactive and allow users to choose which jurisdictions to include for comparisons displayed in this way. For example, one participant thought that this would be a nice way to show comparisons among states regionally (e.g., the Midwest or the Southeast).

In Figure E-11, participants saw the percentage of students at or above each achievement level for mathematics grade 8 for National Public students and Michigan students from 1990 to 2005 were displayed as a sequence of bars on a bar chart, with the bars aligned by the division between the below Basic and Basic achievement levels. The participants found it disconcerting that the repeating pattern of Michigan results followed by National Public results was broken up in 1996 for whatever reason, although they eventually figured out the percent of students in each achievement level without difficulty. When asked for broader interpretations of Michigan’s performance over time, while most indicated that the 15-year trend was of improvement, one member of this group noted that in the past five years or so, the trend was largely flat, and especially of note was no change in the number of students below Basic, a pattern that that focus group member said was consistent with what was happening with that person’s own state.

Figure E-11.



The Figure E-12 display shown to participants in this focus group was a grade 4 item map from 2005. While participants were not asked questions about this display due to time limitations in the focus group meeting, this display was shown to them to provide information following up on questions by participants asked earlier when Figure E-2 was shown. By and large, this group was unfamiliar with item maps, and they were eager to learn more about the meaning that could be associated with scores by illustrative items. Perhaps as mathematics educators, they were particularly interested in the idea of response probabilities and likelihood curves being associated with test items and using that information to aid them in making test score interpretations. But the reasons for locating items on the scale and what interpretations were to be made from the item maps were not known to participants.

Figure E-12.



¹ Each grade 4 mathematics question in the 2005 mathematics assessment was mapped onto the NAEP 0-500 mathematics scale. The position of a question on the scale represents the average scale score attained by students who had a 65 percent probability of successfully answering a constructed-response question, or a 74 percent probability of correctly answering a four-option multiple-choice question. Only selected questions are presented. Scale score ranges for mathematics achievement levels are referenced on the map. For constructed-response questions, the question description represents students' performance rated as completely correct.

NOTE: Regular type denotes a constructed-response question. Italic type denotes a multiple-choice question.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment.

Discussion

Several important themes were identified in the course of conducting the focus group. The focus group setting was conducive to generating discussion among participants and with researchers from UMass. Furthermore, this study indicated that with respect to some materials for audiences there may be a need to revise the NAEP score reports to make them more user-friendly or a need for more explanatory materials for persons using the NAEP reports. The participants in this group expressed a clear preference for NAEP data displays that allowed for them to make quick visual evaluations of data (such as the clickable state-by-state comparison maps now available on the NAEP Web site and the version of the pantyhose chart shown in the meeting). When available, these math educators sought to use display keys and footnotes to find information and make interpretations, but in several situations they became bogged down by minor layout and design issues. Below we provide some of the primary findings from this focus group and some directions that may be helpful for NAEP reporting as it continues to evolve with respect to the content of displays and how results data are disseminated to different audiences.

First, while participants indicated they worked with NAEP data with some regularity, the low familiarity with some of the statistics and reporting methods shown in this focus group indicate there are a number of elements of NAEP displays that may be problematic for educators. The idea of scale scores is one such element: While many educators have some knowledge of their own state's reporting scale for state assessments, and many are likewise familiar with SAT or ACT scales, the NAEP scale was not well known among participants in this group. Particularly of interest to these participants is giving meaning to (1) different points on the score scale using tools such as item maps and distracter or response category analyses and (2) average scale score differences observed between groups (e.g., states/jurisdictions, reporting groups).

In the first case, the educators involved in this group were excited to learn about and use the NAEP item maps, and felt that using exemplar items in this way was extremely valuable in helping them to make an abstract idea (a scale score) relevant and logical. The idea of looking at a single item and knowing the scale scores associated with different response categories (constructed-response) and answer choices (multiple choice) was also seen as valuable information, which helped quantify differences between different points on the score scale. Participants sought out information to help them understand what (for example) a 1-, 5-, or 10-point difference in scores really *meant* with respect to student performance and were interested in using items as to enrich the interpretations they wanted to make. Clearly, the use of item maps and their interpretations should be addressed in subsequent up-dates of the Web sites and training opportunities.

Another element that seemed to be a source of difficulty was layout of different aspects of some of the displays. Perhaps unexpectedly, as a group the participants in this focus group paid considerable attention to footnotes and keys to figures. They were very clear in communicating to the researchers that when they found these elements of displays to be confusing in some way (not consistent in labeling, poorly laid out), their understanding of the data displays was curtailed. They were particularly frustrated by aspects of some of the figure keys and the screen-capture of the clickable state maps. One recommendation, therefore, for NAEP reporting practices, is to ensure that footnotes and keys are complete, comprehensible, and consistent to the extent possible across different displays.

A related layout point concerned the inclusion of standard error and percent of student information in the distracter or response category tabs from the NAEP Question Tool. While these users of NAEP data reiterated interest in knowing what the standard errors were (although their understanding of these was uneven across the group) they wanted to focus on differences in scale scores for different score and response categories, and as laid out now (Figure E-7), they had to look across and keep track of four to five columns of scale scores, each separated from one another by three other columns of data. There are different potential ways to reformat this information, perhaps by rearranging data and adding shading to allow users to quickly match up response categories and data from different rows that they might wish to compare.

While the score reporting issues that were raised via this particular focus group were not comprehensive in scope, the findings discussed here provide important information about the experience of some users of NAEP data with respect to understanding and use of the results. Among educators and educational administrators such as those participating in this group, there seem to be two types of knowledge that help these users work with the NAEP data displays included in this study: First, a broad familiarity with test scores and the jargon of assessment (e.g., standard errors, scale scores, etc.), and second, a familiarity with common NAEP terms and reporting mechanisms (e.g., “at or above,” the NAEP achievement levels, the interactive online tools). This is an important consideration for a testing program such as NAEP, in which the audiences for the data and data products differ widely and even within an audience (“teachers,” “the public”) the range of interest and comprehension varies. Providing background information targeted to different groups, such as the “Educators” link on the NAEP home page in the “About NAEP” menu, is an important step in promoting use of assessment information in user-friendly ways.

Next Steps

There are several important directions for follow-up research. First, while the use of group discussion clearly yielded much useful information, a logical next step is to carry out one-on-one explorations of several data displays with NAEP data users as they navigate themselves through different portions of the NAEP Web site. In this way, we could gather more information about how different individuals fare with respect to both knowledge and interpretation of several of the more interactive features of the site, including the clickable state maps, the Question Tool, and the NAEP Data Explorer.

Second, additional focus groups of this nature with other interested audiences of NAEP data would be useful. Other groups to consider include another group of educators (for replication purposes), media representatives, state or federal policymakers or legislative aides, and the general public.

A third extension of this research would be the development of several redesigns of the displays shown here, with research participants comparing current and revised displays for clarity and understanding. This idea was pursued by Wainer, Hambleton, and Meara (1999) and produced some interesting findings. While NAEP is clearly at the forefront of testing programs with respect to its investment in methods for disseminating results, the results of this focus group indicated that there remain some sources of confusion among audiences who have some familiarity and regular use of NAEP. Clearly substantially more research and development work is needed in the near future.

References

- Goodman, D. P., and Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17, 145–220.
- Hambleton, R.K., and Slater, S.C. (1996). Are NAEP executive summary reports understandable to policy makers and educators? Paper presented at the meeting of the National Council on Measurement in Education, New York.
- Herman, J., and Gribbons, B. (2001). *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation* (CSE Technical Report 535). Los Angeles: UCLA Center for the Study of Evaluation.
- Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress (NAEP Validity Studies Panel Report)*. Palo Alto, Calif.: American Institutes for Research.
- Sharkey, N. S. and Murnane, R. J. (2003). Learning from student assessment results. *Educational Leadership*, 61, 77–81.
- Wainer, H., Hambleton, R.K., and Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335.

This page left intentionally blank

Appendix F: State Reading Content Specialists and NAEP Data Displays

April L. Zenisky, Jill Delton, and Ronald K. Hambleton

Center for Educational Assessment
University of Massachusetts Amherst.

Abstract

The National Assessment of Educational Progress (NAEP) reports results in a variety of subject areas and grade levels, and uses a range of data display methods to communicate results (including line graphs, bar charts, pantyhose charts, clickable state maps, etc.). To evaluate the use and understanding of several of the common strategies for communicating NAEP results, a focus group of state reading content specialists was convened to discuss several current NAEP Reading displays. The group, consisting of eight Reading specialists from multiple states, reviewed 15 NAEP data displays from recent reports, the NAEP Question Tool, and the NAEP Data Explorer and discussed their understandings and impressions. Findings include the need for clarification in footnotes, legends, and keys, as well as simplification of some displays to minimize clutter. In addition, participants sought additional information about the practical meaning of some displays, especially when significant test results were present. Future directions for research include the use of one-on-one conversations with users of the NAEP Web site about selected data displays and using suggestions from the focus group to redesign some displays for tryout with focus groups.

Introduction

Results from the National Assessment of Educational Progress (NAEP) are significant indicators of U.S. students' performance in academic subject areas and are of interest to many audiences including educators, the general public, legislators, and state education officials. However, reporting test results to any stakeholder group is challenging because of the need to consider the density and accuracy of the information to be communicated. This is certainly true with respect to NAEP (Jaeger, 1998; Simmons and Mwalimu, 2000).

Among the many audiences to whom NAEP results are important are educators and state education personnel. To the extent that NAEP can be used as one of multiple gauges of student achievement, NAEP reports should be designed and developed with those individuals in mind. In fact, Jaeger (1998) suggested that NAEP reports focus on the dimensions of *use* and *understanding*. *Use* centers on how different users of NAEP results access information and how they use that information. The *understanding* dimension involves how results are represented and understood.

The purpose of this study was to explore the extent to which educators (in this case, state reading content specialists) were familiar with current methods of displaying NAEP reading results and what kinds of inferences they might make on the basis of those displays. This study is an important part of a broader evaluation of the utility of score reporting methods used by NAEP in gathering information as to how members of different audiences both use and understand NAEP findings. Guiding this study are questions as to 1) how NAEP results are displayed, particularly in electronic communications with respect to principles of good reporting (e.g., Goodman and Hambleton, 2004), 2) what are the ways in which users understand and do not understand the data presented, and 3) the development of alternative displays that may alleviate misunderstandings and misconceptions when they exist. This report also serves as a complement to the evaluation study presented in Appendix E in which the participants of interest were state mathematics content specialists and the displays discussed were similar but reflective of Mathematics results.

Method

A focus group of state reading content specialists was convened to gain information about the meaningfulness of different NAEP data displays.

Participants

The eight participants in this focus group, all reading curriculum specialists from various states attending a NAEP reading item review meeting in Bethesda, Md., were drawn from a list of 13 reading content specialists attending the Baltimore meeting. All agreed to participate in a two-hour focus group in the evening after the conclusion of the item review discussions for the day (May 24, 2006). In return for their participation in the focus group, the participants were provided with dinner and a small honorarium of \$150.³⁸

The eight participants represented seven different states. Based on the Census Bureau reporting regions used by NAEP, three of the participants were from the West (one from Arizona and two from Idaho), three were from the South (Delaware, Virginia, and Florida), one was from the Midwest (Illinois), and one was from the Northeast (Connecticut). There were seven females and one male. All of the participants had at least 11–20 years of experience in the field of education and three had more than 20 years. There were three English language arts coordinators, two education program specialists, one elementary reading specialist, one education consultant, and one director of secondary reading. All participants had a background in reading.

All focus group participants were familiar with NAEP and three reported they were very familiar. The participants had varying degrees of experience working with NAEP data including

³⁸ Three of the eight declined the honorarium because of requirements imposed on them by their state employer.

a couple of times a year (two participants), several times a month (two participants), and weekly (three participants). Only one had no prior experience working with NAEP data.

Most of the participants who had worked with NAEP data in the past had conducted item reviews and studied trends. Several had also compared NAEP data with their own state's curriculum or frameworks, while others reported working with educators, administrators, and NAEP coordinators to interpret and analyze data. One of the participants also reported using the tools on the NAEP Web site for professional development.

Procedure

At the focus group meeting, participants were provided with an overview of the project and asked to complete a brief demographic survey. Next, a series of data displays consisting of both tables and figures from several recent releases of NAEP reading results (National and State results, 2005; Trial Urban District Assessment, 2005; and the NAEP Question Tool) were projected on a screen in color (and also given to participants as full-page handouts in black-and-white). We focused on displays of Reading results to increase interest among participants. As each figure or table was displayed, participants were asked questions about the displays. The displays shown at this meeting were chosen as a sampling of the types of tables and figures seen throughout recent NAEP reports, including:

- line graphs,
- stacked bar charts,
- clickable state comparison maps of average scale scores and percents of students at or above achievement levels,
- tabs from the NAEP question tool with item text, student item performance, a distracter analysis,
- “pantyhose” charts,
- ways of displaying score gaps between reporting subgroups,
- bar graphs, and
- item maps.

As each display was projected on the screen, participants were asked to reflect on each display for a few minutes, and then they were asked questions about the data display by one of the two meeting facilitators. Questions ranged from those that were informational in nature (“What was the average score for eighth graders in 2005 in reading?”) to opinion (“What, if anything, do you find confusing or not clear about this display?”). The focus group discussion format was appropriate for this study because this format served to stimulate some broader conversations among the participants and facilitators about the data displays, building on what was being displayed on the screen, and allowed the participants to answer some of the more difficult data interpretation questions collaboratively.

The last task asked of participants in the meeting was for them to respond to a sequence of discussion questions that focused on broader issues of score reporting and NAEP. These questions included reflection on the collection of displays presented throughout the meeting, their preference for receiving information themselves and how educators want information, and ways of representing gaps in subgroup performance.

Results

Figure F-1 displays a line graph of average reading scale scores showing the performance of grade 4 students nationally from 1992 to 2005. The graph displays the average scale score for each year that the assessment was given. An “*” was placed next to the scores for the years in which the difference between that year and 2005 was found to be statistically significant. A dashed line is used to represent the years in which accommodations were permitted. All of the panelists understood that an “*” next to a score meant that score was statistically different from the 2005 scores. However, there was some confusion about NAEP reporting it as “statistically

different” as opposed to statistically significant (one participant commented, “The graph says statistically different, is that the same as statistically significant?”). The participants easily recognized that the dashed line meant administrations for which accommodations were permitted.

Figure F-1.

Average reading scale scores, grade 4: Various years, 1992-2005

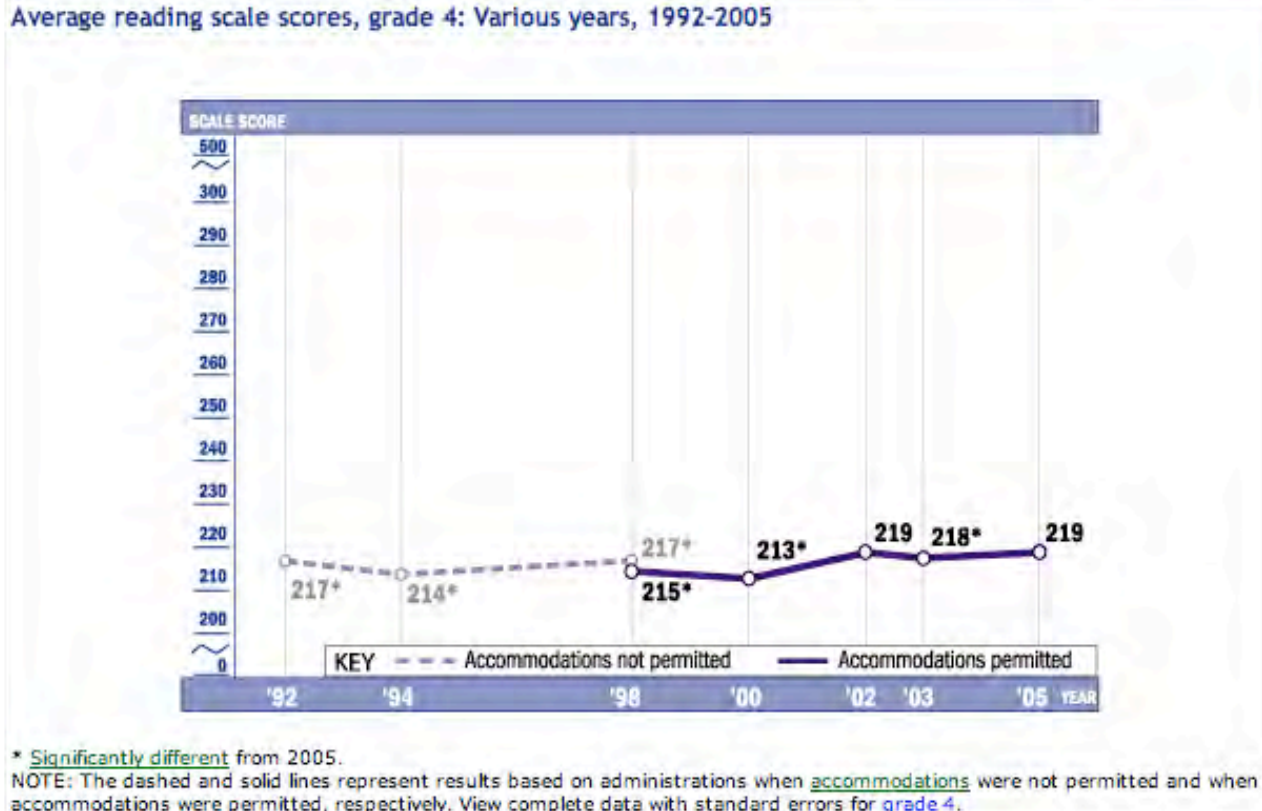
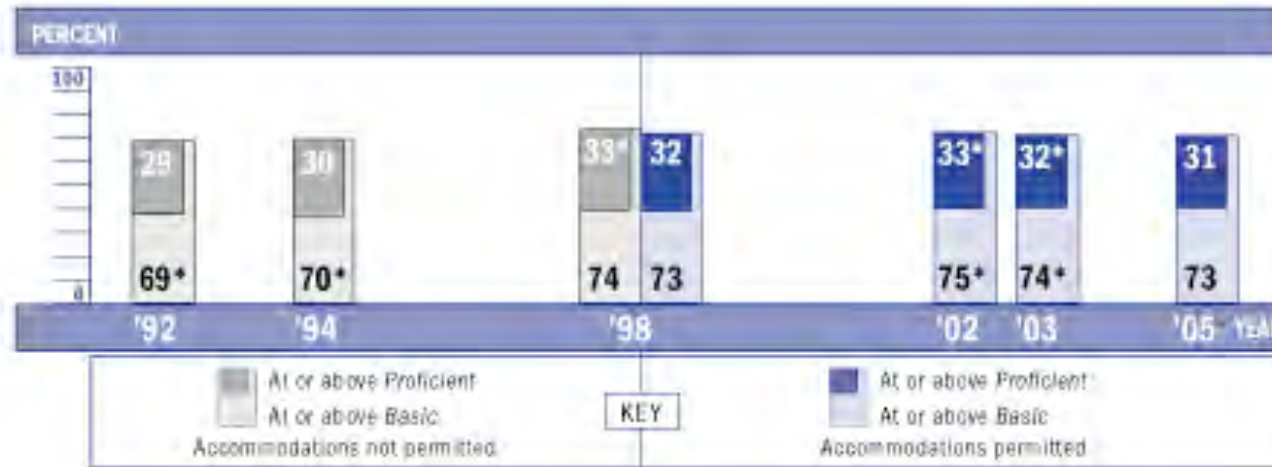


Figure F-2 was a stacked bar graph displaying the percentage of grade 8 students at or above *Basic* and at or above *Proficient* in reading from 1992 to 2005. Only a few of the participants reported seeing something like this before. All of the participants reported they were familiar with the NAEP achievement levels, but they were a little confused about NAEP reporting the percent “at or above” the achievement levels. When asked what it meant to be at or above *Proficient*, panelists responded with comments, such as, “It’s a range of scores,” and “It includes *Basic*, *Proficient*, and *Advanced*.” Although the participants may not have been entirely clear on what it mean to be “at or above,” they were able to correctly identify that 73 percent of the students were at or above *Basic* in 2005. They also realized that this meant that 27 percent were below *Basic*. When asked about the change between 1998 and 2005, participants reported “at or above *Proficient* went down by 1 percent.” Some of the participants did not understand why there were two bars for 1998, but other participants realized that one was with accommodations and one without.

Overall, the group found this graph to be difficult to read. They thought the box inside the bar was a very confusing way to display the data. Panelists commented, “It’s too much information in one place,” and “The box inside the bar is confusing.” It was suggested that it would be easier to read if the two bars were side by side rather than one inside the other. The group also thought it would be less confusing if the percentage of students below *Basic* was displayed somewhere on the graph rather than having to subtract to obtain it.

Figure F-2.

Percentage of students at or above *Basic* and at or above *Proficient* in reading, grade 8: Various years, 1992-2005



* Significantly different from 2005.

NOTE: The gray shaded boxes represent results based on administrations when accommodations were not permitted. View complete data with standard errors for grade 8.

The next three figures (Figures F-3a, F-3b, and F-4), were all examples of clickable state comparison maps. This interactive tool on the NAEP Web site allows the user to select a focal state or jurisdiction in which to compare the other states to. The map is color-coded to reflect how each state compares to the selected focal state or jurisdiction. The user also selects if the comparisons are to be based on either scale scores or at or above *Proficient*.

Figure F-3a was a screen capture of a clickable state comparison map of average reading scale scores in 2005 for grade 4 public schools. In this particular figure, the focal group selected was Oregon. Most of the participants were familiar with this type of figure and they were able to identify the focal state quickly. They also recognized that the comparisons being made were based on average scale score. However, there was some uncertainty as to whether the differences in scale scores were significant because the key did not specify at which point the differences were significant, only when they were not (one panelist commented, “I would assume the darker green is significant because the lime green is labeled ‘not significantly different’”).

Overall, the group thought this was a good way to convey information for quick comparisons across states. There were a few things that the panelists thought could have been clearer, however. One panelist commented, “I don’t like that the focal state is a total different color than the states that have similar scores,” and another stated, “The legend could be clearer.” Another panelist wanted to know what made Oregon the focal state, and they were informed that the Web site allows the user to select any state as the focal state.

In Figure F-3b, another capture of a clickable state comparison map, the percentage of students at or above *Proficient* in reading in 2005 for grade 4 public schools was compared. The focal group for this map was again Oregon. The group was aware that although the focal group was Oregon for each of these figures, they were comparing different things. They recognized that this map was showing percentage of students at or above *Proficient*. Comments were also made about how the second map shows a more favorable image than the first one (e.g., “There are fewer dark green states,” and “Oregon can choose to report this map because it looks better than the last one.”).

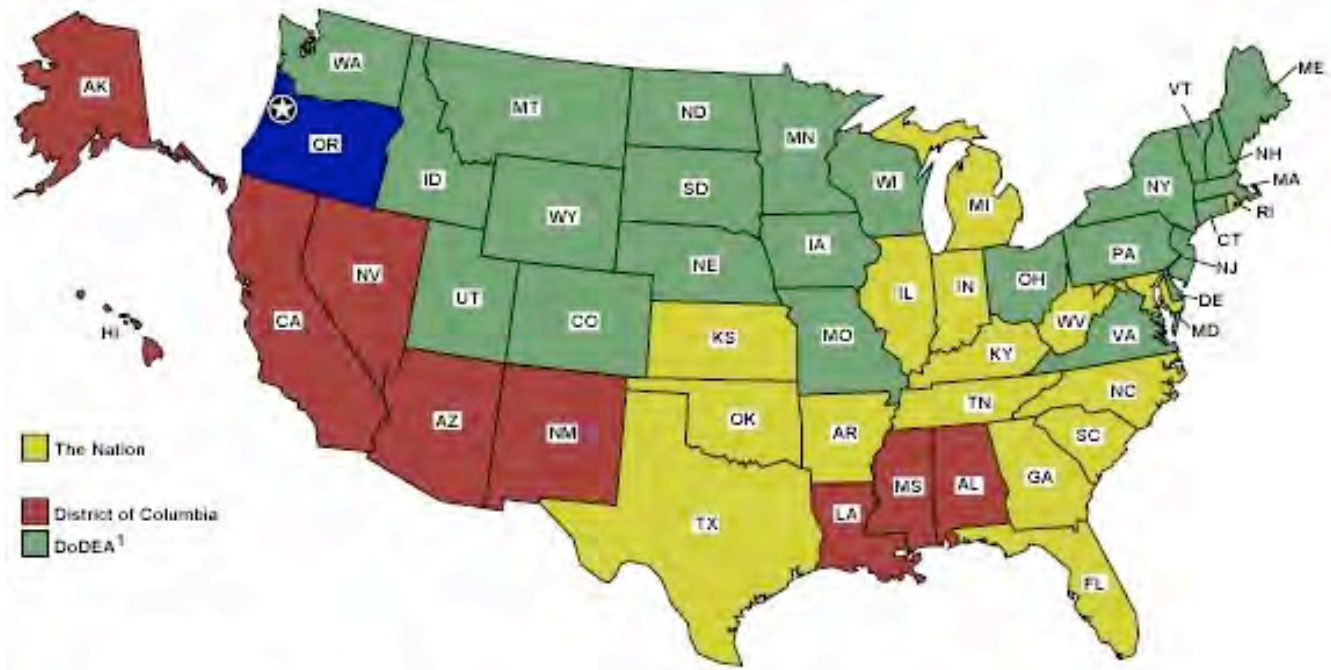
Figure F-4 was also a capture of a clickable state comparison map of state comparisons of average reading scale scores in 2005 for grade 8 public schools. The focal group for this map was the nation. The group identified the focal group very quickly but said they found this one to be more confusing than the other two maps. When asked to describe the results being displayed, the panelists reported, “It looks like the nation is in the middle,” and “It seems like there’s a divide between the scores in the northern and southern parts of the country.”

Figure F-3a.

Cross-state comparisons of average reading scale scores, grade 4 public schools: 2005

NAEP Reading Grade 4 - Reading
 Difference in Average Scale Score Between Jurisdictions
 for All students [TOTAL] = All students
 2005

Color

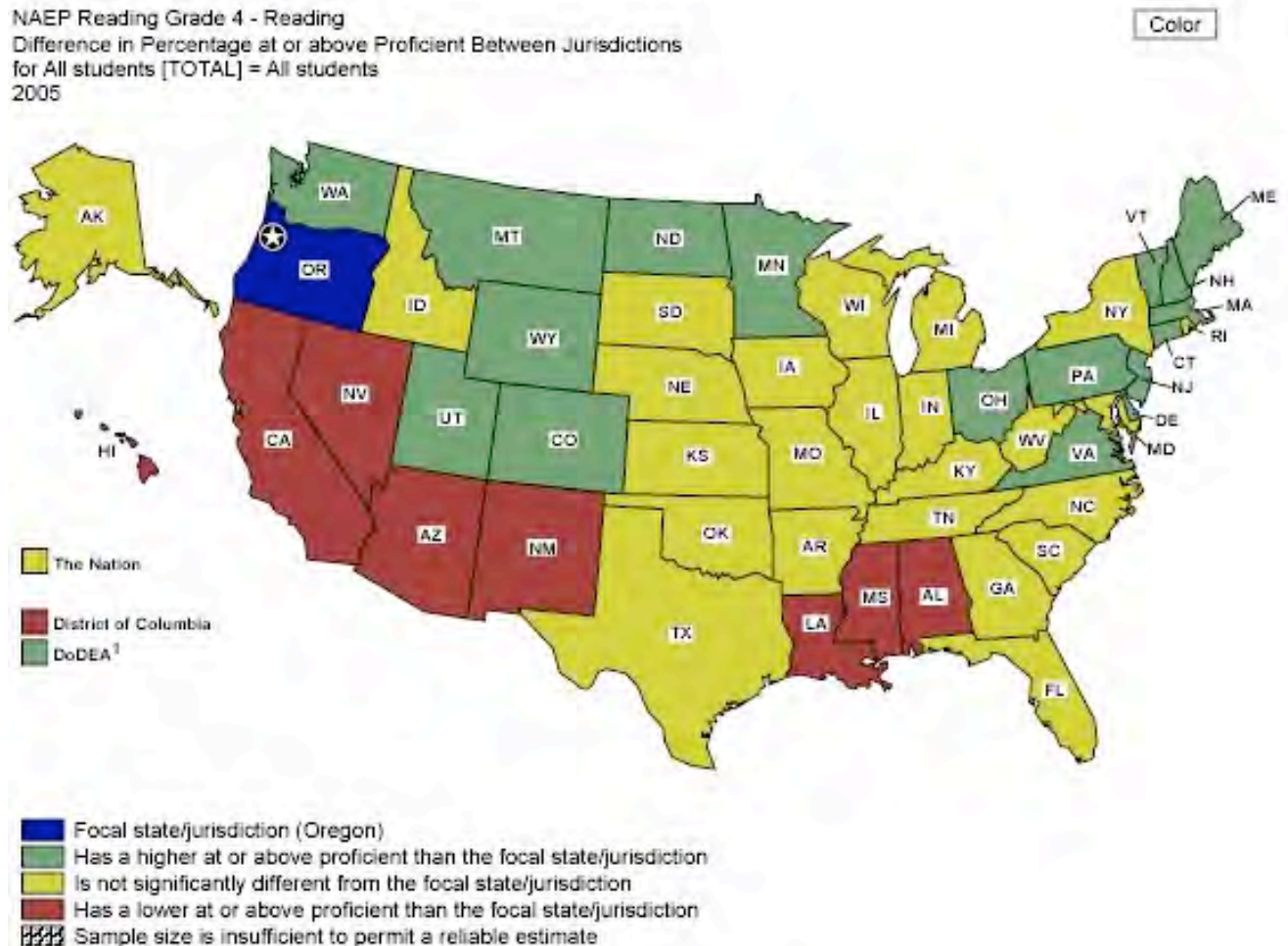


- Focal state/jurisdiction (Oregon)
- Has a higher average scale score than the focal state/jurisdiction
- Is not significantly different from the focal state/jurisdiction
- Has a lower average scale score than the focal state/jurisdiction
- Sample size is insufficient to permit a reliable estimate

¹ Department of Defense Education Activity schools (domestic and overseas).

Figure F-3b.

Cross-state comparisons of percentage of students at or above *Proficient* in reading, grade 4 public schools: 2005



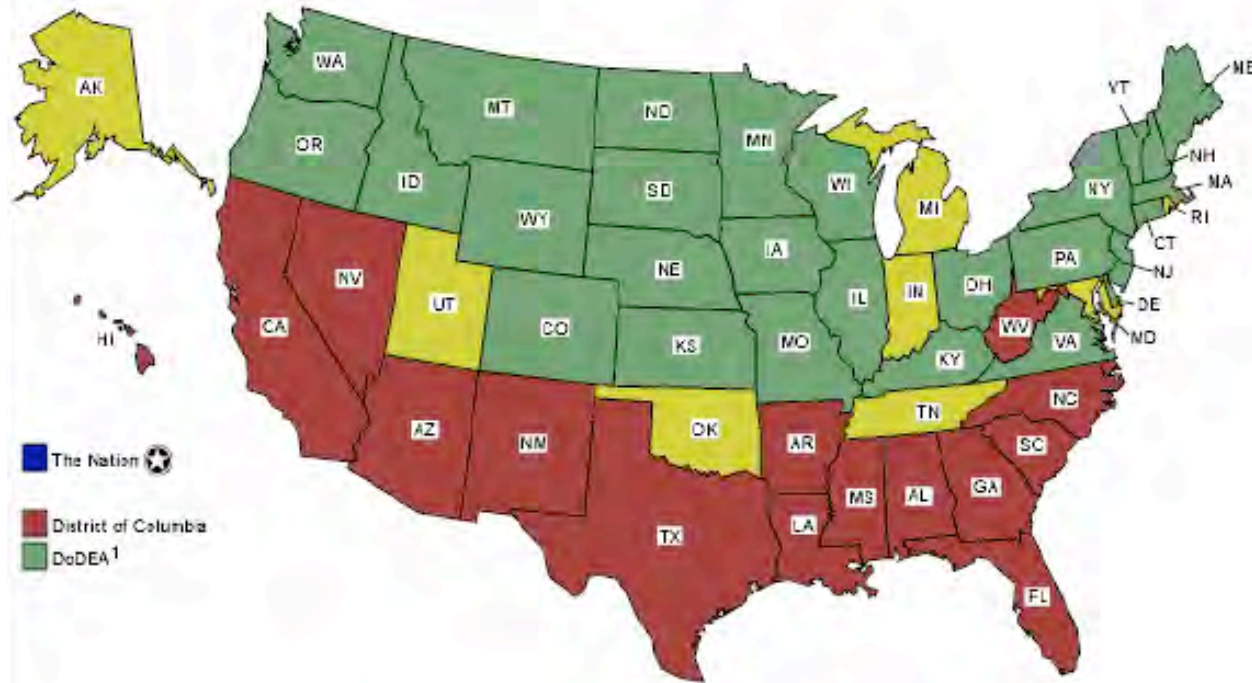
¹ Department of Defense Education Activity schools (domestic and overseas).

Figure F-4.

Cross-state comparisons of average reading scale scores, grade 8 public schools: 2005

NAEP Reading Grade 8 - Reading
 Difference in Average Scale Score Between Jurisdictions
 for All students [TOTAL] = All students
 2005

Color



- Focal state/jurisdiction (National Public)
- Has a higher average scale score than the focal state/jurisdiction
- Is not significantly different from the focal state/jurisdiction
- Has a lower average scale score than the focal state/jurisdiction
- Sample size is insufficient to permit a reliable estimate

¹ Department of Defense Education Activity schools (domestic and overseas).

NOTE: View complete data with standard errors for [grade 8](#).

Figures F-5a and F-5b were included to open discussion about the participants' use of the NAEP Question Tool, located on the NAEP Web site at <http://nces.ed.gov/nationsreportcard/itmrls/>. Figure F-5a included a passage on blue crabs from the NAEP Grade 4 Reading assessment of 1998, and Figure F-5b was one of the multiple-choice questions associated with that passage. The group was first asked about their familiarity with and previous usage of the Question Tool. Four of the reading specialists in the group indicated they had used the tool before. When those individuals were asked to describe their activities, professional development of their states' teachers was cited by two of the specialists. As noted by the specialist from one state: "We have the teachers estimate how many kids they think will answer the questions right, then show them the actual percentage." Another participant reported that because her state assessment program does not release questions, teachers are encouraged to use NAEP passages to prepare students for the state test and to make do when necessary for grade levels not tested by NAEP, such as using *Advanced* fourth grade questions for the fifth-graders, and *Basic* eighth grade items for the sixth-graders. This individual also reported using the NAEP item map to identify "*Advanced*" or "*Basic*" items. Another use of the NAEP question tool was to provide additional writing prompts for classroom activities.

Figure F-5a. Example grade 4 reading passage

NAEP Questions
The Nation's Report Card (home)

[Tool Help](#)

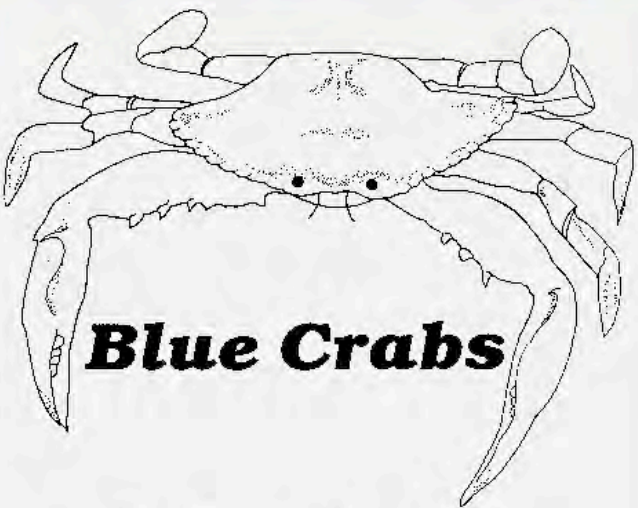
[New Search](#) [Previous Search Results](#) [← Question 26 of 36 →](#)

[To Print Folder: Empty](#) [Add Question](#)

Subject: [Reading](#) [[Subject Info](#)] Grade: 4 Block: 1998-4R6 No. 2
Description: **Blue Crabs: Common with arthropods-skeleton**

[Question](#) [Performance Data](#) [Content Classification](#) [Scoring Guide/Key](#) [Student Responses](#) [More Data](#)

[Printable Version](#) [Return to Question](#)



Blue Crabs

By George W. Frame

Nearly every day last summer my nephew Keith and I went crabbing in a creek on the New Jersey coast. We used a wire trap baited with scraps of fish and meat. Each time a crab entered the trap to eat, we pulled the doors closed. We cooked and ate the crabs we caught.

Blue crabs are very strong. Their big claws can make a painful pinch. When cornered, the crabs boldly defend themselves. They wave their outstretched claws and are fast and ready to fight. Keith and I had to be very careful to avoid having

Figure F-5b. Example multiple choice item grade 4 reading

The screenshot shows the NAEP Questions website interface. At the top, there are navigation menus for 'ABOUT NAEP...', 'SUBJECT AREAS...', 'HELP', 'SITE MAP', 'CONTACT US', 'GLOSSARY', and 'NEWSFLASH'. Below these are links for 'SAMPLE QUESTIONS', 'ANALYZE DATA', 'STATE PROFILES', and 'PUBLICATIONS'. The main heading is 'NAEP Questions' with a subtitle 'The Nation's Report Card (home)'. A navigation bar indicates 'Question 26 of 36' and includes 'Tool Help', 'New Search', and 'Previous Search Results' buttons. The question details are: 'To Print Folder: Empty', 'Subject: Reading [Subject Info]', 'Grade: 4 Block: 1998-4R6 No.: 2', and 'Description: Blue Crabs: Common with arthropods-skeleton'. There are links for 'Question', 'Performance Data', 'Content Classification', 'Scoring Guide/Key', 'Student Responses', and 'More Data'. A 'Printable Version' link is also present. The question text is: '2. According to the passage, what do blue crabs have in common with all other arthropods?'. The options are: 'A) They have a skeleton on the outside of their bodies.', 'B) They hatch out of a shell-like pod.', 'C) They live in the shallow waters of North America.', and 'D) They are delicious to eat.' At the bottom left, it says 'Last updated 24 June 2005 (PCG)'.

ABOUT NAEP... SUBJECT AREAS... HELP SITE MAP CONTACT US GLOSSARY NEWSFLASH

SAMPLE QUESTIONS ANALYZE DATA STATE PROFILES PUBLICATIONS

NAEP Questions

The Nation's Report Card (home)

Tool Help

← Question 26 of 36 →

New Search Previous Search Results

To Print Folder: Empty Add Question

Subject: Reading [Subject Info] Grade: 4 Block: 1998-4R6 No.: 2

Description: Blue Crabs: Common with arthropods-skeleton

Question Performance Data Content Classification Scoring Guide/Key Student Responses More Data

Printable Version View Reading Passage

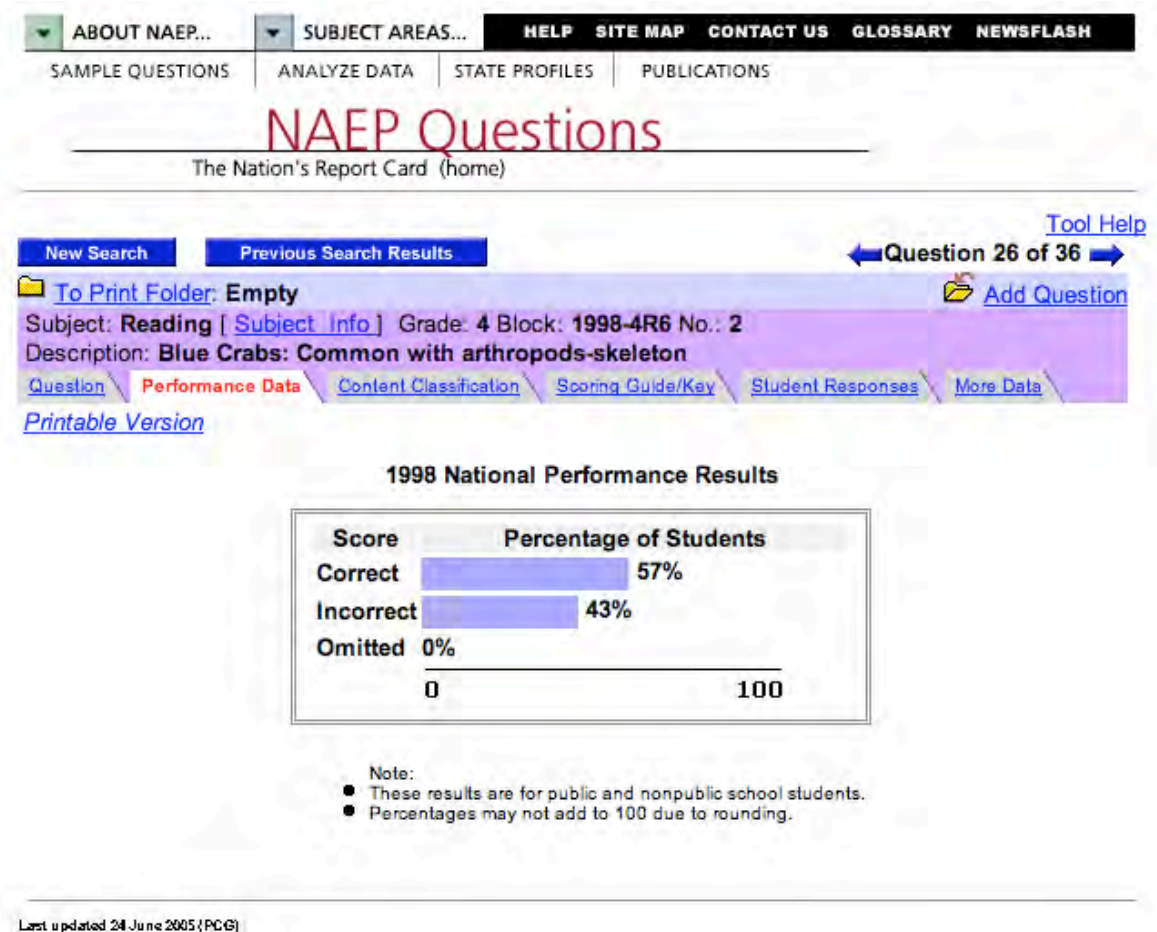
2. According to the passage, what do blue crabs have in common with all other arthropods?

A) They have a skeleton on the outside of their bodies.
 B) They hatch out of a shell-like pod.
 C) They live in the shallow waters of North America.
 D) They are delicious to eat.

Last updated 24 June 2005 (PCG)

The first results-oriented display associated with the Question Tool was the performance data (Figure F-6). Participants easily identified the percent of students answering the item correct and incorrect, and when asked how difficult this item was, responses included “ok,” “not too bad,” and “more or less medium difficulty.” When asked to explain “not too bad,” one specialist indicated that it wasn’t that hard, and “as a rule of thumb you need a certain percent of hard, easy, and medium difficulty when making a test.”

Figure F-6. Example multiple choice item results from grade 4 reading



Moving on to Figure F-7, which was a capture of part of the distracter analysis provided by one of the tabs in the Question Tool, most participants said they had seen that kind of break-out by response option before, mostly on their state tests but sometimes during NAEP item reviews. One reported using a similar breakdown for each item to look for “bias, like gender differences.” They recognized A, B, C, D, Omitted, and Missing as the different response options and ways categorizing nonresponses, and the “average score” on the display associated with each response category was the “average score of the people who chose that answer.” One participant said “The higher average score for students choosing Option A tells us something about those people, while 199 for the Option C people tells us something about them.” As a group, they liked being provided with this information, particularly the average scale score for each response choice (“You don’t always see that,”) and generally found it helpful. Uses they mentioned for these data included choosing items and looking for bias (gender, ethnicity, free or reduced-price lunch, etc.).

Figure F-7. Example scale score results from grade 4 reading

NAEP National Reading Grade 4 1998 Accommodations Not Permitted
 Average Scale Score with Percentages (with Standard Errors in Parentheses), Reading
Blue Crabs: Common with arthropods-skeleton [R012202]

All students													
	A *		B		C		D		Omitted			Missing	
	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	
All students	230 (1.2)	57 (1.4)	205 (2.2)	19 (1.1)	199 (3.4)	10 (0.8)	203 (2.7)	14 (1.0)	‡ (‡)	# (0.1)	‡ (‡)	1 (0.2)	
Gender													
	A *		B		C		D		Omitted			Missing	
	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	Avg. Score (S.E.)	Row Pct. (S.E.)	
Male	227 (1.6)	59 (1.9)	203 (3.6)	19 (1.6)	192 (5.2)	10 (1.1)	200 (4.1)	13 (1.4)	‡ (‡)	# (***)	‡ (‡)	1 (0.4)	
Female	233 (1.6)	55 (1.7)	207 (2.6)	20 (1.3)	205 (4.1)	9 (1.1)	206 (3.2)	15 (1.2)	‡ (‡)	# (0.2)	‡ (‡)	1 (0.3)	

^ Accommodations were not permitted for this assessment.

Percentage rounds to zero.

‡ Sample size is insufficient to permit a reliable estimate.

(***) Standard error estimates cannot be accurately determined.

NOTE: The NAEP Reading scale ranges from 0 to 500. Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.

Figures F-8a through F-8g all related to a constructed-response item (also associated with the blue crabs passage shown in Figure 5a). As participants were shown the item, the performance data, the scoring guide, and examples of extensive, essential, partial, and unsatisfactory student responses, they were asked to think about the intended audience and the amount and nature of the information that could be accessed through the NAEP Question Tool. The group felt this information had value and use for teachers, parents, and students in particular, and one participant noted that exemplar items like these were used in that state's testing program. When considering the level of data and content that NAEP provides for MC and CR questions relative to the states, most of the specialists in the focus group agreed on the considerable value of the data that are reported and that "NAEP is extensive"—some states do this, but to a lesser degree, as most states do not have anything nearly as comprehensive. NAEP was considered by the group to be a huge resource with respect to released items and information about those items ("This is what states should aspire to"). One indicated that they are trying to get their teachers to use the NAEP site, and another said while they do not come across many people using it, it is hard to keep track.

Figures F-8a and F-8b. Example constructed response item and results from grade 4 reading

[ABOUT NAEP...](#) | [SUBJECT AREAS...](#) | [HELP](#) | [SITE MAP](#) | [CONTACT US](#) | [GLOSSARY](#) | [NEWSFLASH](#)
[SAMPLE QUESTIONS](#) | [ANALYZE DATA](#) | [STATE PROFILES](#) | [PUBLICATIONS](#)

NAEP Questions

The Nation's Report Card (home)

[New Search](#) | [Previous Search Results](#) | [Tool Help](#)
← Question 65 of 82 →

📁 To Print Folder: Empty | 📄 Add Question

Subject: **Reading** | [Subject Info](#) | Grade: **4** Block: **1998-4R6** No.: **4**
 Description: **Blue Crabs: Tell things learned-paragraph**

[Question](#) | [Performance Data](#) | [Content Classification](#) | [Scoring Guide/Key](#) | [Student Responses](#) | [More Data](#)

[Printable Version](#) | [View Reading Passage](#)

4. Write a paragraph telling the major things you learned about blue crabs.

[ABOUT NAEP...](#) | [SUBJECT AREAS...](#) | [HELP](#) | [SITE MAP](#) | [CONTACT US](#) | [GLOSSARY](#) | [NEWSFLASH](#)
[SAMPLE QUESTIONS](#) | [ANALYZE DATA](#) | [STATE PROFILES](#) | [PUBLICATIONS](#)

NAEP Questions

The Nation's Report Card (home)

[New Search](#) | [Previous Search Results](#) | [Tool Help](#)
← Question 65 of 82 →

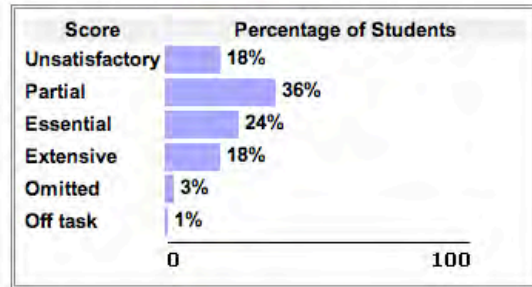
📁 To Print Folder: Empty | 📄 Add Question

Subject: **Reading** | [Subject Info](#) | Grade: **4** Block: **1998-4R6** No.: **4**
 Description: **Blue Crabs: Tell things learned-paragraph**

[Question](#) | [Performance Data](#) | [Content Classification](#) | [Scoring Guide/Key](#) | [Student Responses](#) | [More Data](#)

[Printable Version](#)

1998 National Performance Results



Note:

- These results are for public and nonpublic school students.
- Percentages may not add to 100 due to rounding.

Last updated 24 June 2005 (PCG)

Figure F-8c.

Scoring Guide
<p>Score & Description</p> <p>Extensive</p> <p>These responses demonstrate a more in-depth understanding of the characteristics of blue crabs by providing information about at least three of the major points in the passage (i.e., external skeleton, molting, mating, reproduction, development, regeneration and habitat), or by providing a thorough discussion of at least two of the important characteristics. For example:</p> <ol style="list-style-type: none"> "The passage is mainly about how blue crabs look, how they shed their shells, and reproduce, and grow new legs if they lose the old ones." "The passage is mostly about how blue crabs mate and grow new shells. A male and female blue crab dance as they mate only once; and when the eggs are fertilized, the female crab carries them around until they hatch and then she dies. Blue crabs can also get rid of their shells. The old shell splits apart and the crab sort of walks out of it. Underneath the old shell is a new soft shell which gets very hard."
<p>Essential</p> <p>These responses demonstrate a basic understanding of the characteristics of blue crabs by providing information about at least two of the following: external skeleton, molting, mating, reproduction, regeneration, or habitat; or by providing a thorough discussion of one of the important characteristics. A thorough discussion includes more than two pieces of information on one characteristic. For example:</p> <ol style="list-style-type: none"> "If blue crabs lose a leg, they can grow another one. The female blue crab dies after she lays her eggs." "The passage is about how blue crabs shed their shells, and grow new legs if they lose the old ones." "The passage tells about how a male and female blue crab dance as they mate only once; and when the eggs are fertilized, the female crab carries them around until they hatch and then she dies."
<p>Partial</p> <p>These responses demonstrate an understanding of some disparate facts about blue crabs or any one piece of important information (i.e., any facts about the blue crab's external skeleton, molting, reproduction, regeneration, or habitat). For example:</p> <ol style="list-style-type: none"> "This passage is about where blue crabs live, how to catch them, and how many crabs the author caught one summer." "This passage is about how they grow new legs if they lose one."
<p>Unsatisfactory</p> <p>These responses demonstrate little or no understanding of the characteristics of blue crabs by providing inappropriate information, or personal opinions, or giving only one disparate fact. For example:</p> <ol style="list-style-type: none"> "The crabs in this article are blue." "Blue crabs are blue." "Blue crabs are cool." "Blue crabs have big claws."

Figure F-8d.

Extensive - Student Response

4. Write a paragraph telling the major things you learned about blue crabs.

I learned that blue crabs are very strong. They can make a painful pinch. They have outstretched claws. They are arthropods. Blue crabs belong to a group of crustaceans. Their shells are strong and legged that molting means getting rid of a shell. As blue crabs get older they move into shallower water. Predators are raccoons, alligators, and people. Crabs can lose a leg and grow a new one.

<http://www.rites.ed.gov>

Scorer's Commentary

These responses demonstrate an "extensive" understanding by referring to three major topics discussed in the article. Both responses also include additional information, for example "their powerful pinch," which is correct but unrelated to one of the major topics of the article.

[Close Window](#)

4. Write a paragraph telling the major things you learned about blue crabs.

I learned that the blue crabs lose their shell about twenty times during their life. When a crab is cornered in, to protect themselves, they stretch their claws and crab at people. Crabs have an external skeleton and jointed legs. If a crab is caught it loses its leg or claw to escape. When the crab gets its leg or claw back it is usually smaller than it was before.

Scorer's Commentary

Figure F-8e.

Essential - Student Response

4. Write a paragraph telling the major things you learned about blue crabs.

I learned that blue crabs live in the shallow waters of marsh and that when they hatched they are use tiny use a little speck of dust.

4. Write a paragraph telling the major things you learned about blue crabs.

I learned that blue crabs have a real scientific name called Callinectes Sapidus they are real delicious to eat. It changes from a little larva to a blue crab. It is always laying eggs on the real shallow water. It drops off its legs and grows new ones.

http://nces.ed.gov

Scorer's Commentary

The first response presents only two pieces of information; however both of them, reproduction and regeneration, are major topics from the article. The second response provides more information from the article, but still only two major topics are presented: habitat and development. The beginning of the second response with the mention of the crabs being delicious is not a major topic in the article.

[Close Window](#)

Scorer's Commentary

Figure F-8f and F-8g.

Partial - Student Response

4. Write a paragraph telling the major things you learned about blue crabs.

What I have learned about blue crabs are they are strong. You have to be real carefully with them or they will pinch you. And lots of more.

http://nces.ed.gov...

Scorer's Commentary

Both responses provide accurate details from the article, but no reference to any major topics.

[Close Window](#)

4. Write a paragraph telling the major things you learned about blue crabs.

I learned that a blue crab pinches hurts. I also learned that many fishermen catch them to sell.

Scorer's Commentary

Unsatisfactory - Student Response

4. Write a paragraph telling the major things you learned about blue crabs.

They are blue. And I've never seen a blue crab.

4. Write a paragraph telling the major things you learned about blue crabs.

I learn they eat blue crabs for dinner because they are good and you can eat them for lunch and they will be good for a dinner.

http://nces.ed.gov...

Scorer's Commentary

Both responses fail to provide more than one disparate fact about blue crabs. The first response merely repeats the title, and the second response provides only one fact about people eating crabs.

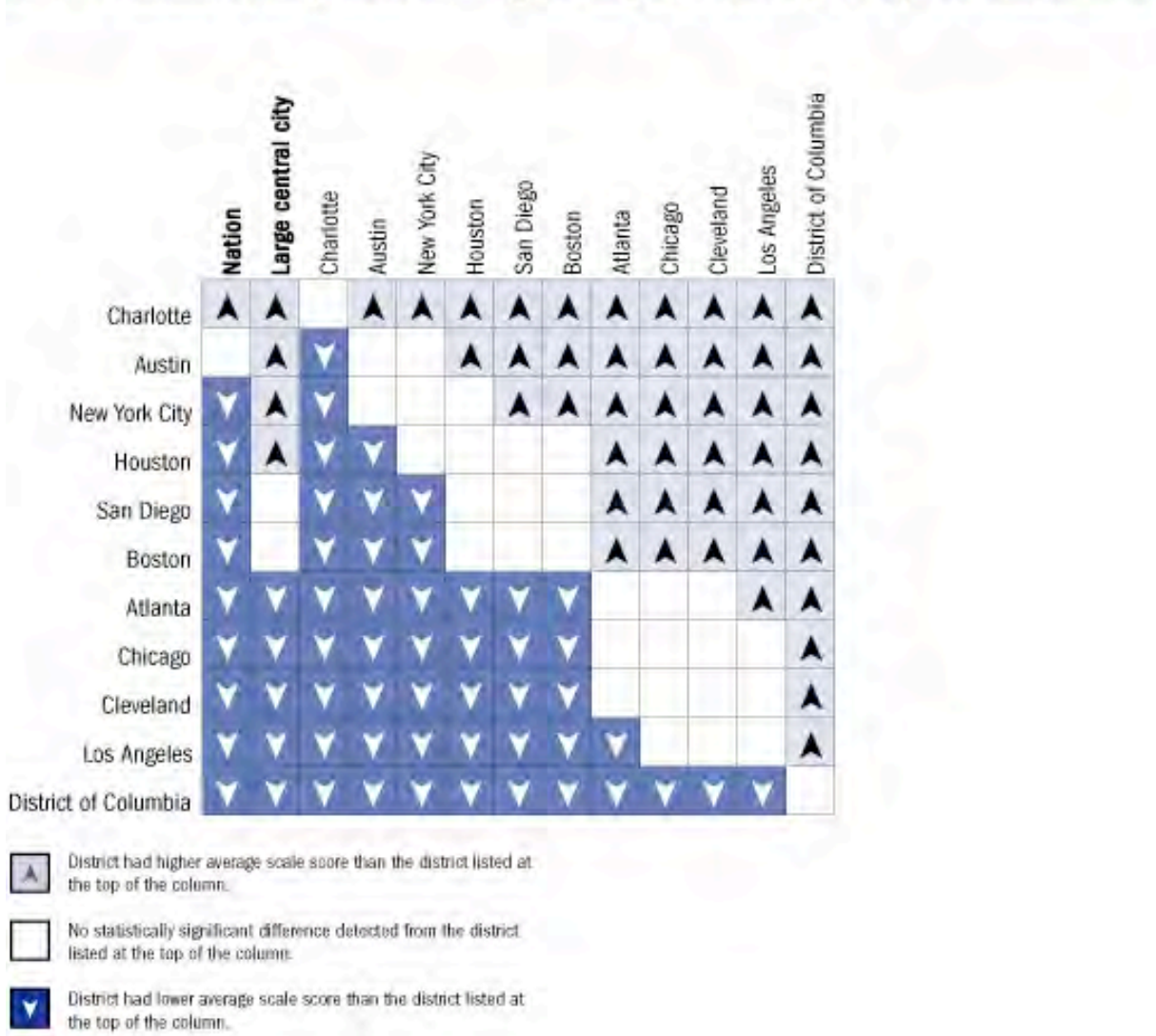
[Close Window](#)

Scorer's Commentary

Figure F-9 was a version of a “pantyhose” chart displaying the overall cross-district comparisons of average reading scale scores in 2005 for grade 4 public school students. Very few of the participants reported they had seen a chart like this before. When asked what kind of information the chart was conveying, panelists responded with, “It’s comparing cities,” and “It’s comparing districts.” They recognized that the lightly shaded boxes with the dark arrows pointing up represented that the city to left was scoring higher than the corresponding city at the top of the chart and that the darker boxes with the white arrows pointing downward meant the opposite. When the participants were asked which cities would be happiest with the results displayed in this chart and why, responses included “Charlotte and Austin” and “Charlotte has more up arrows, so they are performing better.”

Figure F-9.

Overall cross-district comparisons of average reading scale scores, grade 4 public schools: 2005



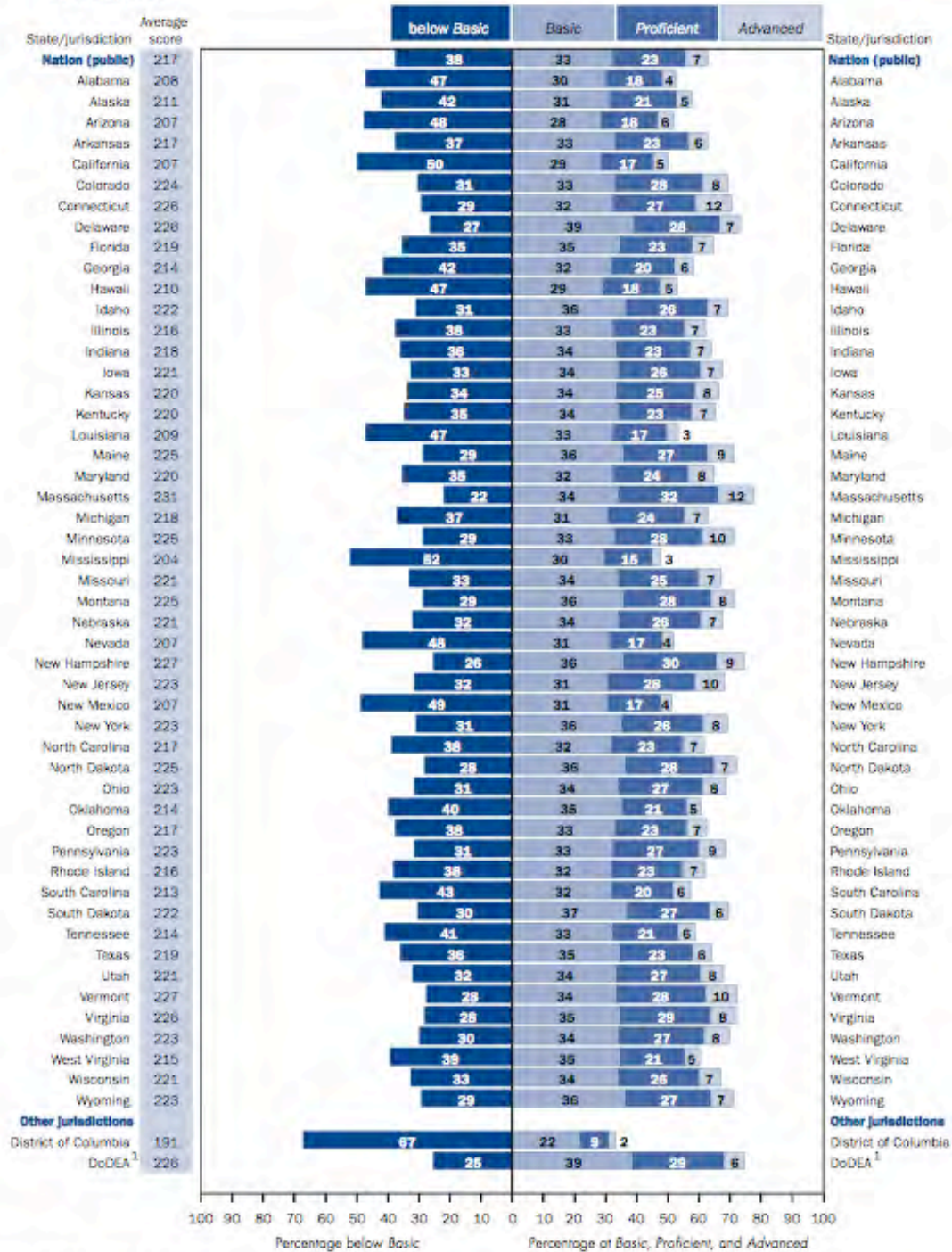
DATA: View complete data with standard errors for scale scores in [districts and the nation](#) or [large central cities](#).

Although the participants managed to read the chart appropriately, they found this type of chart to be confusing at first because the cities are listed on both axes. It was also noted that the chart showed which cities had average scale scores that were significantly higher or lower than other cities, but it did not provide information on how big the difference really was (one participant commented, “This figure doesn’t provide information on how much of a difference there is between districts. It could be a very small difference.”) The participants also commented that a chart like this would not be valuable to states that have no large cities.

Figure F-10 was not discussed in the large group due to time limitations.

Figure F-10.

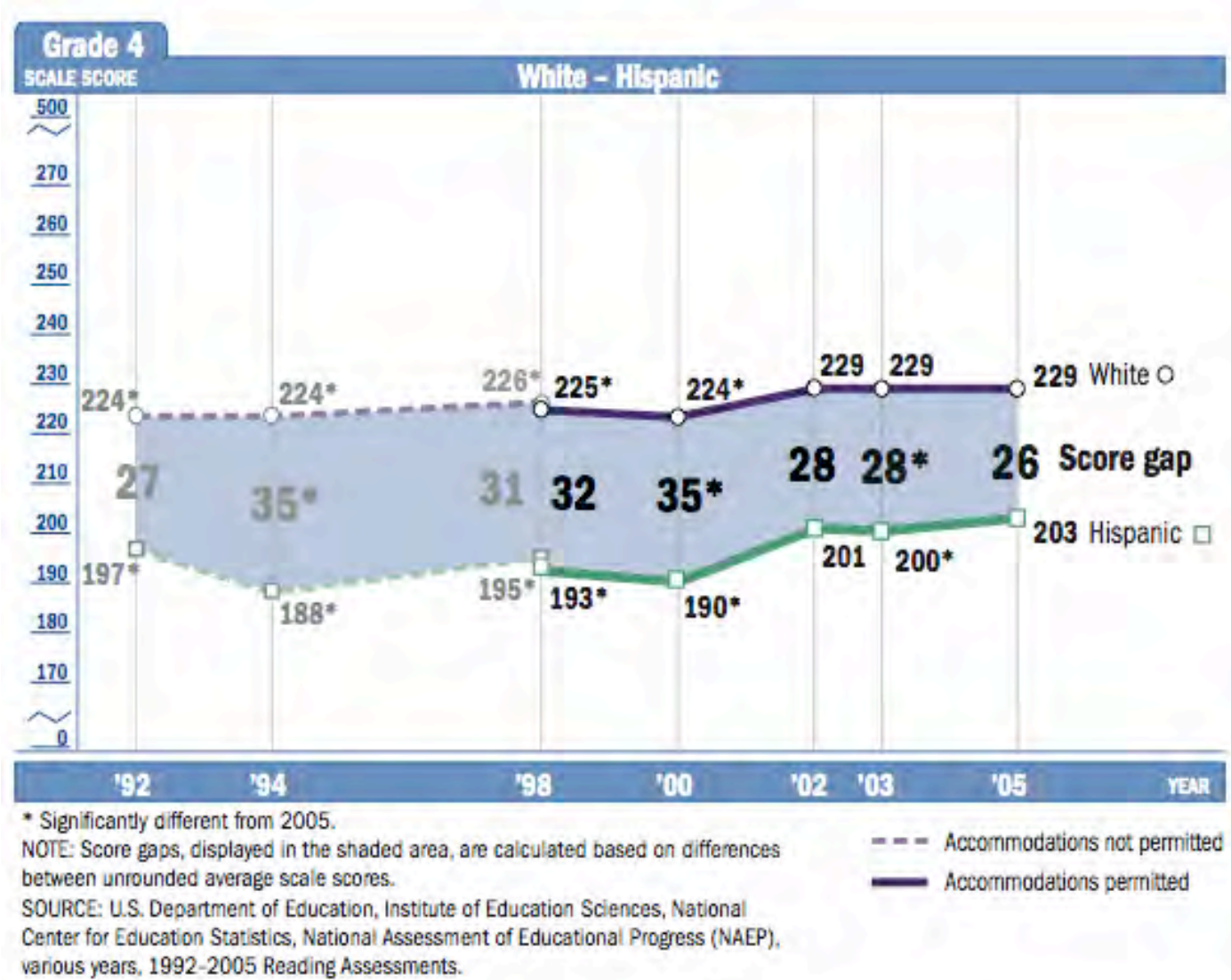
Figure 11. Average reading scale scores and percentage of students within each achievement level, grade 4 public schools: By state, 2005



¹ Department of Defense Education Activity.
 NOTE: The NAEP reading scale ranges from 0 to 500. Detail may not sum to totals because of rounding. The shaded bars are graphed using unrounded numbers.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment.

In Figure F-11, two lines representing the performance of white and Hispanic fourth-graders overall from 1992 to 2005 were presented, and the figure also shaded the differences in performance for the two groups as a way to illustrate score gaps over time. This was the first of several displays shown to participants that focused on methods used by NAEP to communicate score gaps. Most participants in the group commented that they had seen displays such as Figure F-11 before. When asked the purpose of this figure, one participant replied, “To show the gap is getting smaller,” to which another specialist responded, “The scores are not really going up though.” They characterized the white students’ performance as very flat, and noted general similarities for the Hispanic students, when even though it moved up and (mostly) down a little in the intervening years until about 2000, the overall trend for those kids from 1992 to 2005 was pretty flat. Interpreting the gap was the next task: Most of the group noted that while the gap is decreasing, it hadn’t really changed all that much. One participant pointed out that if the figure began reporting in 1994 rather than 1992, “It looks more impressive,” which led to this comment: “That’s how statistics can be manipulated.” The group took particular notice of the asterisks (denoting significance of a particular year’s result relative to 2005) in concluding that there really was not much change in the overall score gap between white and Hispanic grade 4 students.

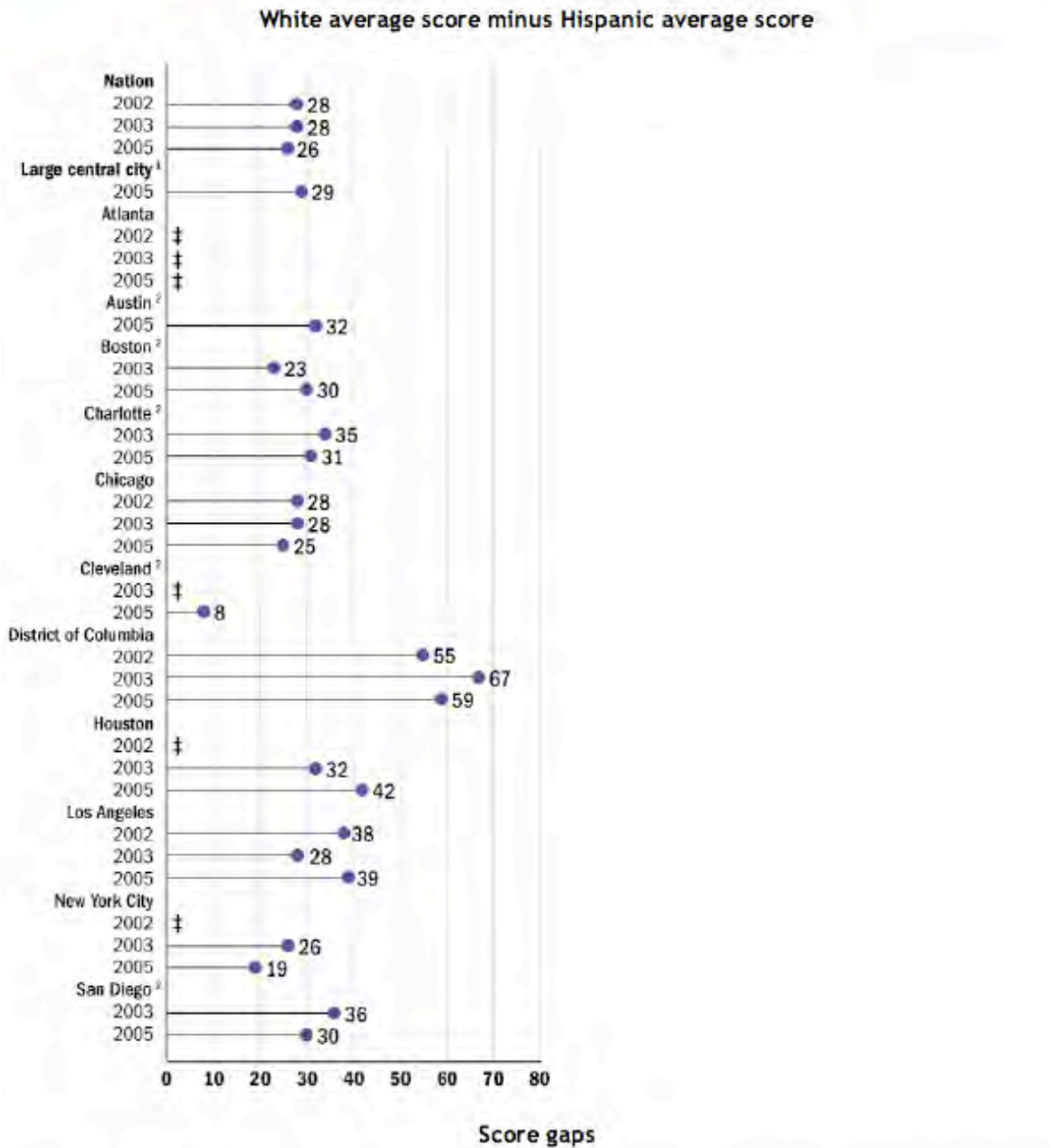
Figure F-11. White–Hispanic scale score comparison on grade 4 reading



Moving on to Figure F-12, which was another way of illustrating gaps (the data in this display were from the Trial Urban District Assessment (TUDA) 2005), most participants were not familiar with this specific NAEP display (when the difference between average scale scores for two subgroups were the values plotted as horizontal bar graphs). They identified the purpose as showing the white-Hispanic gap in urban districts, and with a similar aim as Figure F-11, but displayed differently. However, the group generally concurred in expressing a preference for the appearance of Figure F-11, when gaps were presented in a more visual way and comparisons within a district are easier to make. The group did note that Figure F-12 does display different information, because results are not only provided within a single reporting jurisdiction over time but also for multiple jurisdictions at once, which permits informal (not informed by results of significance testing) comparisons between those jurisdictions. Specific difficulties noted with this display included reading down the axis of years to make those kind of informal comparisons, which the participants found challenging due to the lack of results for some districts in some years because the reporting standards were not met. For example, they wanted to mark off the patterns to the results from 2005, and because there were not three years' worth of data represented for each district, the pattern was not as simple as looking at every third line. A suggestion made was to include the third year (often 2002) for each district whether the district participated or not, or if the reporting standards were met or not, to facilitate the consistency of the reporting and thereby ease reading of the display.

Figure F-12.

White-Hispanic gap in average reading scores, grade 4 public schools: By urban district, various years, 2002-2005



† Reporting standards not met.

¹ Data for large central city schools are not included for years prior to 2005 because the application of the definitions of the types of location has changed. For 2005, "large central city" includes nationally representative public schools located in large central cities (population of 250,000 or more) within a Metropolitan Statistical Area (MSA).

² The district did not participate in 2002 or 2003.

NOTE: Score gaps are calculated based on differences between unrounded average scale scores.

DATA: View complete data with standard errors for scale scores in each district: [Atlanta](#), [Austin](#), [Boston](#), [Charlotte](#), [Chicago](#), [Cleveland](#), [District of Columbia](#), [Houston](#), [Los Angeles](#), [New York City](#), and [San Diego](#).

The next figure presented to the group was the Grade 8 Item map (Figure F-13). Discussion of this figure was relatively brief, as most in the room indicated they were not familiar with item maps. Among those who had seen them, one used them to give meaning to student performance and to see what types of questions were used on the assessment. Another stated that the NAEP item maps were used to write performance level descriptions for that individual's state. When asked how an item comes to be placed on the map at a particular location (the example given to the group was the score point 287), most participants were not certain, other than to suggest, "It is a higher cognitive demand than *Proficient*."

Figure F-13.



¹ Each grade 8 reading question in the 2005 reading assessment was mapped onto the NAEP 0-500 reading scale. The position of a question on the scale represents the average scale score attained by students who had a 65 percent probability of successfully answering a constructed-response question, or a 74 percent probability of correctly answering a four-option multiple-choice question. Only selected questions are presented. Scale score ranges for reading achievement levels are referenced on the map. For constructed-response questions, the question description represents students' performance at the scoring level being mapped.

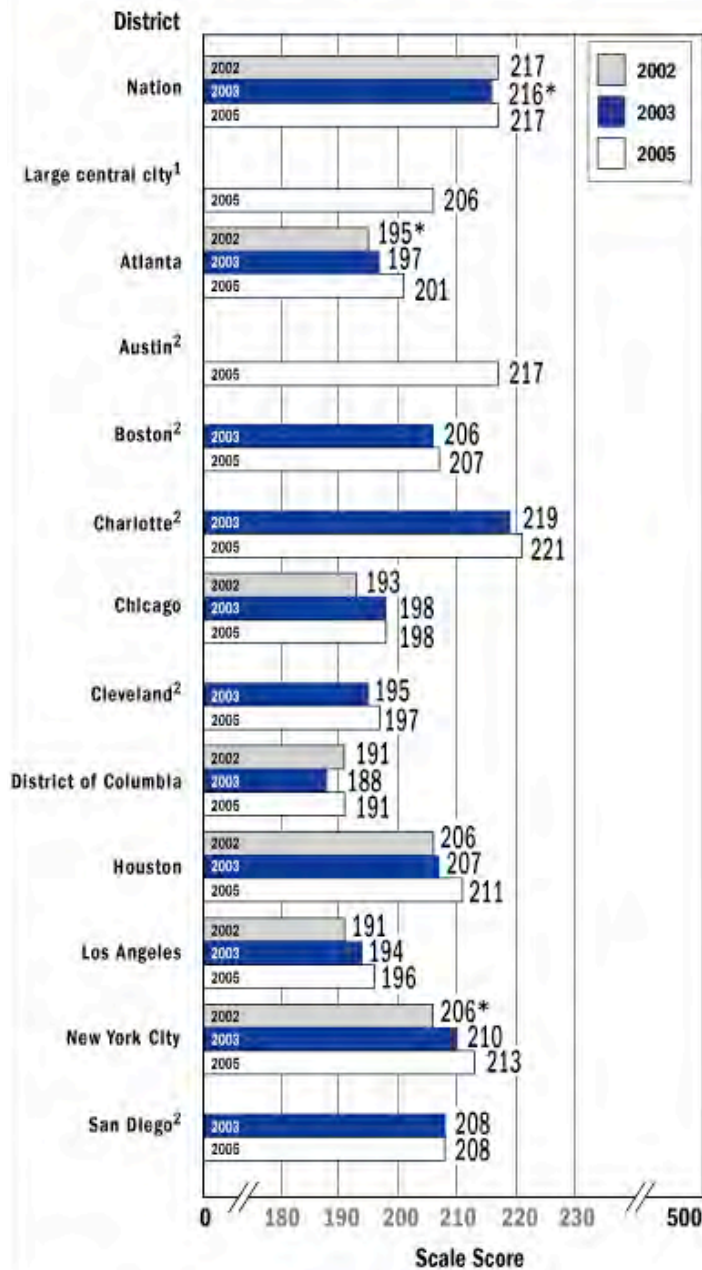
NOTE: Regular type denotes a constructed-response question. *Italic* type denotes a multiple-choice question.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment.

Figure F-14 was not discussed in the large group due to time limitations.

Figure F-14.

Average reading scale scores, grade 4 public schools: By urban district, various years, 2002-2005



* Average scale score is significantly different from 2005.

¹ Data for large central city schools are not included for years prior to 2005 because the application of the definitions of the types of location has changed. For 2005, "large central city" includes nationally representative public schools located in large central cities (population of 250,000 or more) within a Metropolitan Statistical Area (MSA).

² The district did not participate either in 2002 or 2003.

NOTE: Significance tests were performed using unrounded numbers.

DATA: View complete data with standard errors for scale scores in [districts and the nation](#) or in [large central cities](#).

In Figure F-15, the percentages of students at each achievement level for Reading Grade 4 in five states (California, Florida, Illinois, New York, and Texas) for 2003 and 2005 were represented in “stacked” horizontal bars. This display was unique among the figures presented to the group because the states included in the display were chosen by the researchers and the display itself was generated using the NAEP Data Explorer, although a similar display method was used in Figure F-10 (which was not discussed in the large group due to time considerations). The responses from the participants in the group were quite varied, with some finding it easy to see the percent of kids in achievement levels, while other reported that it was confusing because each state was listed twice (once for 2003, then again for 2005). A preference for placing the 2003 and 2005 results for a single state next to each other was raised (participants drew a connection in this regard to Figure F-12 and Figure F-14 (the latter was omitted from broad discussion due to time limitations, although screen capture of it was included in the participants’ packets)). One participant asked about the line between *Basic* and *Proficient* on this graph, and it was explained that when working within the NAEP Data Explorer, users could choose where to put the zero point on the axis by clicking on an achievement level. Another specialist indicated that such a choice was common practice in some states, to see who is above and below *Proficient*. The axis on this graph did lead to some minor confusion among the group, given that it there were 80 percentage points on one side, and 60 on the other, but once participants were directed to look within a jurisdiction to focus on the percentages within the achievement levels, the meaning became clearer for all.

This page intentionally left blank

Appendix F-1: Communication for gathering documents

Hi [COR(s)]

As you know, I am the lead person on the audit part of the NAEP evaluation project. The purpose of this communication is to provide some detailed follow up on the documents we need to complete the part of the audit for [respective contractor]. In addition to the documents we asked you to provide to us, we have identified some additional documents that we believe will give us more information about the procedures and processes used by [respective contractor]. Attached to this communication is a list of specific documents that we would like to receive that are specific to the agency/contractors that you work with. We found these documents listed on the NCES Web site. We made our selection of these documents based on the very brief description provided on the Web site about the purpose of the study/project. This list of documents only represents a limited resource for the complete set of documents that we need from [respective contractor]. Therefore, this list is not meant to be exhaustive, but only one of a number of lists of documents that you have access to that would inform our work. In addition to these documents we have already requested information that is available in the quality control documents prepared by [respective contractor] and scope of work statements. We also need, of course, technical reports and project documents that are prepared and submitted to NCES by [respective contractor] as part of the regular reporting expectations as contractors on the project.

In addition to gathering the documentation that will provide evidence needed to conduct the audit on your contractor's roles in NAEP, we will also be conducting brief on-site meetings with the agencies and contractors. We would like to conduct these meetings in a 1 – 2 day session in [respective month(s)], 2005. In order to make these site visits as productive as possible, we'll need to have reviewed the documentation in advance so we can prepare specific questions to address during the site visits. Therefore, we will need these documents from [respective contractor] very soon. We would appreciate if these materials could be made available to us by April 30, 2005.

It would be very helpful if you could provide us with some suggested dates for when we could conduct the on-site visits with [respective contractor]. We want to be sensitive to the task demands and timelines of the contractor, while at the same time meeting our deadlines for completing the audit.

Please feel free to contact me regarding the audit process or questions about getting these documents to me for our review. I look forward to working with you on this project.

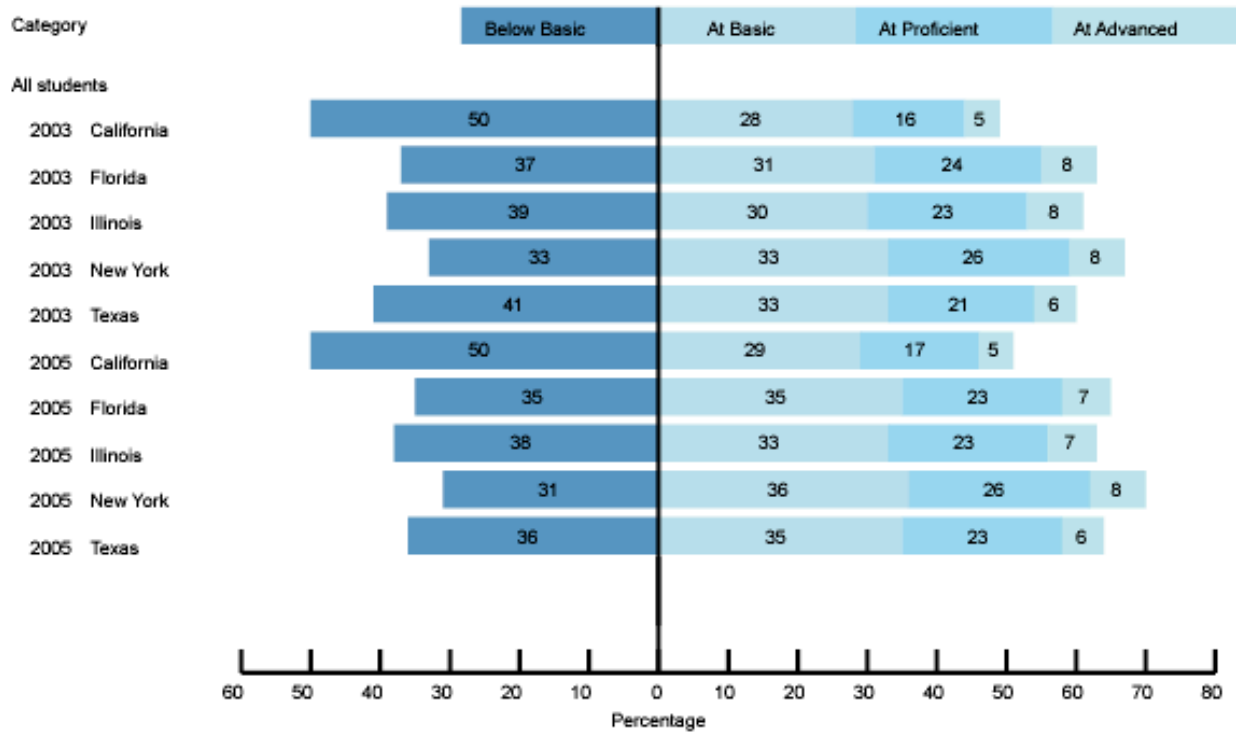
Barbara Plake, NAEP Audit Project Leader
Buros Center for Testing

Attachments:
Revised Audit Components and Sources of Evidence
Specific document request

This page intentionally left blank

Figure F-15.

Percentages of students at each achievement level for reading, grade 4
 All students (TOTAL)
 By jurisdiction, 2003 and 2005



NOTE: Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.
 SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and

After focusing on this display, the group moved on to a broader discussion about NAEP and state score reporting in general. A general agreement was expressed among the reading specialists in attendance that performance levels (in NAEP, below *Basic*, *Basic*, *Proficient*, and *Advanced*) were preferred over scale scores. The kinds of results that states are often interested in varies, with some choosing to emphasize state comparisons based on geographical proximity, while others were more concerned with demographic-group performance or state comparisons chosen to emphasize peers with demographically similar student populations.

When asked to reflect back on the range of displays shown throughout the focus group, several different items were highlighted. Color was mentioned as an important thing to retain and include whenever possible, particularly on the state comparison maps. Overall, the participants found many of the displays particularly useful to them, including the ones showing the score gaps (Figures F-11 and F-12), the state maps (Figures F-3a, F-3b, and F-4), and the line graphs (Figure F-1). Figure F-15 was also cited. Some discussion of the utility of the distracter data in Figure F-7 for different audiences ensued, with some attendees mentioning that while that information was useful for them as state reading specialists, teachers would probably want it more simplified. Also with respect to the Question Tool, the actual student responses are valued for their instructional and scoring uses. The participants were somewhat divided on the pantyhose chart. Consensus was reached, however, as to status of NAEP as a tremendous resource for these individuals in their work and in their states more generally.

In discussing reporting methods in general, most found that executive summaries are generally useful for them, and the NAEP state snapshot reports for a content area at a given grade level was mentioned as being particularly useful for giving results clearly with just enough context. Speaking to the specialists' knowledge of schools' and teachers' use of the results, executive summaries were identified as unlikely to be read, and participants stated, "Schools want to know how their kids are doing compared to other schools and districts," but "Not all schools care as much about NAEP results because their kids don't take it," and "Districts want individual results but that's not NAEP." Another attendee described the state's use for NAEP as being important at the state level but not at the district level, to "justify our state assessment is on target."

Discussion

This focus group of state Reading content specialists provided considerable information about how different NAEP data displays are used and understood by state education personnel. By and large, the group was comfortable with many of the ways that NAEP currently reports results, but identified several possible directions for clarification. One such area was the process they used to look at the data represented in the graphs, with an eye toward recognizing patterns to the results. The group expressed a preference for consistency in the layout of reporting results for different states or jurisdictions, and as a consequence found Figures F-12 and F-14 more challenging to understand in that way, because not each year was represented on the axis and it made interpreting results for different years more problematic for some. That said, the general idea of both figures was generally found to useful for the group, but the suggestion is made to leave space for each year for each jurisdiction on those types of figures and denote it accordingly if a jurisdiction did not participate or meet reporting standards is no reportable data are available.

Another source of complexity for this group was in how at or above a given achievement level (e.g., *Basic* or *Proficient*) was represented in some figures. For example, though the purpose of Figure F-2 was to illustrate the percents at or above *Basic* and at or above *Proficient* some participants found themselves wanting to know the proportion of students below *Basic* in this figure, and wanted it to be on the figure somewhere instead of having to subtract out the percent at or above *Basic* from 100. There was also interest in breaking this figure into two graphs instead of nested bar charts, with one figure focusing on the percents below and at or above *Basic* and another with the percents below and at or above *Proficient*.

Some of the more innovative displays, including the clickable state maps and TUDA's use of the pantyhose chart, were largely well received by this group. These displays in particular

are visual and readily interpretable for most in the group. Several small suggestions were received for improving the clarity of the state maps, such as making the color scheme more distinct and enhancing the legend. For the pantyhose chart, there is a bit of a learning curve required, but for making pair-wise comparisons between jurisdictions when the number of jurisdictions is about 10 or fewer, this display can be effective. However, questions were raised as to the practical meaning that could be ascribed to this display when no scale scores are shown, just significance test results. There was one further issue with the language in the legend for this display (noted in Appendix E), as users have to infer that up arrows correspond to *significantly* better performance and down arrows indicate *significantly* lower performance.

As to the NAEP Question Tool, the reading content specialists in this group reported consistent, strong positive feelings about the nature and layout of the data represented in those figures. They noted that the Question Tool is likely an underutilized resource. To the extent possible, the Question Tool was regarded by this group as a model to which states should aspire to create with their own items or, barring that, encourage teachers to draw on more fully.

Among the key findings of this report, this focus group reinforced the notion that information along the axes of displays and found in keys or legends is read by participants and deserves careful scrutiny in the preparation of NAEP data displays. When an axis in a figure generated by the NAEP Data Explorer (Figure F-15) shows a horizontal axis labeled “Percentage” with 140 possible points on it, this potentially increases the difficulty some users may have with making correct interpretations. The labeling for a similar display given in Figure 10 may be instructive in this regard. Footnotes that denote statistical significance should make that point consistently. When labeling axes, consistency for years should be maintained (if three years’ worth of results is to be reported for some jurisdictions, then leave spacing for three years for all jurisdictions even if the full data is not available. Color whenever possible is encouraged, but should be chosen and used with a clear purpose and to differentiate results plainly.

Next Steps

There are several important directions for follow-up research. First, while the use of group discussion clearly yielded much useful information, a logical next step is to carry out one-on-one explorations of several data displays with NAEP data users as they navigate themselves through different portions of the NAEP Web site. In this way, we could gather more information about how different individuals fare with respect to both knowledge and interpretation of several of the more interactive features of the site, including the clickable state maps, the Question Tool, and the NAEP Data Explorer.

In addition, given the kinds of suggestions made about these displays, one additional direction for future research includes the development of several redesigns of the displays shown here, with research participants comparing current and revised displays for clarity and understanding. This idea was pursued by Wainer, Hambleton, and Meara (1999) and produced some interesting findings. While NAEP is clearly at the forefront of testing programs with respect to its investment in methods for disseminating results, the results of this focus group indicated that there remain some sources of confusion among audiences who have some familiarity and regular use of NAEP. Clearly substantially more research and development work is needed in the near future.

References

- Goodman, D. P., and Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17, 145-220.
- Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress* (NAEP Validity Studies Panel Report). Palo Alto, Calif.: American Institutes for Research.
- Simmons, C., and Mwalimu, M. (November, 2000). What NAEP's publics have to say. In M. L. Bourque and S. Byrd (eds.), *Student performance standards on the National Assessment of Educational Progress: A study initiated to examine a decade of achievement level setting on NAEP* (pp. 185–219). Washington, D.C.: National Assessment Governing Board.
- Wainer, H., Hambleton, R.K., and Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335.

**Appendix G: Displaying NAEP Results Effectively:
Exploring Stakeholder Understanding of Selected NAEP Graphs**

Jill Delton and April L. Zenisky
Center for Educational Assessment
University of Massachusetts Amherst

Feb. 13, 2007

Introduction

The old adage says that a picture can express a thousand words, and this is certainly the case when the topic at hand is reporting results for a large-scale assessment program such as the National Assessment of Educational Progress (NAEP). Graphs and figures can be used quite effectively to communicate information about student test performance. Indeed, as noted by Wainer (2000), well-constructed graphical displays of data put the data front and center to illustrate a clear and complete story. For NAEP, there are many stories of interest. For example, some people are curious about how students in the nation overall did or how students in one or more states compare. Some consumers of the data want to understand results in terms of scale scores and/or achievement levels. Still others focus their interest on subgroup performance, and may pay particular attention to comparing score for groups or computing score gaps.

With so much information to report, there is an equally vast array of display options available to NAEP. The most recent Mathematics report (Perie, Grigg, and Dion, 2005) illustrated results using a number of different graphical displays including line graphs, horizontally and vertically stacked bar charts, and line graphs with multiple lines for subgroups illustrating gaps in group performance. In addition, tables are used to communicate other information. Another recent Mathematics report (Rampey, Lutkus, and Dion, 2006) incorporates several of these display strategies as well as clustered bar charts and a “pantyhose” chart.

As described in the reports contained in Appendices E and F, two focus groups of state content specialists were convened to discuss selected data displays from NAEP reports in Reading and Mathematics, respectively. The purpose of these focus groups was to learn more about how a subset of the broader audience for NAEP data understands and uses NAEP results, specifically the graphical information disseminated in NAEP publications, both in print and on the Web. In this report, report on a third focus group convened to focus the discussion on specific findings of the previous two groups and to explore in greater depth specific recommendations for improving NAEP data displays.

Method

A focus group of state reading content specialists and teachers was convened to gather information about the meaningfulness of different NAEP score reporting methods.

Participants were selected from a list provided by state testing personnel across New Hampshire, Massachusetts, Connecticut, Vermont, and Rhode Island. From this list, about 15 individuals were contacted and seven confirmed attendance at the meeting. However, weather conditions and personal circumstances on the day of the meeting resulted in only four of the participants attending. The participants were provided with a small honorarium of \$200 for their participation in the focus group.

The four participants represented two different states, Vermont and Massachusetts. Two were males and two were females. All of the participants had at least 5–10 years of experience in the field of education and most had 11–20 years. Two of the participants were teachers, one a school psychologist, and one a reading assessment coordinator.

All of the participants reported that they were somewhat familiar with NAEP. Most of the participants work with NAEP data or information a few times a year and one reported never working with NAEP data. The participants that reported working with NAEP data in the past have studied trends and one of the participants had conducted item reviews.

Procedure

At the focus group meeting, participants were provided with an overview of the project and asked to complete a brief demographic survey. Next, a series of data displays consisting of both tables and figures from several recent releases of NAEP reading results (National and State results, 2005; Trial Urban District Assessment, 2005; and the NAEP Question Tool) were

projected on a screen in color (and also given to participants as full-page handouts in black-and-white). We focused on displays of both Reading and Mathematics. As each figure or table was displayed, participants were asked questions about the displays. The displays shown at this meeting were chosen as a sampling of the types of tables and figures seen throughout recent NAEP reports, including:

- line graphs,
- stacked bar charts,
- clickable state comparison maps of average scale scores and percents of students at or above achievement levels,
- tabs from the NAEP question tool with item text, student item performance, a distracter analysis,
- “pantyhose” charts,
- ways of displaying score gaps between reporting subgroups,
- bar graphs, and
- item maps.

As each display was projected on the screen, participants were asked to reflect on each display for a few minutes, and then they were asked questions about the data display by one of the two meeting facilitators. Questions ranged from those that were informational in nature (“What was the average score for eighth graders in 2005 in reading?”) to opinion (“What, if anything, do you find confusing or not clear about this display?”). The focus group discussion format was appropriate for this study because this format served to stimulate some broader conversations among the participants and facilitators about the data displays, building on what was being displayed on the screen, and allowed the participants to answer some of the more difficult data interpretation questions collaboratively.

The last task asked of participants in the meeting was for them to respond to a sequence of discussion questions that focused on broader issues of score reporting and NAEP. These questions included reflection on the collection of displays presented throughout the meeting, their preference for receiving information themselves and how educators want information, and ways of representing gaps in subgroup performance.

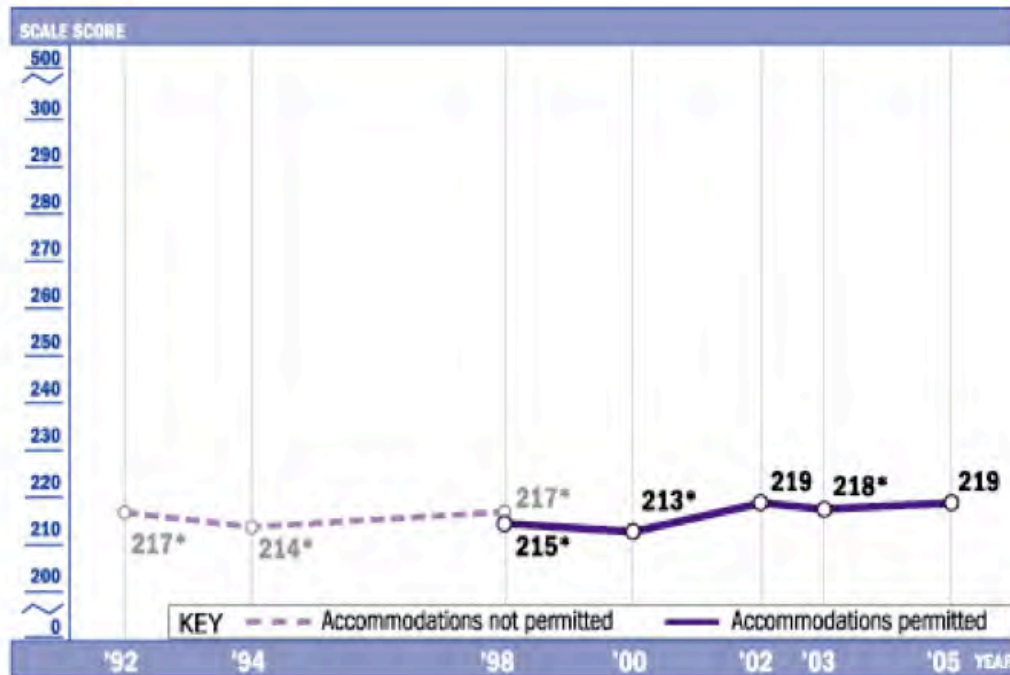
Results

Figure G-1 presents a line graph of average reading scale scores showing the performance of fourth-grade students from 1992 to 2005. The graph displays the average scale score for each year the assessment was given. An asterisk was placed next to the scores for the years in which the difference between that year and 2005 was significantly different. A dashed line was used to represent the years in which accommodation were permitted.

The participants thought the line graph was a good way to display data over time and found it easy to read. There was some confusion about what the scale score meant because most of the participants were not familiar with the NAEP scale. Questions were also raised about the accommodations. One of the participants asked, “What is the difference between accommodations versus non accommodations and does it make a difference when comparing the scores?” The participants also found the time line confusing because the years in which NAEP was not administered did not appear on the graph (one of the participants commented, “Certain years are missing, so it looks incomplete.”)

Figure G-1.

Average reading scale scores, grade 4: Various years, 1992-2005



* Significantly different from 2005.

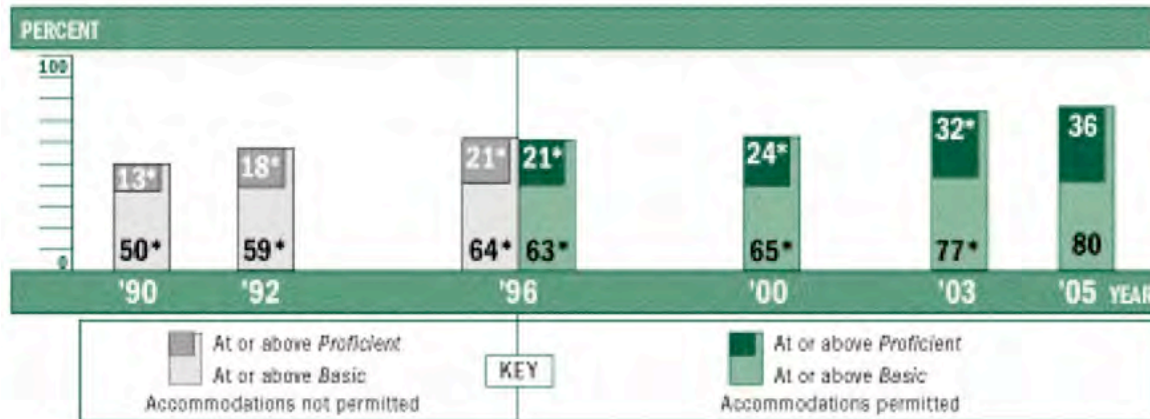
NOTE: The dashed and solid lines represent results based on administrations when accommodations were not permitted and when accommodations were permitted, respectively. View complete data with standard errors for [grade 4](#).

Figure G-2 was a stacked bar graph displaying the percentage of grade 4 students at or above *Basic* and at or above *Proficient* in Mathematics from 1990 to 2005. In this figure, a vertical line divided the years in which accommodations were permitted and the years in which they were not.

The participants were easily able to read and interpret this graph. They found it easier to understand the percentages than it was to understand the scale score, as they were unfamiliar with the NAEP scale (one participant commented, “Figure G-2 is easier to interpret than Figure G-1 because percentages are more familiar to people.”) Questions were again raised about the affect the accommodations have on the comparability of the scores.

Figure G-2.

Percentage of students at or above *Basic* and at or above *Proficient* in Mathematics, Grade 4 Various Years, 1990-2005



* Significantly different from 2005.

NOTE: The gray shaded boxes represent results based on administrations when [accommodations](#) were not permitted. View complete data with standard errors for [grade 4](#).

Figure G-3 was a table displaying the percentage of grade 4 students at or above each achievement level in mathematics from 1990 to 2005. This figure also used an asterisk to signify which percentages were significantly different from 2005 and separated the years in which accommodations permitted and not permitted.

This table was more difficult for the participants to read and interpret. The participants found the abundance of percentages featuring an asterisk to be overwhelming. Because so much of the data here was significantly different, they also wanted more information about the actual meaning of “statically different.” The discussant informed them the definition was available on the Web site by clicking on “statistically different” on the table; however they thought it would be more convenient if the definition appeared in the figure.

Another point of confusion was the reporting method used by NAEP because the percentages are reported “at or above” the achievement levels. Many of the participants did not understand why the percentages for each year summed to over 100 percent. Participants commented, “This one could be easily misinterpreted because the numbers add to over 100 percent,” and “It is difficult to see where the increase is because the data is reported by percentage at or above each level.”

Figure G-3.

Percentage of students, by mathematics achievement level, grade 4: Various years, 1990–2005				
Assessment year	Below <i>Basic</i>	At or above <i>Basic</i>	At or above <i>Proficient</i>	<i>Advanced</i>
<u>Accommodations</u> not permitted				
1990	50*	50*	13*	1*
1992	41*	59*	18*	2*
1996	36*	64*	21*	2*
<u>Accommodations</u> permitted				
1996	37*	63*	21*	2*
2000	35*	65*	24*	3*
2003	23*	77*	32*	4*
2005	20	80	36	5

* Significantly different from 2005.

NOTE: Rows will not sum to 100 percent because of cumulative categories. View complete data with standard errors for [grade 4](#).

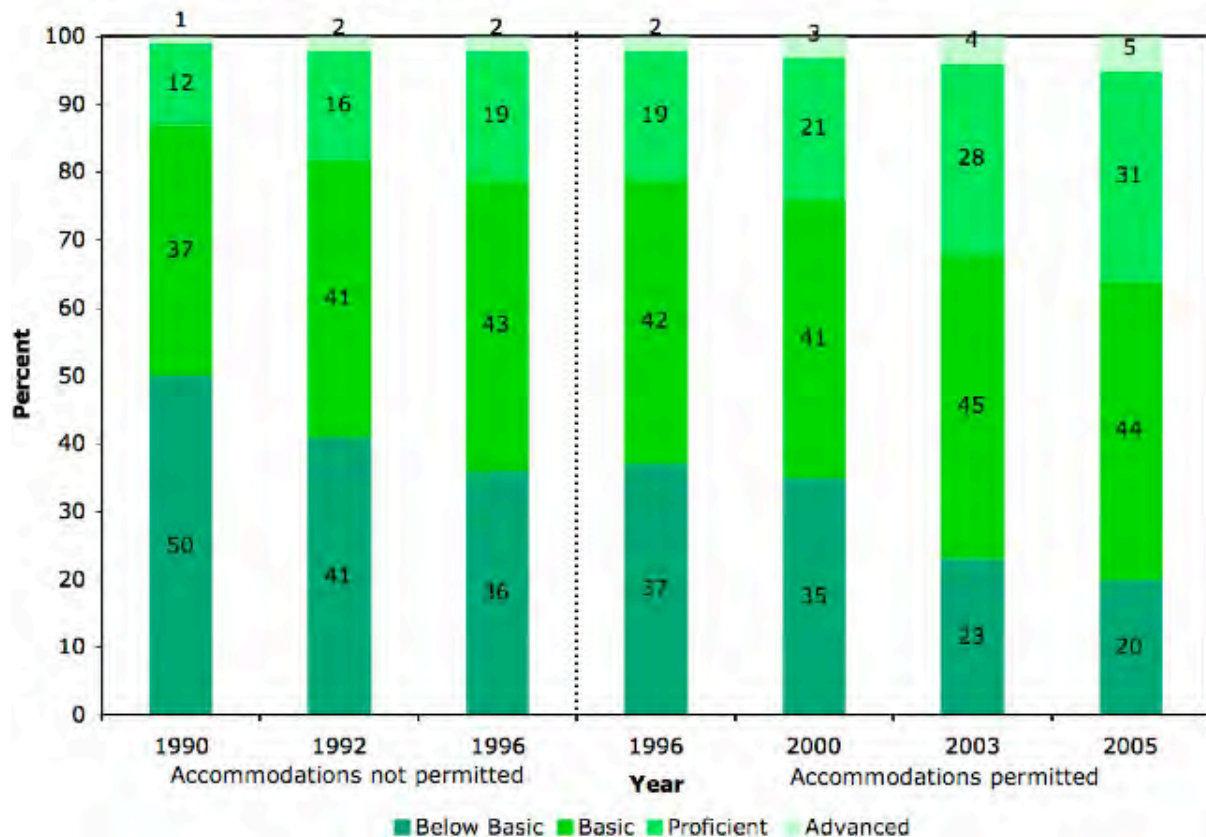
Figure G-4 is a stacked bar graph showing the percentage of grade 4 students in each achievement level category in mathematics from 1990 to 2005. This figure was not taken from the NAEP Web site; it was modeled after NAEP graphs to gauge participants’ preferences for “at or above achievement level” or discrete achievement levels reporting. In this graph, a dashed line was used to separate years in which accommodations were permitted and years in which they were not.

The participant provided a lot of positive feedback for this figure (e.g. “It is easier to interpret because the numbers sum to 100 percent,” and “I like how the information is conveyed with the bars and the numbers”). They all agreed that this type of reporting was easier to interpret than reporting the percentage “at or above” each level as was featured in the previous figure. They also noted that the overall layout of this figure made it easier to read and spot score increases at each of the different levels.

Although the participants liked this graph, they voiced a few concerns about it. First, they would have liked to see information about the statistical significance provided the definition was included. They also thought the spacing of the years should be more to scale like it was in Figures G-1 and G-2 to take into account the years in which NAEP was not administered.

Figure G-4.

Percentage of students in achievement level categories in Mathematics, Grade 4 Various Years, 1990-2005



The next three figures (Figures G-5, G-6, and G-7) were all examples of clickable state comparison maps. This interactive tool on the NAEP Web site allows the user to select a focal state or jurisdiction in which to compare the other states to. The map is color-coded to reflect how each state compares to a focal state or jurisdiction. The user also selects if comparisons are to be based on percentage at or above *Proficient* or scale scores.

Figure G-5 was a screen capture of a clickable state map showing the cross-state comparisons of average reading scale scores for grade 4 public schools in 2005. In this screen capture, the focal state selected was Oregon. The participants were easily able to identify the focal state and realized that the comparisons being made were based on scale scores. Overall the participants found the map to be a good way to convey information. However, they had some problems distinguishing between the colors displayed on the map. One participant commented, “The colors blend together. There is not enough contrast; different colors would make it easier to see.”

Figure G-6 was another screen capture of a clickable state map with Oregon as the focal state. However, this figure displayed the cross-state comparisons of the percentage of students at or above *Proficient* for eighth-grade public schools in 2005. The participants understood that the two maps were comparing different things; the first map was comparing the percentage at or above proficient and this one was comparing the average scale scores. The participants thought the maps were an innovative way to convey this type of data. Though one participant asked, “Why is it only compared to a focal state? I would like to see a map like this comparing the states to a neutral score.”

This question prompted Figure G-7, which was the last screen capture of the clickable state comparison maps. This figure displayed the cross-state comparisons of average reading scale scores for eighth-grade public schools in 2005 with the nation as the focal group.

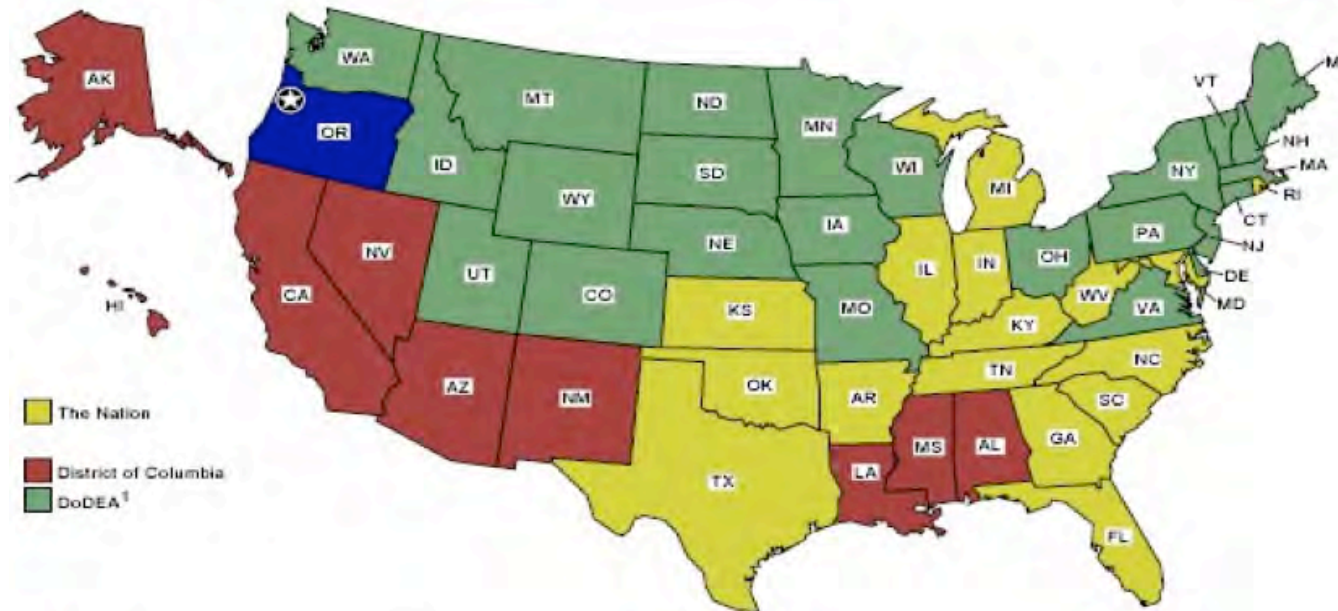
The participants were very impressed with the flexibility of the clickable state maps (one participant commented, “I like that you still see the same map, but you can manipulate the focal group and the information being reported”). There were concerns about this type of reporting for states that were performing more poorly than the focal state though. When asked if they liked these displays, participants commented, “It depends if your state is in the red [performing significantly lower than the focal state] or the green [performing significantly higher than the focal state]; I would not want to use this map if my state was in the red.”

Figure G-5.

Cross-state comparisons of average reading scale scores, grade 4 public schools: 2005

NAEP Reading Grade 4 - Reading
 Difference in Average Scale Score Between Jurisdictions
 for All students [TOTAL] = All students
 2005

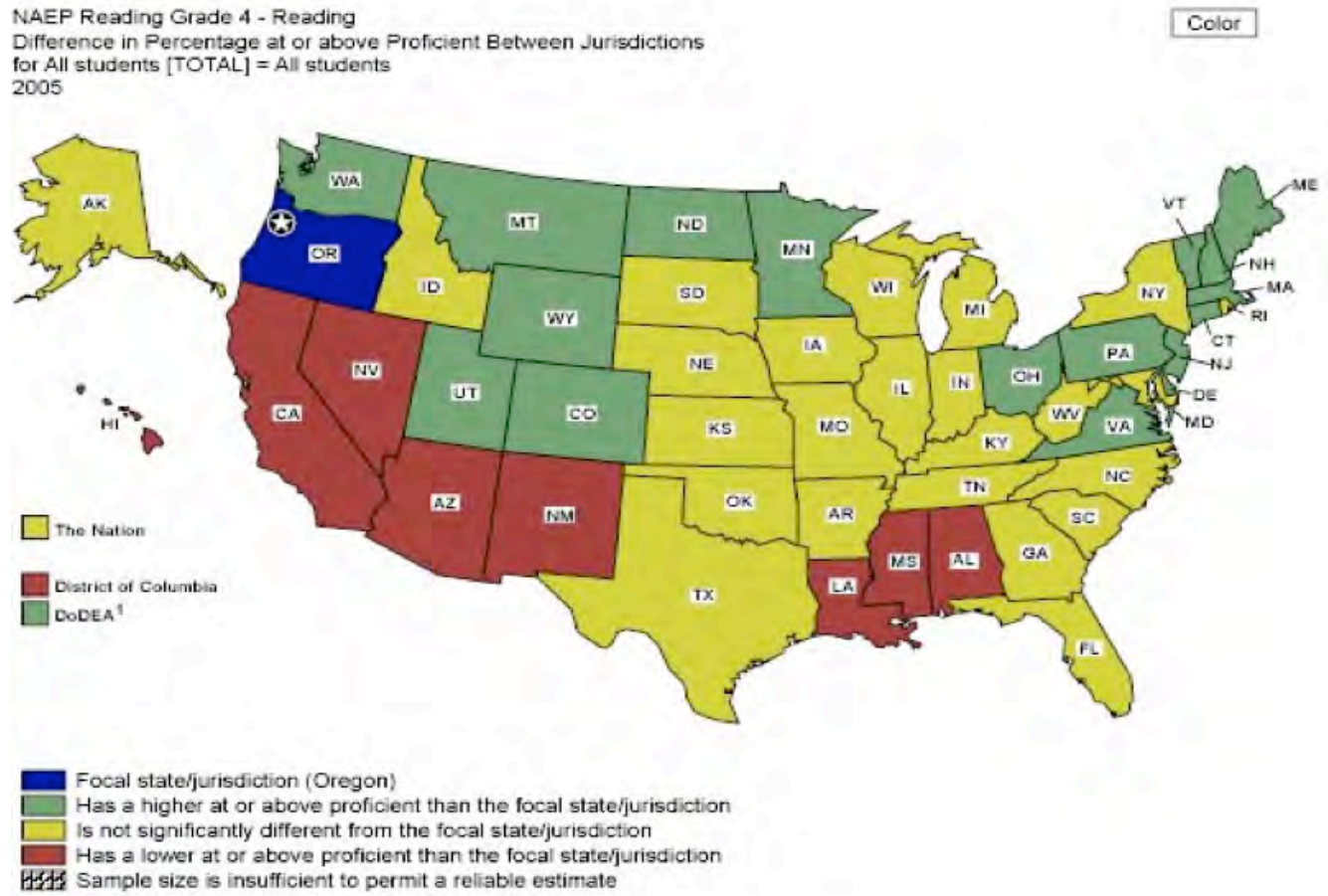
Color



- Focal state/jurisdiction (Oregon)
- Has a higher average scale score than the focal state/jurisdiction
- Is not significantly different from the focal state/jurisdiction
- Has a lower average scale score than the focal state/jurisdiction
- Sample size is insufficient to permit a reliable estimate

¹ Department of Defense Education Activity schools (domestic and overseas).

Figure G-6.
Cross-state comparisons of percentage of students at or above *Proficient* in reading, grade 4 public schools: 2005



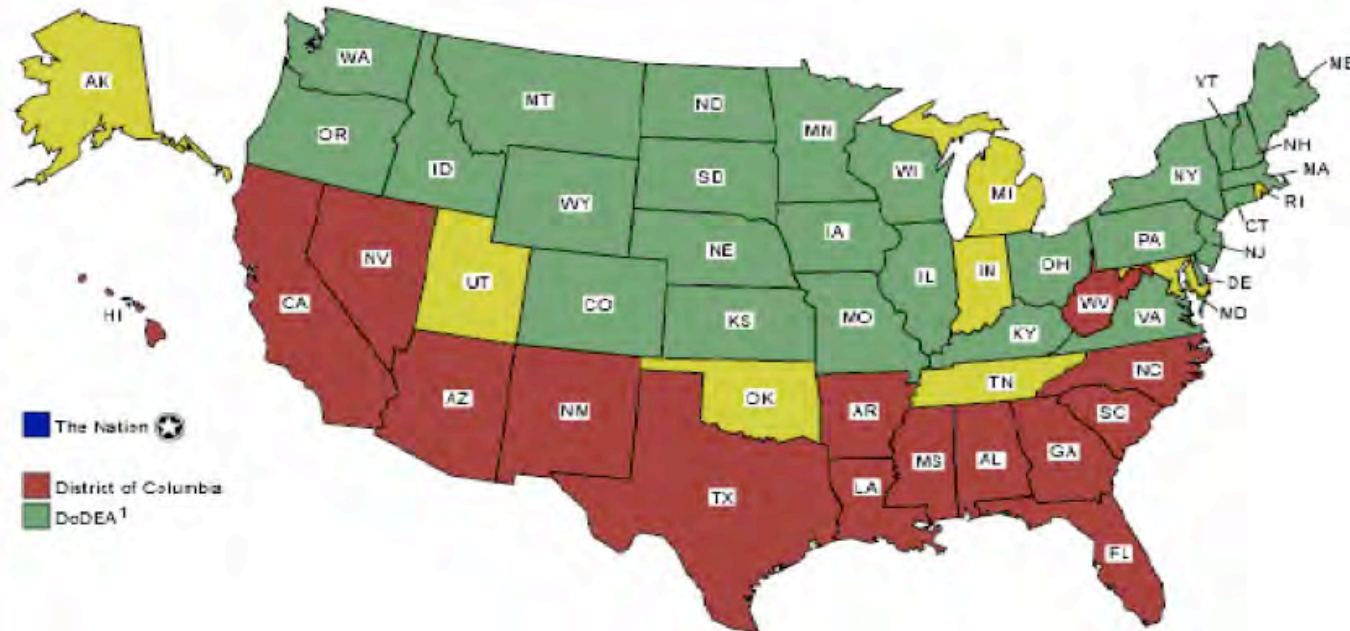
¹ Department of Defense Education Activity schools (domestic and overseas).

Figure G-7.

Cross-state comparisons of average reading scale scores, grade 8 public schools: 2005

NAEP Reading Grade 8 - Reading
 Differences in Average Scale Score Between Jurisdictions
 for All students [TOTAL] = All students
 2005

Color



- Focal state/jurisdiction (National Public)
- Has a higher average scale score than the focal state/jurisdiction
- Is not significantly different from the focal state/jurisdiction
- Has a lower average scale score than the focal state/jurisdiction
- Sample size is insufficient to permit a reliable estimate

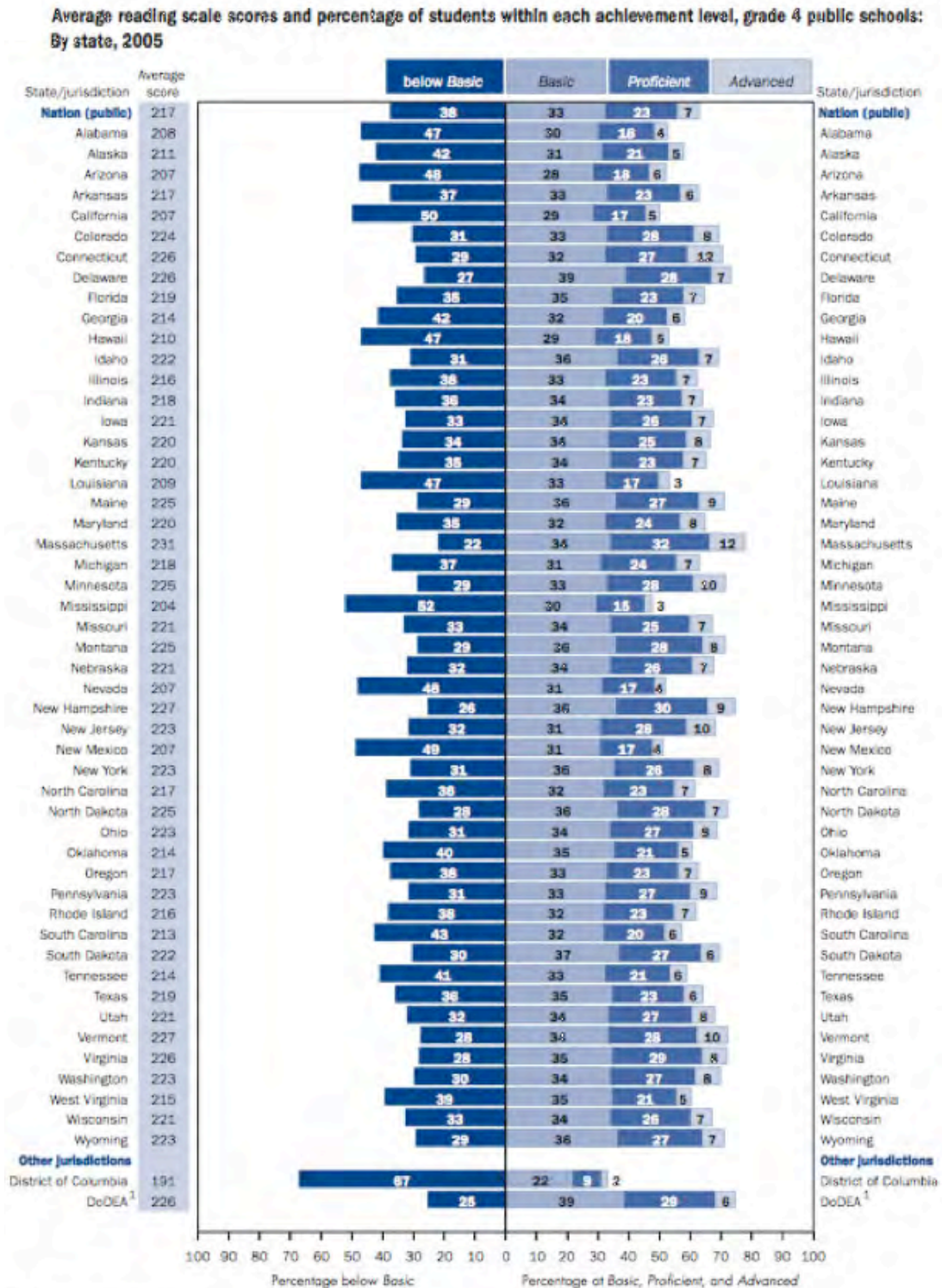
¹ Department of Defense Education Activity schools (domestic and overseas).

NOTE: View complete data with standard errors for [grade 8](#).

Figure G-8 was a bar graph showing the average reading scale scores and percentages of students within each achievement level for fourth-grade public schools by state for 2005. This graph included information for all 50 states and additional jurisdictions. The user is allowed to select which achievement level they want the percentage bars to line up at. In this example, the percentage bars lined up at the basic cut point.

The participants initially found this display overwhelming, but after examining it for a brief period, it was agreed that the graph provided a lot of good information. Some of the participants commented, “It is overwhelming; there is too much to look at and it is too close together” and “The information is useful, but it would be better if it was presented in a different way.” In response to this, one of the participants suggested, “It would probably be easier to see if the states were sorted by performance rather than alphabetically.”

Figure G-8.



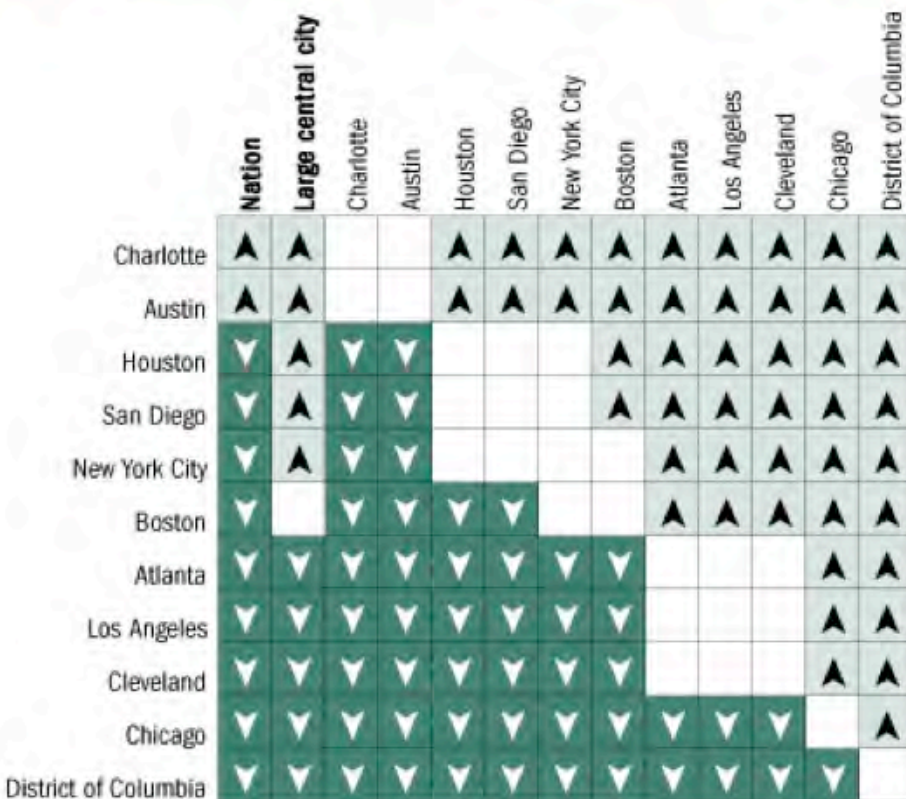
¹ Department of Defense Education Activity.
² NAEP reading scale ranges from 0 to 500. Detail may not sum to totals because of rounding. The shaded bars are graphed using unrounded numbers.
 U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment.

The next display was a version of a “pantyhose” chart showing the overall cross-district comparisons of average mathematics scale scores for fourth-grade public schools in 2005 (Figure G-9). When asked what kind of information the chart was conveying, panelists were able to tell it was comparing districts. They recognized that the lightly shaded boxes with the dark arrows pointing up represented that the city to left was scoring higher than the corresponding city at the top of the chart and that the darker boxes with the white arrows pointing downwards meant the opposite. When the participants were asked how Boston was performing relative to the other districts there was some uncertainty (one participant responded, “It is in the middle, right?”).

Although the participants managed to read the chart appropriately, it required closer scrutiny than many of the previous displays. One participant commented, “it’s confusing; I thought you could read it down or across at first.” It was also noted that the information displayed in this chart would not be valuable to states that have no large cities. The participants all agreed that the state maps were a better tool for making comparisons.

Figure G-9.

Overall cross-district comparisons of average mathematics scale scores, grade 4 public schools: 2005






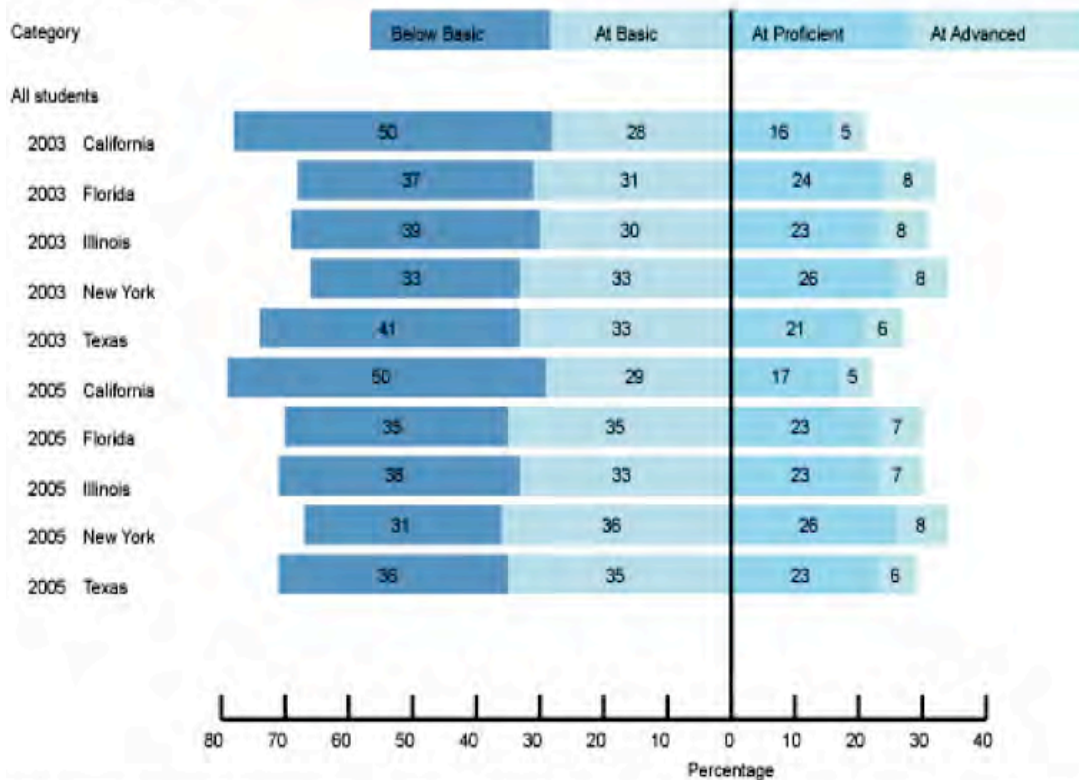
-  District had higher average scale score than the district listed at the top of the column.
-  No statistically significant difference detected from the district listed at the top of the column.
-  District had lower average scale score than the district listed at the top of the column.

Figure G-10 was a bar graph showing the percentages of fourth-grade students at each achievement level for reading for 2003 and 2005. This figure was created by the researchers using the NAEP Data Explorer, which permits the user the flexibility to choose the states/jurisdictions to be included (here, five states were selected: California, Florida, Illinois, New York, and Texas) for comparison. In addition, the researchers chose to have the percentage bars aligned at the proficient level.

The participants were easily able to read this graph, but they had several comments about the layout of the graph. One participant suggested, “There should be a line clearly dividing the 2 years.” Another participant raised a question about the significance of the differences between the percentages because it was not noted on the graph. Overall, the group liked the flexibility of being able to select which states to display and which achievement level they would like the percentage bars aligned at.

Figure G-10.

Percentages of students at each achievement level for reading, grade 4
All students [TOTAL]
By jurisdiction, 2003 and 2005



NOTE: Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.

SOURCE: U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. National Assessment of Educational Progress (NAEP).

At this point, the discussant asked the participants to look back over the previous ten data displays figures and share some comments about which figures they believe are the best way to convey state comparison information. The participants liked the state maps for comparing states at one point in time. For comparing over time, they preferred Figure G-1, although participants commented, “They should either display percentages or define the scale” and another suggested displaying a “target zone” on the graph showing what level the students should be working towards. Another display of interest was Figure G-4. The participants also stressed that they would like to see information about the size of the samples tested in each state on the displays. When asked about how they would prefer to receive the results, either online or a paper report, the general consensus was that online reports were much more flexible and readily available than printed reports.

The next displays shown were all examples of the type of information users can extract from the NAEP Question Tool, located on the NAEP Web site at <http://nces.ed.gov/nationsreportcard/itmrls/>. The images shown included a sample passage from the NAEP fourth-grade reading assessment administered in 1998 and a sample question corresponding to the passage, as well as the percentage of students that responded to this question correct, incorrect, or omitted, and the distracter analysis. Participants also saw examples of open response passages, questions, scoring guides, and sample student responses.

The discussion focused mainly on Figure G-11, the distracter analysis. Overall, the participants did not think people would find the distracter analysis interesting because the test is reported at the state level, not the individual level. The participants were impressed with immense amount of information provided, though many were not familiar with the question tool.

Figure G-11.

NAEP National Reading Grade 4 1998 Accommodations Not Permitted
 Average Scale Score with Percentages (with Standard Errors in Parentheses), Reading
Blue Crabs: Common with arthropods-skeleton [R012202]

All students																
	A *			B			C			D			Omitted		Missing	
	Avg.	Row		Avg.	Row		Avg.	Row		Avg.	Row		Avg.	Row	Avg.	Row
	Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)	Score (S.E.)	Pct. (S.E.)
All students	230 (1.2)	57 (1.4)		205 (2.2)	19 (1.1)		199 (3.4)	10 (0.8)		203 (2.7)	14 (1.0)		‡ (‡)	# (0.1)	‡ (‡)	1 (0.2)
Gender																
	A *			B			C			D			Omitted		Missing	
	Avg.	Row		Avg.	Row		Avg.	Row		Avg.	Row		Avg.	Row	Avg.	Row
	Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)		Score (S.E.)	Pct. (S.E.)	Score (S.E.)	Pct. (S.E.)
Male	227 (1.6)	59 (1.9)		203 (3.6)	19 (1.6)		192 (5.2)	10 (1.1)		200 (4.1)	13 (1.4)		‡ (‡)	# (***)	‡ (‡)	1 (0.4)
Female	233 (1.6)	55 (1.7)		207 (2.6)	20 (1.3)		205 (4.1)	9 (1.1)		206 (3.2)	15 (1.2)		‡ (‡)	# (0.2)	‡ (‡)	1 (0.3)

† Accommodations were not permitted for this assessment.

Percentage rounds to zero.

‡ Sample size is insufficient to permit a reliable estimate.

(***) Standard error estimates cannot be accurately determined.

NOTE: The NAEP Reading scale ranges from 0 to 500. Observed differences are not necessarily statistically significant. Detail may not sum to totals because of rounding.

Figure G-12 was an item map for the eighth-grade reading scale. At first the participants were unsure exactly how this information would be useful. After further examination they realized that this item map would be useful in addition to several of the other figures that display scale scores because this explains what the scale scores mean. The participants were slightly confused on how to interpret this figure. The discussant pointed out the caption and the foot note at the bottom of the figure which explain that the skills listed next to the scale scores correspond to items that students performing at the specified level had a high probability of correctly answering. After realizing how to interpret it, the participants all agreed that it was necessary to display this type of data if other charts and graphs were reporting scale scores (one participant commented, “It adds context to the graphs and charts that report average scale scores.”)

Figure G-12.



¹ Each grade 8 reading question in the 2005 reading assessment was mapped onto the NAEP 0-500 reading scale. The position of a question on the scale represents the average scale score attained by students who had a 65 percent probability of successfully answering a constructed-response question, or a 74 percent probability of correctly answering a four-option multiple-choice question. Only selected questions are presented. Scale score ranges for reading achievement levels are referenced on the map. For constructed-response questions, the question description represents students' performance at the scoring level being mapped.

NOTE: Regular type denotes a constructed-response question. Italic type denotes a multiple-choice question.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment.

The last few displays (shown in Figures G-13 through G-17) were all examples of ways NAEP reports gap data. In Figure G-13, two lines representing the performance of male and female students aged 9 from 1973 to 2004 were presented. The region between the two lines was shaded to represent the score gap and the difference in scores at each administration was displayed just above the two lines. The participants liked the concept of displaying gap data this way, but they found this display confusing because it was difficult to distinguish which of the lines was intended to represent females and males, especially because the lines crossed in the late 1980s. They suggested adding a legend clearly explaining which line represented each group as well as making the colors of the lines more distinguishable.

Figure 14 was another example of male/female gap data. This figure displayed the differences in average scale score between the two groups from 1990 to 2003. A gray box was used to denote the administration in which accommodations were permitted. The “#” symbol was used to signify years in which the estimated difference in scale scores rounded to zero. All of the participants agreed that Figure G-13 was a better way to display gap data because it provided more information and it was easier to see the gaps.

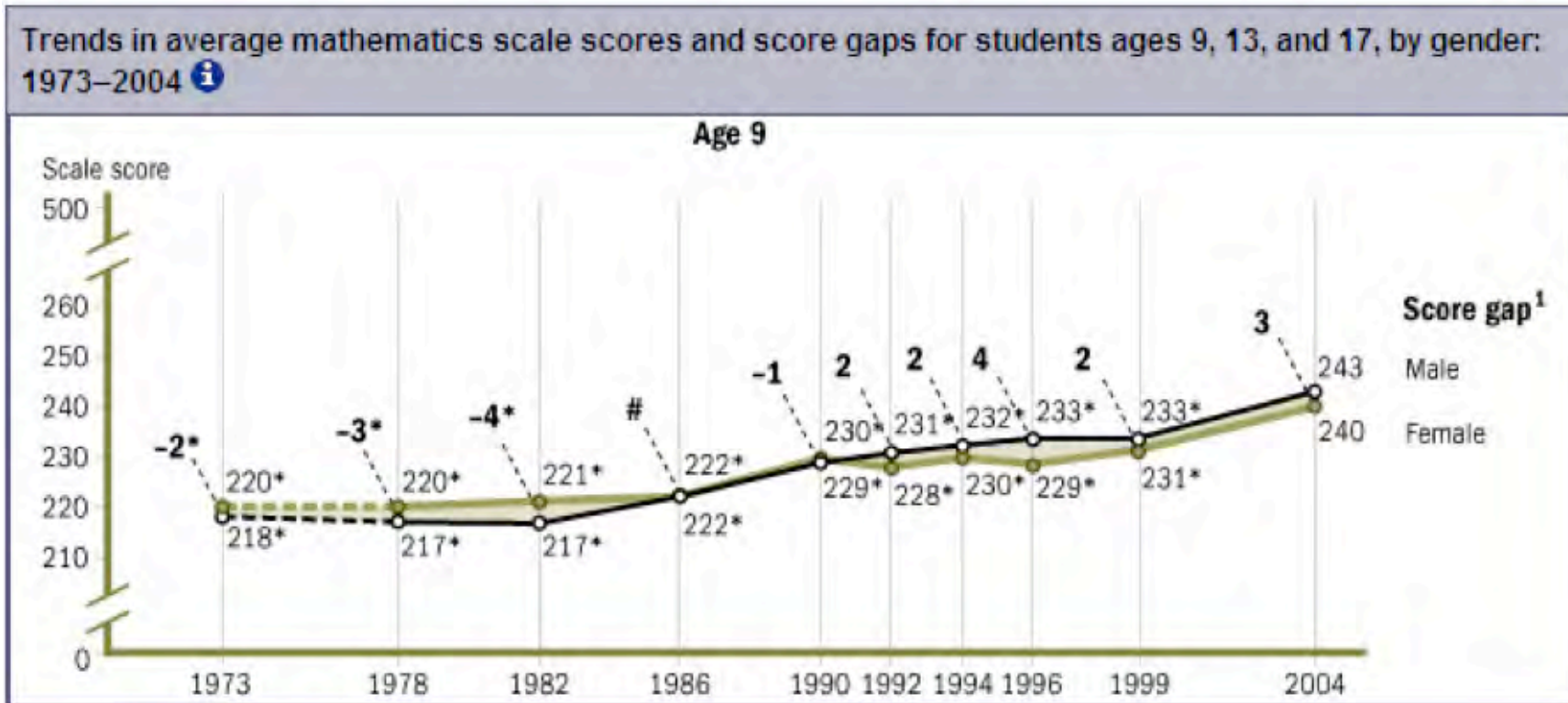
The next figure was another line graph showing gap data. Figure G-15 displayed two lines, representing the performance of white and Hispanic students on the grade 8 mathematics assessment from 1990 to 2005. The region between the lines was shaded and the difference in scale scores between the two groups was displayed. A dashed line denoted the years in which accommodations were permitted and an asterisk was again used to show values that were statically different from 2005. The participants found this graph much easier to read than the previous two. They liked that it included a legend with different line types so it was easy to tell which line represented which group because the colors were still difficult to distinguish between.

The last figure shown to the participants (Figure G-16) was a bar graph displaying male and female gap information from 1990 to 2005. The bars represented the percentage of students at or above *Basic* and at or above *Proficient* with the male percentages on the top and the female percentages below. Comments on this graph ranged from positive (“This is similar to figure G-13, but this one shows the achievement levels; I like Figure G-16 better”) to negative (“It is too busy; with both male and female information and the two boxes; it is just too much.”) Participants suggested it would be easier to interpret if the male and female data was side by side on the same axis rather than one graph on top of the other.

Figure G-17 also displayed gap data; however, this figure was not discussed due to time constraints.

The discussant then asked the participants which figures they felt were best to display gap information. Overall the participants agreed line graphs (Figures G-13 and G-15 in particular) were the best way to present the gap data.

Figure G-13.



The estimate rounds to zero.

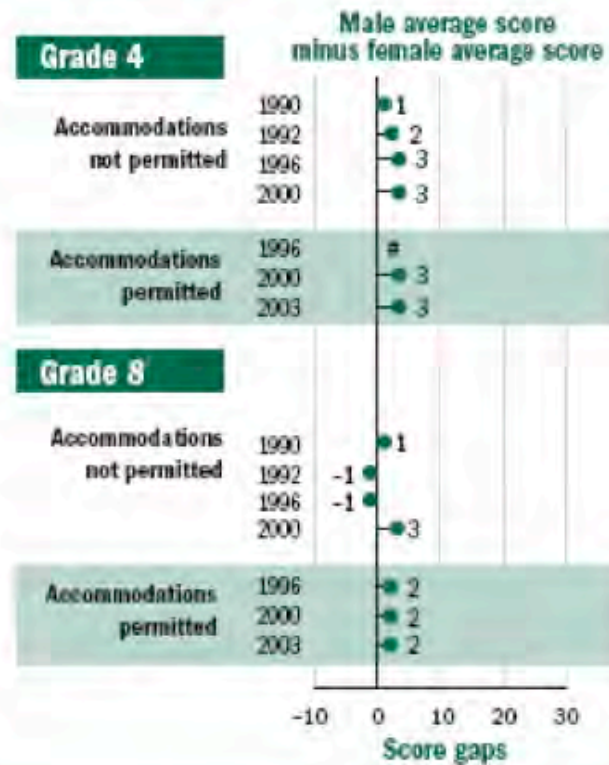
* Significantly different from 2004.

¹ Male average scale score minus female average scale score. Negative numbers indicate that the average scale score for male students was lower than the score for female students.

NOTE: Dashed lines represent extrapolated data. Score gaps are calculated based on differences between unrounded average scale scores. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), selected years, 1973–2004 Long-Term Trend Mathematics Assessments.

Figure G-14.

Gaps in average mathematics scale scores, by gender, grades 4 and 8: 1990-2003



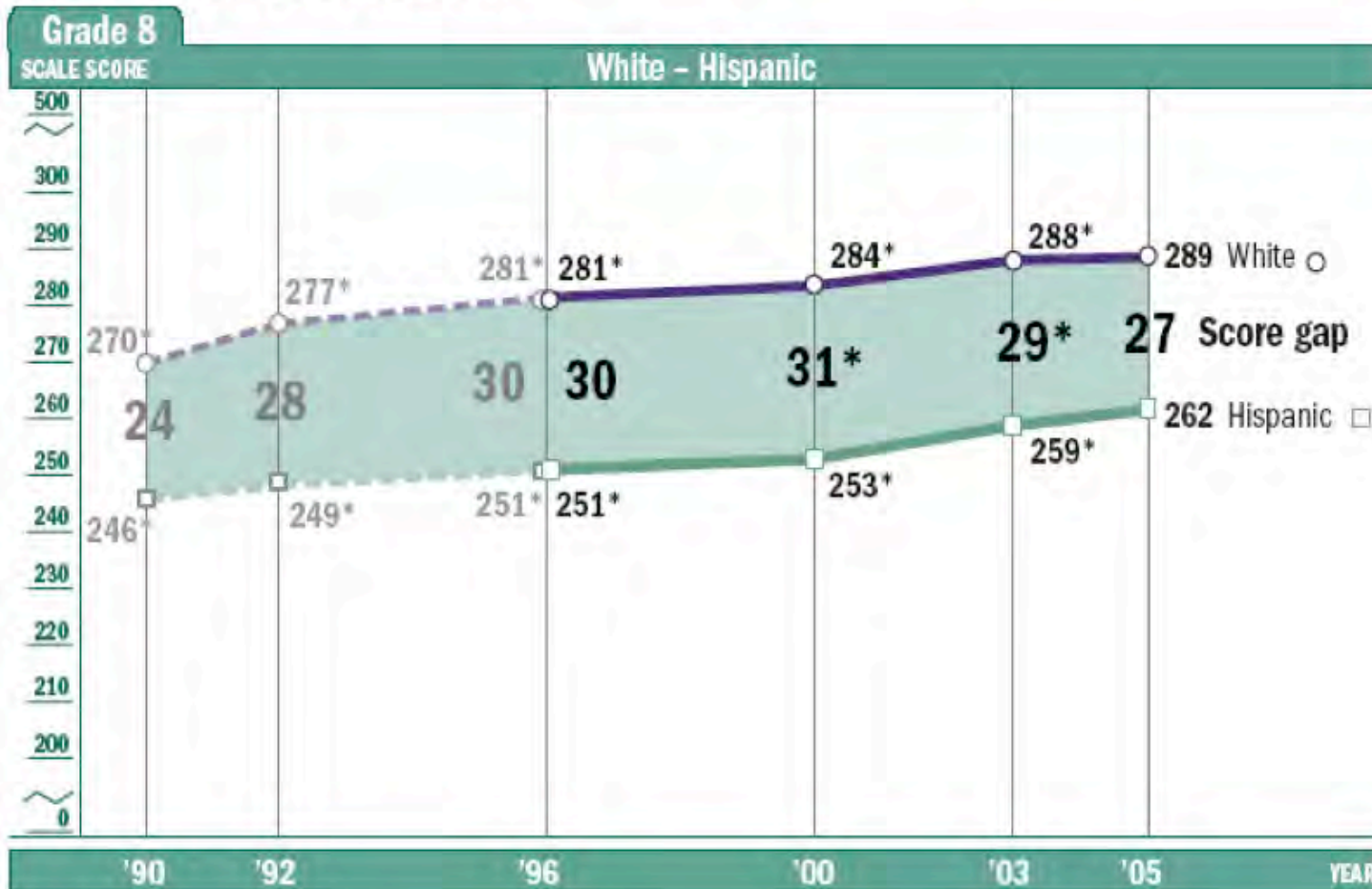
The estimate rounds to zero.

NOTE: In addition to allowing for accommodations, the accommodations-permitted results (1996-2003) differ slightly from previous years' results, and from previously reported results for 1996 and 2000, due to changes in sample weighting procedures. See appendix A for more details. Score gaps are calculated based on differences between unrounded average scale scores. NAEP sample sizes have increased in 2003, compared to previous years, resulting in smaller detectable differences than in previous assessments. Negative numbers indicate that the average score for male students was lower than the score for female students.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1990, 1992, 1996, 2000, and 2003 Mathematics Assessments.

Figure G-15.

Average mathematics scale scores and score gaps for White-Hispanic students, grade 8: Various years, 1990-2005



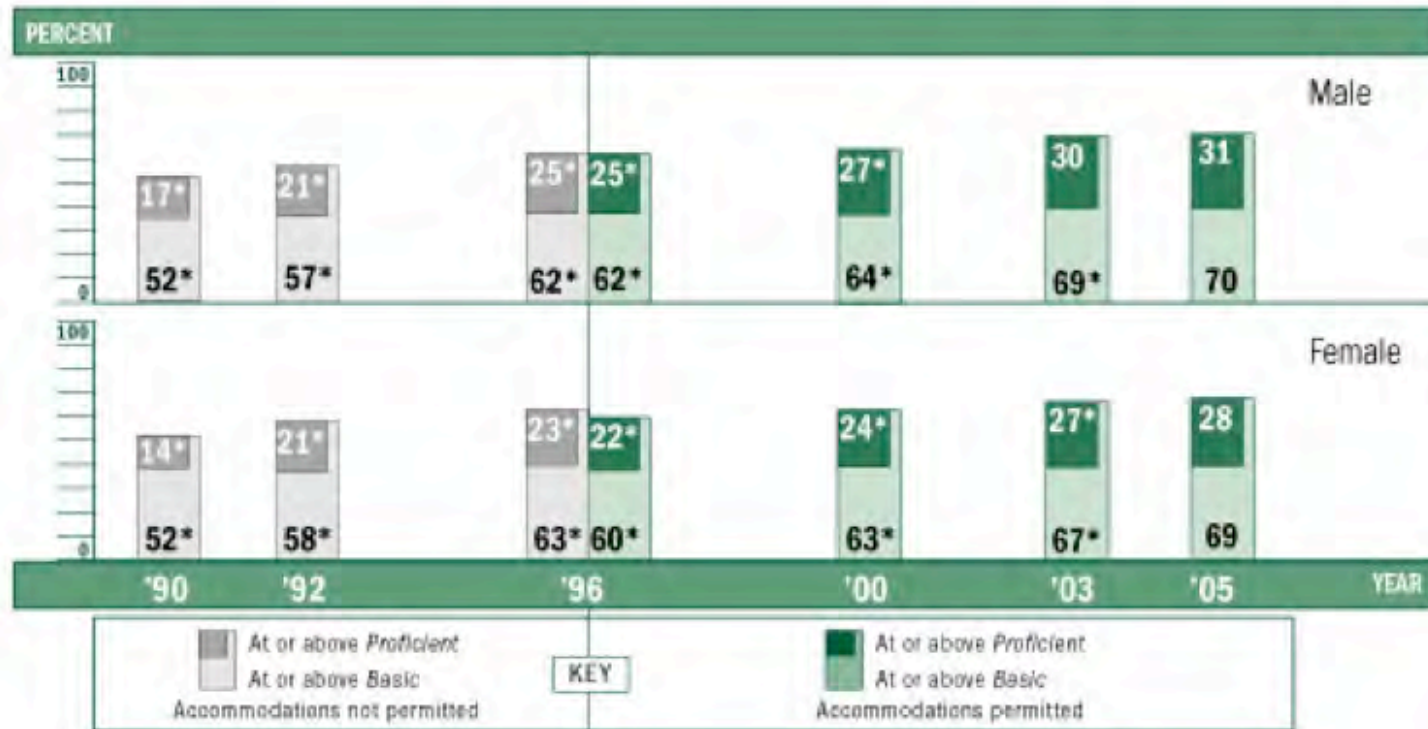
* Significantly different from 2005.

NOTE: Score gaps, displayed in the shaded area, are calculated based on differences between unrounded average scale scores.

- Accommodations not permitted
- Accommodations permitted

Figure G-16.

Achievement-level results in mathematics, by gender, grade 8: Various years, 1990-2005

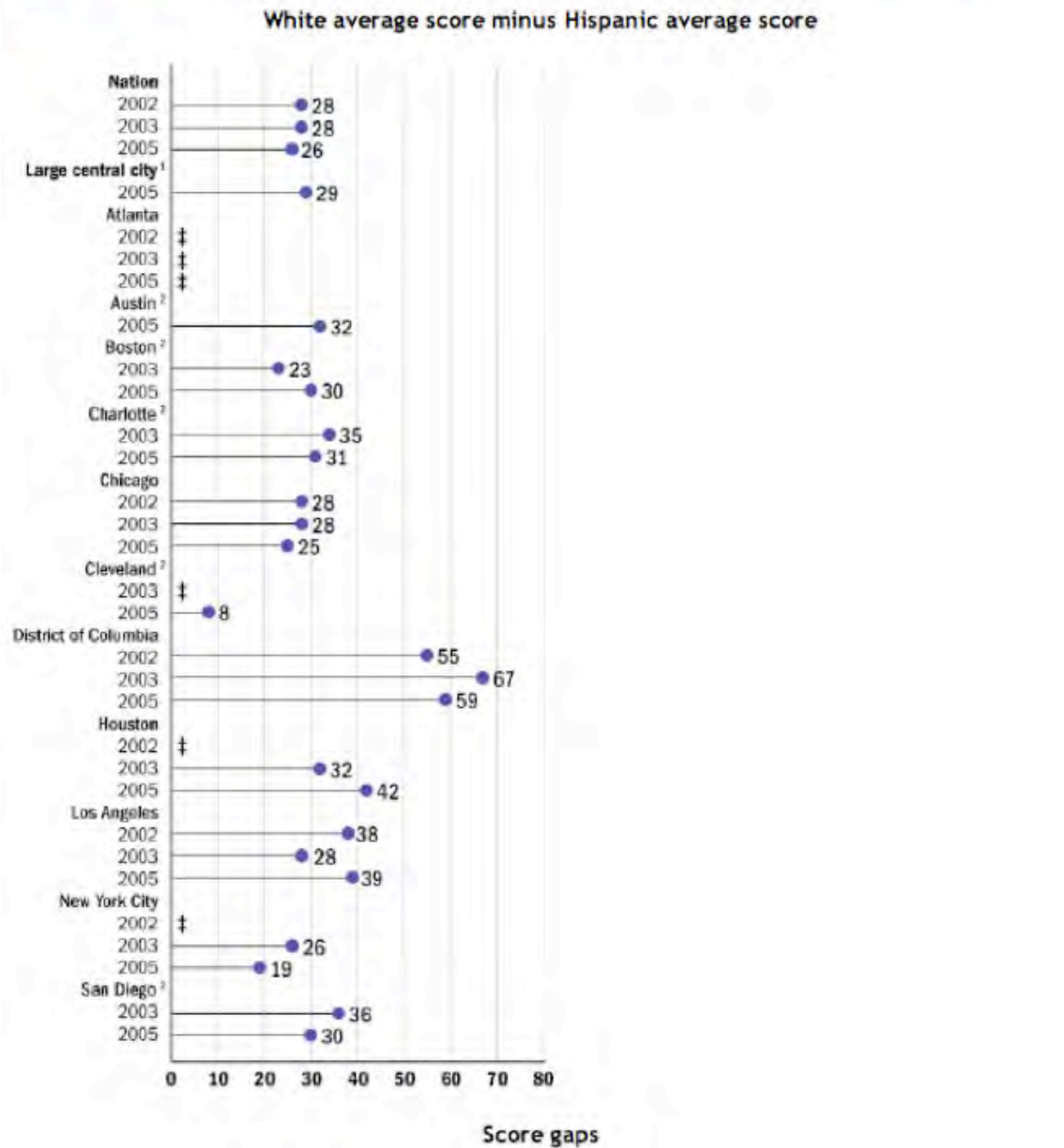


* Significantly different from 2005.

NOTE: The gray shaded boxes represent results based on administrations when accommodations were not permitted. View complete data with standard errors for [grade 8](#).

Figure G-17.

White-Hispanic gap in average reading scores, grade 4 public schools: By urban district, various years, 2002-2005



‡ Reporting standards not met.

¹ Data for large central city schools are not included for years prior to 2005 because the application of the definitions of the types of location has changed. For 2005, "large central city" includes nationally representative public schools located in large central cities (population of 250,000 or more) within a Metropolitan Statistical Area (MSA).

² The district did not participate in 2002 or 2003.

NOTE: Score gaps are calculated based on differences between unrounded average scale scores.

DATA: View complete data with standard errors for scale scores in each district: [Atlanta](#), [Austin](#), [Boston](#), [Charlotte](#), [Chicago](#), [Cleveland](#), [District of Columbia](#), [Houston](#), [Los Angeles](#), [New York City](#), and [San Diego](#).

For the remainder of the meeting, the discussant asked the participants more general questions to open discussion about NAEP and state score reporting methods. The participants were asked to reflect upon all of the displays they were shown and comment on the figures they thought would be useful for reporting scores. There was a general agreement that figures that were less data rich, but easy to read and interpret (e.g. Figures G-4, G-10, and G-15) would be more useful for all audiences. Participants commented, “They [Figures G-4 and G-15] are not too overwhelming,” and “These [Figures G-4, G-5, G-10, and G-15] are good for presentations when there is limited time to make a point, especially Figures G-4 and G-15 because they are easily interpretable and lead to conversation.” The participants also thought the item map, shown in Figure G-12, was valuable especially as a complement to several of the other figures that reported the scale scores.

The discussant also asked if the participants would like to see additional information included in score reports. Overall, the participants expressed interest in seeing more information on subgroup performance. The group also wanted to see information about the sample sizes and subgroup sample sizes for the data reported.

Discussion

The participants in this focus group provided researchers with important insights about how educators use and interpret NAEP data displays. Overall, the group expressed a preference for graphs that are less data rich, but more readily interpretable (e.g. the clickable state comparisons maps; Figures G-5, G-6, and G-7). Also, one area that caused confusion among the group was the terminology and reporting mechanisms used by NAEP. The participants found it difficult to interpret the graphs in which the scores were reported at or above a given achievement level. Although Figures G-2 and G-4 were displaying similar information, all of the participants agreed that Figure G-4 was easier to read and interpret simply because the scores were reported within each level and the percentages summed to 100 percent. Overall the participants were more interested in the percentage of students performing within each performance level and found it more intuitive to interpret information displayed in this manner.

Another concept that caused confusion among the group was the NAEP scale scores. All of the participants had reported they were somewhat familiar with NAEP, yet none of them seemed familiar with the NAEP scale. When figures displaying scale score data were shown, a lot of questions were raised about the comparability of the NAEP scale to the scales used by state testing programs. Initially the participants found these figures confusing, but when presented with the item map (Figure G-12) they recognized that they could use it to interpret the meaning of the scale scores. Participants expressed concern that people may not understand how the item map could be utilized and recommended that it be displayed with graphs that report score scale data to add context to the scores.

Another recurring source of difficulty throughout the meeting was the inconsistencies in the layout of various figures. First of all, the axes on the graphs were not consistent when displaying units of time. The axes should be consistent across graphs and level appropriate spacing for years in which NAEP is not administered. Also, when the axis is measuring percent and it sums to over 100 percent as in the figures reporting at or above a given level, this should be emphasized and clearly explained on the graph.

A lot of attention was focused on the footnotes and display keys. The participants found it challenging to interpret several of the footnotes and legends found on the displays. They questioned the inconsistency of the terminology throughout the various figures (e.g. some figures used the term “district” and others used “jurisdiction”). There were also a lot of questions about how NAEP reports statistically significant results. Some of the figures made it explicit significant results were being reported and others provided no information and participants had to make assumptions.

The participants provided feedback about the displays as well as some additional information they would like to see in future score reports. First, all of the participants preferred the timeliness and flexibility of the online score reporting. They offered a lot of positive

feedback about several of the innovative displays and the Question Tool. However, there were also several things they felt were lacking in the displays. The participants were interested in knowing the sample sizes for each administration, as well as the sample sizes of the subgroup where it was applicable. There was uncertainty about how varying sample sizes may affect the scores. They also expressed great interest in seeing more data on subgroup performance.

In sum, recommendations from this focus group reinforced and clarified the findings of the studies described in Appendices E and F. Overall, the participants expressed a desire for consistency of layout and terminology within and across graphs. The data should be displayed in a straightforward manner with explicit and easy to understand footnotes and legends. These should also include links to additional information where it is applicable (e.g. if a graph is reporting scale score, there should be a link to an item map to add meaning to the scores). The use of color was encouraged; however, in several of the displays it was difficult to distinguish between the colors. Adding more contrast to the colors would clarify several of the graphs.

Next Steps

There are several important directions for follow-up research. First, while the use of group discussion clearly yielded much useful information, a logical next step is to carry out one-on-one explorations of several data displays with NAEP data users as they navigate themselves through different portions of the NAEP Web site. In this way, we could gather more information about how different individuals fare with respect to both knowledge and interpretation of several of the more interactive features of the site, including the clickable state maps, the Question Tool, and the NAEP Data Explorer.

In addition, given the kinds of suggestions made about these displays, one additional direction for future research includes the development of several redesigns of the displays shown here, with research participants comparing current and revised displays for clarity and understanding. This idea was pursued by Wainer, Hambleton, and Meara (1999) and produced some interesting findings. While NAEP is clearly at the forefront of testing programs with respect to its investment in methods for disseminating results, the results of this focus group indicated that there remain some sources of confusion among audiences who have some familiarity and regular use of NAEP. Clearly substantially more research and development work is needed in the future.

References

- Perie, M., Grigg, W. S., and Dion, G. S. (2005). *The Nation's Report Card: Mathematics 2005* (NCES 2006-453). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U. S. Department of Education.
- Rampey, B. D., Lutkus, A. D., and Dion, G. S. (2006). *The Nation's Report Card: Trial Urban District Assessment Mathematics 2005* (NCES 2006-457r). U.S. Department of Education, National Center for Education Statistics. Washington, D.C.: U. S. Department of Education.
- Wainer, H. (2000). Cholera, rocket ships, and Tom's veggies: Contemporary and historical ideas toward the effective communication of school performance. *Evaluation and Research in Education*, 14, 148–180.
- Wainer, H., Hambleton, R.K., and Meara, K. (2000). Alternative displays for communicating NAEP results: a redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335.

Chapter 5:
Evaluating Score Equity Across Selected States
for the 2005 Grade 8 NAEP Math and Reading Assessments

Craig S. Wells, Su Baldwin, Ronald K. Hambleton, Stephen G. Sireci,
Ana Karantonis, Stephen Jirka, Robert Keller and Lisa A. Keller
Center for Educational Assessment
University of Massachusetts Amherst

This page intentionally left blank

Contents

List of Figures and Tables.....	5-v
Executive Summary.....	5-vii
Introduction.....	5-1
NAEP Equating Procedures.....	5-2
Method.....	5-3
ETS's Operational Procedure.....	5-5
Item Parameter Calibration.....	5-5
Obtaining Plausible Values.....	5-5
Placing the Proficiency Estimates Onto the NAEP Reporting Scale.....	5-6
Study 1: Classification Consistency.....	5-7
Item Parameter Calibration.....	5-7
Obtaining Plausible Values.....	5-8
Study 2: Item Parameter Invariance Across Selected States.....	5-13
Study 3: Comparison of Test Characteristic Curves Within a Year.....	5-19
Empirical Distribution for RMSD.....	5-26
Study 4: Comparison of TCCs Between 2005 and 2003.....	5-29
Conclusion.....	5-31
References.....	5-33
Appendixes.....	5-35
Appendix A: <i>a</i> - and <i>b</i> -plots for 2005, Grade 8, Math Assessment.....	5-37
Appendix B: <i>a</i> - and <i>b</i> -plots for 2005 and 2003, Grade 8, Reading Assessment.....	5-69
Appendix C: Empirical Distribution for <i>RMSD</i> , Based on the Math Assessment, for Three Sample Sizes.....	5-131
Appendix D: Empirical Distribution for <i>RMSD</i> , Based on the Reading Assessment, for Four Sample Sizes.....	5-135
Appendix E: Test Characteristic Curves for 2005 and 2003, Grade 8, Math Assessment.....	5-141
Appendix F: Test Characteristics Curves for 2005 and 2003, Grade 8, Reading Assessment.....	5-149

This page left intentionally blank

Figures and Tables

Figures

Figure 1. A comparison of the smoothed proficiency distribution for Texas based on the state and national item parameter estimates for the Reading assessment.....	5-10
Figure 2. <i>b</i> -parameter estimates for California and the national sample.....	5-13
Figure 3. Test characteristic curves for each state compared to the national sample for the for the Math assessment.....	5-20
Figure 4. Test characteristic curves for each 2005 state sample compared to the 2005 national sample for the Reading assessment.....	5-23
Figure A-1. 2005 NAEP Math Gr 8 a- and b-plots: Selected States vs. National.....	5-38
Figure A-2. 2005 NAEP Math Gr 8 a- and b-plots: Selected States vs. States.....	5-48
Figure B-1. 2005 NAEP Reading Gr 8 a- and b-plots: Selected States vs. National....	5-70
Figure B-2. 2005 NAEP Reading Gr 8 a- and b-plots: Selected States vs. States.....	5-80
Figure B-3. 2003 Reading Grade 8 a- and b-plots: Selected States vs. National.....	5-100
Figure B-4. 2003 Reading Grade 8 a- and b-plots: Selected States vs. States.....	5-110
Figure C-1. Empirical Distribution for RMSD, Based on the Math Assessment for Three Sample Sizes.....	5-132
Figure D-1. Empirical Distribution for RMSD, Based on the Reading for Four Sample Sizes.....	5-136
Figure E-1. TCCs for the 2005 and 2003, Grade 8, Math Assessment for the National Sample.....	5-142
Figure E-2. TCCs for the 2005 and 2003, Grade 8, Math Assessment for Florida.....	5-143
Figure E-3. TCCs for the 2005 and 2003, Grade 8, Math Assessment for Massachusetts.....	5-144
Figure E-4. TCCs for the 2005 and 2003, Grade 8, Math Assessment for California	5-145
Figure E-5. TCCs for the 2005 and 2003, Grade 8, Math Assessment for North Carolina.....	5-146
Figure E-6. TCCs for the 2005 and 2003, Grade 8, Math Assessment for Oklahoma.....	5-147
Figure F-1. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for the National Sample.....	5-150
Figure F-2. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for Florida.....	5-151
Figure F-3. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for Massachusetts.....	5-152
Figure F-4. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for California	5-153
Figure F-5. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for North Carolina.....	5-154
Figure F-6. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for Oklahoma.....	5-155

Tables

Table 1. Weights used to construct the composite score for the Math assessment.....	5-6
Table 2. Weights used to construct the composite score for the Reading assessment.....	5-6
Table 3. Correlation coefficients between item parameter estimates we calibrated and ETS's estimates.....	5-8

Table 4. Achievement level results for Math: National versus State item parameter estimates.....	5-9
Table 5. Achievement level results for Reading: National versus State item parameter Estimates.....	5-10
Table 6. Proportion of 2005 Grade 8 Math items exhibiting item parameter estimate differences beyond three standard errors.....	5-14
Table 7. Proportion of 2005 Grade 8 Reading items exhibiting item parameter estimate differences beyond three standard errors.....	5-15
Table 8. Mean difference between a- and b-parameter estimates identified to function differentially for Grade 8 Math, 2005.....	5-16
Table 9. Mean difference between a- and b-parameter estimates identified to function differentially for Grade 8 Reading, 2005.....	5-16
Table 10. Proportion of 2003 Grade 8 Math items exhibiting item parameter estimate differences beyond three standard errors.....	5-16
Table 11. Proportion of 2003 Grade 8 Reading items exhibiting item parameter estimate differences beyond three standard errors.....	5-17
Table 12. Mean difference between a- and b-parameter estimates identified to function differentially for Grade 8 Math, 2003.....	5-17
Table 13. Mean difference between a- and b-parameter estimates identified to function differentially for Grade 8 Reading, 2003.....	5-17
Table 14. RMSD for each comparison between the TCC for each respective state and the national sample for the 2005 Math assessment.....	5-25
Table 15. RMSD for each comparison between the TCC for each respective state and the national sample for the 2005 Reading assessment.....	5-26
Table 16. Percentile of the observed RMSD for each state relative to the appropriate empirical distribution for the Math assessment.....	5-27
Table 17. Percentile of the observed RMSD for each state relative to the appropriate empirical distribution for the Reading assessment.....	5-27
Table 18. RMSD and UAD values summarizing the difference between raw scores in the conversion table for the national sample and for each state.....	5-30

Executive Summary

Score equity assessment is an important concept that examines whether a test measures the same construct across distinct subgroups. The purpose of the present evaluation was to assess the extent to which the Grade 8 NAEP Math and Reading assessments for 2005 were equivalent across selected states that varied with respect to the alignment of the state assessment with NAEP. Score equity was assessed by examining invariance of the NAEP reporting scale for the selected states when treated separately versus as part of the entire national sample. Since developing the NAEP reporting scale is complex, we examined score equity via four separate but conceptually related studies.

Study 1 examined the consistency of the achievement level results (i.e., Basic, Proficient, and Advanced) across the five selected states for each assessment when they were treated separately versus as part of the national sample. Study 1 was the primary evaluation because it highlighted the practical consequences that any lack of invariance may have on the lack of score equity. Following the operational procedure closely (i.e., item parameter calibration, proficiency estimates via plausible value methodology, and equating), the proportion of examinees within each achievement level was obtained using the item statistics for the respective state and the national sample. The achievement level proportions were comparable for each state comparison for the Math and Reading assessment. Because the percentages of examinees within each category were similar for each state determined separately versus as part of the national sample, it appeared that any lack of score equity with respect to the five studied states was minimal. In addition, it appeared that the alignment of the state's assessment program to NAEP did not have a meaningful impact.

Study 2 examined whether the item parameter values (i.e., difficulty and discrimination) differed across the five states when calibrated separately versus as part of the national sample. Overall, there appeared to be a small to moderate number of items that exhibited a lack of invariance with respect to the item parameter values across states. This is important because the item parameter values define the scale; therefore, any instability in the item parameter values may lead to a lack of score equity. Furthermore, because the average difference between the estimates for a particular comparison was not far from 0 for either parameter, the overall effect on score equity due to a lack of item parameter invariance is likely to be small as was observed in the comparison of the achievement level results described Study 1. This was an important finding in the context of the four studies that were carried out since the item parameter estimates are instrumental in each case.

Study 3 compared the test characteristic curve (TCC) for each respective state to the TCC based on the national sample within 2005. The TCC represents the score scale based on the group's item statistics; thus, any difference in the TCCs, after equating, would represent a lack of score equity. The difference between the TCCs was summarized using the root mean square difference (*RMSD*). A nonparametric bootstrapping procedure was used to construct an empirical distribution for the *RMSD* statistic to assess whether the TCCs indicated a lack of invariance. Although one state (California) produced an *RMSD* statistic that may indicate a lack of strict invariance, the TCCs were mainly comparable between each state and the national sample. Therefore, it appeared that any lack of invariance that may have been present was small resulting in a minimal impact on the score equity across the selected states.

As part of the operational procedure, the item statistics for 2005 are equated to the 2003 scale. Study 4 examined the equating between 2005 and 2003 for each state and the national sample via differences in conversion tables constructed using the TCCs for the respective year. The difference between the 2005 and 2003 TCCs were summarized using the *RMSD* and unsigned average difference (*UAD*) for each state and the national sample. The *RMSD* and *UAD* values were comparable for each state and the national sample

indicating that the equating relationship between 2005 and 2003 were similar across the states.

Overall, the four studies support the conclusion that the scores and achievement level classifications for the selected states appeared to have approximately the same meaning regardless of whether they were treated separately or as part of the national sample.

Introduction

A primary concern for testing programs is whether the inferences drawn from test scores are comparable across various populations. Several psychometric analyses are routinely conducted to assess whether aspects of a testing program may be unfair to certain subgroups of students. In test development, for example, differential item functioning analyses and sensitivity reviews are conducted to evaluate whether specific items may contain construct-irrelevant material that would provide an unfair advantage or disadvantage to specific types of students. In admissions testing, studies of differential predictive validity are often conducted to evaluate the degree to which the predictive utility of a test is consistent across subgroups of students.

More recently, psychometricians have suggested analyzing the equating aspects of a testing program for potential bias. Equating is one of the most complex aspects of a large-scale, standardized testing program and refers to the process of placing scores on two or more test forms onto a common scale (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Through equating, scores on one test form are statistically adjusted for difficulty to match the difficulty of scores on a second test to which it is being equated. Dorans and Holland (2000), Dorans (2002), and Kolen and Brennan (2004) proposed the use of “population invariance” as a criterion for evaluating the results of equating. Using this criterion, tests are considered equitable to the extent that the same equating function is obtained across significant subpopulations (Dorans, 2002). This issue of fairness across subgroups could include groups such as those defined by gender, ethnicity, geographic location, or administration year.

Although “population invariance” was the original term used for evaluating equating equity across subgroups (Lord, 1980), Liu, Cahn, and Dorans (2006) proposed the term *score equity assessment* to describe the evaluation of the invariance of equating functions across subgroups of the population (in the context of male and female students taking the SAT). Given that this term is more likely to be understandable to a non-psychometric audience, we use this term (or its abbreviation—SEA) in the remainder of this report.

In most testing situations, multiple forms of tests are administered and results on the different forms are compared. This is certainly the case in NAEP, in which one of the primary purposes is to monitor the progress of important subgroups (e.g., ethnicity, sex, states) over time. For those comparisons to be valid, the scores from the various forms must be comparable. Equating is a statistical process used to place test scores or item parameters from different test forms onto the same scale so that the appropriate inferences can be made using scores from the different forms (e.g., across time). Specifically, a score conversion is conducted so that scores from two assessments become comparable. In this study, we evaluated the degree to which the placement of scores onto a common scale was consistent (invariant) across subgroups of the NAEP population. The subpopulations we chose for analysis were specific states.

The purpose of this score equity study is to ascertain the extent to which NAEP equating is consistent across selected states. Specifically, we examined the Grade 8 NAEP Math and Reading assessment score scales from 2003–2005 across selected states (described below). This task is complex as there are many places where test score scales are equated in NAEP to make the appropriate comparisons.

In practical terms, the question studied was this: Are particular states at an advantage or disadvantage when national item statistics are used to produce state-level results? If national NAEP item statistics are not, in the main, invariant across states, perhaps because of curriculum differences that are present, then it is possible that some states may be advantaged and others disadvantaged when national NAEP item statistics are used to compile state results. The validity of state to state comparisons of results is highly dependent, therefore, on the important property of item parameter invariance over states, and therefore this property should be carefully examined.

NAEP Equating Procedures

Student progress in NAEP is measured primarily using the main NAEP assessment. Main NAEP is administered in grades 4, 8 and 12 in various subject areas. Although vertical equating across grade levels is no longer done within an administration year, each grade is equated back to the 1992 score scale, via the “mean-sigma” approach, which has been vertically equated. The mean-sigma approach allows one to transform the item characteristic functions, ultimately, from one form of a test to the scale of another. As a result, the scores for the 4th, 8th, and 12th grade assessments administered within a single year end up on the same reporting scale.

Method

Our SEA analyses were conducted in three separate studies. All studies focused on the Grade 8 NAEP Math and Reading tests using data from 2003 and 2005. The specific procedures used in each study are described in separate sections of this report. For each subject area, the equating was replicated following essentially the same procedures used by ETS (i.e., the operational equating procedures). Given the nature of the sampling, the plausible values, and the sample weights, the approach is not straightforward. The equating was done with both the entire sample, as well as for specific subgroups.

States were prioritized in this study as the subgroup of interest for checking item parameter invariance and the subsequent consequences of violating invariance. Five states were selected as the unit of analysis for the Math and Reading assessments. Our goal was to select states that differed with respect to the alignment of their state assessment to NAEP. The five states selected for the Math assessment were Florida (Fla.), Massachusetts (Mass.), California (Calif.), North Carolina (N.C.), and Oklahoma (Okla.). The rationale for selecting these five states was based on correspondences with representatives from state departments of education and researchers, and from two reports that included analyses of the alignment of state and the NAEP Grade 8 Math assessment (Kingsbury, Olson, Cronin, Hauser, and Houser, n.d.; Smithson, 2004). Fla. and Mass. were considered to be “highly aligned” with NAEP; N.C. and Okla. were considered to be “not highly aligned;” and Calif. was considered to be in between the two pairs of states regarding alignment.

The five states selected for analysis of the Reading assessment were California (Calif.), North Carolina (N.C.), New York (N.Y.), Oklahoma (Okla.), and Texas (Texas). We originally planned to use the same five states that were used in the Math assessment, but an independent plan to evaluate the alignment of NAEP and state-specific reading assessments was brought to our attention. Because that plan involved the states of Texas and New York, we substituted those states for Florida and Massachusetts in hope that we could compare any departure of score equity to differences in NAEP-State assessment alignment. Unfortunately, the alignment study did not materialize and therefore the comparison was not possible.

Because the procedure to develop a scale within the NAEP context is complex, we examined score equity through four studies with the intention of understanding invariance from the different steps implemented in the operational procedure. The primary evaluation, and ultimately the most important because of the practical consequences of any lack of parameter invariance, examined score equity via the classification consistency of students into the various NAEP achievement levels based on the different equating functions (i.e., using the national sample or using the state sample). In addition, several additional studies, preparatory to Study 1, were performed for the Math and Reading assessments to examine various aspects that influence scale stability. First, item parameter invariance was assessed across the five states and national sample by examining a - and b -plots and item parameter estimate differences. Second, the test characteristic curves for each state were compared to the national sample for both 2005 and 2003, and the Reading and Math assessments. Last, conversion tables equating scores between 2003 and 2005 were compared for each state and the national sample. Each set of analyses illuminated potential problems that a lack of invariance may have on developing a stable scale across states. To further understand the rationale underlying each analysis, we will first describe ETS’ operational procedure.

This page left intentionally blank

ETS's Operational Procedure

The following describes ETS's operational procedure for calibrating the item parameters, obtaining the plausible values, placing the proficiency estimates onto the NAEP reporting scale, and computing the proportion of examinees classified in each performance category.

Item Parameter Calibration

A proprietary version of PARSCALE, developed by ETS, is used to concurrently calibrate the item parameters for 2005 and 2003 using the entire national sample. The three-parameter logistic model (3PLM) is used for dichotomously scored multiple-choice items, while the two-parameter logistic model (2PLM) and generalized partial credit model (GPCM) are used for open-ended responses that are scored either dichotomously or polytomously, respectively. Omitted item responses for the multiple-choice items are treated as fractionally correct ($1/K$, where K =number of options) due to the potential to correctly answer a multiple-choice item by guessing while the omitted responses for the other types of items (i.e., constructed response) receive the lowest score category (i.e., zero). Not-presented items are treated as missing.

The subscale structure of the NAEP, Grade 8, Mathematics and Reading assessments are preserved by scaling each separately, in a simultaneous calibration. The Math assessment contains five subscales (Numbers and Operations, Geometry, Measurement, Data Analytic, and Algebra and Functions) while the Reading assessment contains three subscales (Reading for Literacy Experience, Reading for Information, and Reading to Perform a Task). A weighting factor is implemented to take into account the sampling procedure used in NAEP. The convergence criterion is specified to be 0.005 with a maximum of 100 expectation maximization (EM) cycles. Priors are placed on all of the item parameters; a normal prior ($N(0,2)$) is used for the item threshold (dichotomous) or location (polytomous) parameters; a lognormal prior is used for the slope parameter; and a beta distribution, with a mean equal to the inverse of the number of response options, is used for the lower-asymptote parameter for the items calibrated using the 3PLM.

The item parameters are calibrated using a three-step process. The first calibration incorporates a fixed normal prior on the proficiency distribution; item parameter estimates from a field test are used as starting values. The second step calibrates the item parameters from the first run using an empirically estimated prior distribution for each year. The final calibration saves the scored item responses to a file suitable for input to MGROUP, the software program used to obtain the plausible values.

Obtaining Plausible Values

ETS implements the following general procedure, using plausible value methodology, to obtain a proficiency estimate for each examinee on each subscale (i.e., five proficiency estimates for Math and three estimates for Reading). The software package, MGROUP, is used to obtain five plausible values for each examinee on each subscale (i.e., 25 total plausible values for Math and 15 for Reading) using the item parameter estimates from the national sample, background variables that are previously converted to principal components, and sampling weights. The five plausible values for each subscale are averaged, resulting in subscale proficiency estimates for each examinee.

The proficiency estimates are obtained for private and public school students, separately. First, MGROUP is run on the entire national sample for the purpose of obtaining proficiency estimates for the private school student sample. Then, to obtain the proficiency estimates for the public school student sample, MGROUP is run for each state separately, using only the respective state public school student sample with the national item parameter estimates (the weights are transformed so that the sum equals the total sample size for the particular state). The mean and standard deviation of the proficiency estimates for the aggregated sample of public

school students from each of the state runs are equated to the mean and standard deviation of the proficiency estimates from the national run for each subscale. The final product is a proficiency estimate for each subscale for each examinee, scaled on the z -metric.

Placing the Proficiency Estimates Onto the NAEP Reporting Scale

The purpose of the following procedure is to place the proficiency estimates from 2005 for each subscale onto the NAEP reporting scale. First, proficiency estimates for examinees from 2003 are obtained using the concurrently calibrated item parameters from 2005 and 2003 (Note: MGROUP is only run for the national sample for 2003; i.e., separate state runs are not performed). The resulting proficiency estimates are equated to the 2003 estimates (based on the previous year) using the means and standard deviations to obtain the linear transformation constants that will be applied to the 2005 proficiency estimates to place them onto the NAEP reporting scale. It is important to note that the sampling weights are used in computing means and standard deviations.

Classifying Examinees

Once the proficiency estimates for each examinee on each subscale have been placed onto the NAEP reporting scale, a composite score is obtained for each examinee by applying subscale weights. Tables 1 and 2 report the subscale weights for the Math and Reading assessments.

Table 1. Weights used to construct the composite score for the Math assessment.

Subscale	Weight
Numbers and Operations	0.20
Geometry	0.15
Measurement	0.20
Data Analytic	0.15
Algebra and Functions	0.30

Table 2. Weights used to construct the composite score for the Reading assessment.

Subscale	Weight
Reading for Literacy Experience	0.40
Reading for Information	0.40
Reading to Perform a Task	0.20

The proportion of examinees classified into the four performance categories (Below Basic, Basic, Proficient, and Advanced) is computed using the respective Math and Reading cut scores. The cut scores for the Math assessment 262, 299, and 333 while the cut scores for the Reading assessment are 243, 281, and 323. The sampling weights are included in calculating the proportion of students within a performance category.

Study 1: Classification Consistency

As mentioned earlier, the primary goal of evaluating score equity within the Grade 8 Math and Reading assessments was to determine if the estimates of the percentages of students classified into the four achievement levels were comparable across states when calibrated separately by state versus as part of the national sample. This would provide direct and informative evidence about whether the score scale was stable across states. To accomplish this goal, we followed ETS's operational procedures, described previously, as closely as possible. Any departure from the operational procedure will be noted in the descriptions of the method and results.

Item Parameter Calibration

Although we were able to calibrate the item parameters using the commercial version of PARSCALE (see Studies 2, 3, and 4), we decided to request ETS's version of PARSCALE, hereafter referred to as NAEP PARSCALE, for compatibility reasons. For example, the format of the item parameter file that is used to obtain the plausible values is compatible with the output file produced by NAEP PARSCALE. Furthermore, NAEP PARSCALE provides certain settings during calibration that the commercial version is lacking (e.g., scoring omits as partially correct for multiple-choice items). We obtained NAEP PARSCALE from ETS (with permission and a temporary license from Scientific Software International) in February 2006. The following analyses were performed using NAEP PARSCALE.

Using NAEP PARSCALE, we concurrently calibrated the item parameters for the 2003 and 2005 Grade 8 math and reading assessment data for the national sample. The 3PLM was used for dichotomously scored multiple-choice items, while the 2PLM and GPCM were used for open-ended responses that were scored either dichotomously or polytomously, respectively. The sampling weights, which account for unequal probability of student selection, were used in the calibration. The subscale structure was preserved and calibrated simultaneously. The default priors in PARSCALE were used. The omits for the multiple-choice items were treated as fractionally correct ($1/K$, where K =number of options) while the omits for the other types of items received the lowest score category. Not-presented items were treated as missing. The starting values, based on the field test, were implemented. The convergence criterion was set to 0.005 with a maximum number of EM cycles of 100.

Three sets of calibrations were performed for the national sample. The first calibration incorporated a fixed normal prior on the proficiency distribution. The second calibrated the item parameter estimates from the first run using an empirically estimated prior distribution for the proficiency distribution for each year. The final run saves the scored item responses to a file suitable for input to DESI, which is the software program developed by ETS to obtain the plausible values. All subscales for the Math and Reading assessments converged using the previous specifications.

Because it is crucial to obtain item parameter estimates comparable to ETS's parameter estimates prior to obtaining the plausible values, the correlation coefficients between the two sets of estimates were evaluated. Table 3 reports the overall correlation coefficients between our item parameter estimates and those provided by ETS rounded to the second place.

Table 3. Correlation coefficients between item parameter estimates we calibrated and ETS’s estimates.

	Item Parameter Estimate		
	<i>a</i>	<i>b</i>	<i>c</i>
Math	1.00	1.00	1.00
Reading	1.00	1.00	1.00

The perfect correlation coefficients (i.e., 1.0) indicated that the item parameter estimates produced by the UMass researchers were linearly related to the ETS estimates.

The same procedure and specifications were followed for calibrating the item parameters for each of the five states, except the weights were rescaled so that the sum equaled the sum for the national weights. The item parameter estimates for each state converged successfully for the Math assessment; however, the Reading assessment failed to converge due to two problematic items. We observed two polytomous items with three categories in which few (and sometimes no) examinees responded to the lowest category in the individual states (in fact, a very small proportion of the national sample responded to the lowest category for both items as well). To solve this problem, we collapsed the two lowest categories (for the national and state samples), creating a dichotomous item and modeled the response using the 2PLM. The state samples converged successfully once the categories were collapsed.

Obtaining Plausible Values

Once the item parameter estimates for the national sample and for each of the five states were obtained (resulting in six sets of item parameter estimates for the Math and Reading assessments), the next step was to obtain two estimates of each examinee’s proficiency level on the math assessment in the five states; one estimate was based on the item parameter estimates calibrated using the entire national sample while the second estimate was based on the item parameter estimates calibrated for the specific state of interest (e.g., Florida). The goal was not to compare these specific estimates for each examinee, but rather to compare the distribution these estimates created for each state.

ETS provided the software package DESI, a user-friendly beta version of ETS’s MGROUP, to obtain the plausible values. DESI uses, as input, the item parameter estimates, raw data, and background variables to obtain the plausible values. The background variables are a crucial element of obtaining the plausible values. Due to the large number of background variables and multicollinearity, the background variables are converted to principal components using principal components analysis (PCA). Unfortunately, this procedure is extremely difficult to replicate. Because all of the background variables are treated as categorical, contrast codes are used in the PCA. There are literally thousands of contrast codes representing the main effects and interactions between the variables. Considering the additional difficulties in estimation (e.g., software, convergence), we opted to request, and received, the principal components from ETS so that we could concentrate on computing the plausible values as accurately as possible.

We attempted to use the same specifications and procedure in estimating the plausible values implemented by ETS. For instance, even though the plausible values for the national sample use the same item parameter estimates, ETS obtains the values for each state individually (for the public school sample). Therefore, we obtained the plausible values for the national sample for each state separately; however, the item parameters calibrated using the entire national sample were used for each state. A second set of plausible values were obtained for each state using the respective state item parameter estimates.

For each run, five multiple imputations were performed for each examinee on each subscale. The mean of the five plausible values were used as an estimate of the examinee’s proficiency for the respective subscale. For California and Florida, the maximum number of EM cycles was set to 2000, with a convergence criterion of 0.001 for the standardized coefficients

and 0.01 for the log-likelihood function. For the remaining three states (Massachusetts, North Carolina, and Oklahoma) however, the number of EM cycles was increased to 2,500 and the imputation had to be performed twice due a lack of convergence on the first attempt; the estimates for the regression and residual covariance coefficients (GFILE) from the first run were used as starting values for the second DESI run. Running DESI twice is a strategy that ETS uses when the first attempt fails to converge.

The resulting proficiency estimates for each subscale were placed onto the 2005 NAEP reporting scale via the mean-sigma method for each state separately. A composite score was computed for each examinee using the weights reported in Tables 1 and 2.

Classification Consistency

The percentage of examinees in each of the four NAEP performance categories (Below Basic, Basic, Proficient, and Advanced) were determined for each set of plausible values using the 2005 Math cut scores for Grade 8. Table 4 reports the percentage of examinees classified in each category based on the plausible values using the national item parameter estimates as well as the item parameters calibrated within each state for the Math assessment.

Table 4. Achievement level results for Math: National versus State item parameter estimates.

State	National				State			
	Below Basic	Basic	Prof.	Adv.	Below Basic	Basic	Prof.	Adv.
Florida	34.7	39.5	20.9	4.9	34.9	39.7	20.2	5.2
Massachusetts	19.4	35.9	32.3	12.4	19.4	36.7	31.8	12.1
California	43.0	34.3	17.3	5.3	42.8	34.7	17.3	5.3
North Carolina	27.0	40.5	25.1	7.4	27.1	40.8	24.6	7.5
Oklahoma	34.7	45.4	17.7	2.2	34.9	45.0	18.1	2.0
Mean	31.76	39.12	22.66	6.44	31.82	39.38	22.4	6.42

The percentage of examinees within each category were comparable across the selected states when treated as part of the national sample or individually. The differences between the percentages for the respective categories were very small. The largest difference for Below Basic was -0.2 (Oklahoma and Florida) and 0.2 (California); for the Basic category, the largest difference was -0.8 (Massachusetts); for the Proficient category the largest differences were 0.7 (Oklahoma) while for the Advanced category the largest differences were -0.3 (Florida) and 0.3 (Massachusetts). Therefore, it appeared that any lack of score equity was minimal when the states were treated separately versus as part of the national sample.

The percentage of examinees in each of the four NAEP performance categories (Below Basic, Basic, Proficient, and Advanced) were determined for each set of plausible values using the 2005, Reading cut scores for Grade 8. Table 5 reports the percentage of examinees classified in each category based on the plausible values using the national item parameter estimates as well as the item parameters calibrated within each state for the Reading assessment.

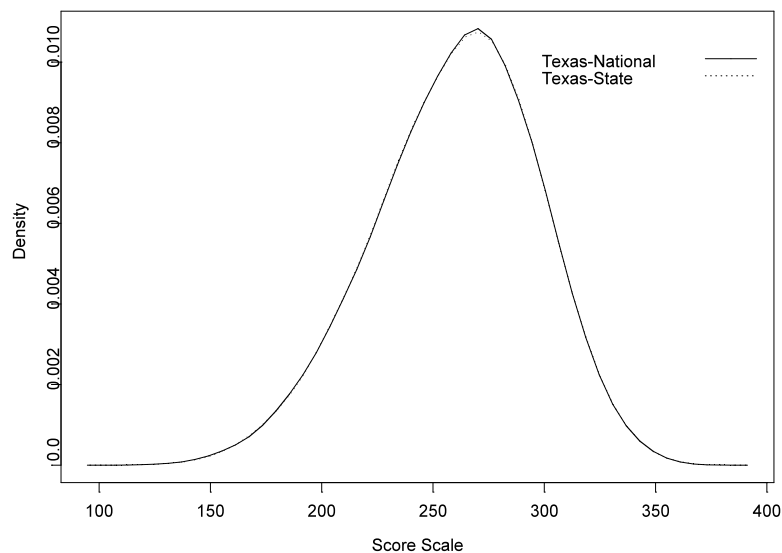
Table 5. Achievement level results for Reading: National versus State item parameter estimates.

State	National				State			
	Below Basic	Basic	Prof.	Adv.	Below Basic	Basic	Prof.	Adv.
California	40.2	37.2	21.0	1.5	40.2	37.3	21.1	1.4
New York	26.7	41.7	29.2	2.4	26.5	41.8	29.3	2.4
North Carolina	26.5	42.4	29.2	1.9	26.4	42.4	29.2	1.9
Oklahoma	25.8	45.2	27.4	1.6	26.1	45.1	27.1	1.7
Texas	27.9	40.0	29.2	2.9	27.9	39.6	29.4	3.1
Mean	29.4	41.3	27.2	2.1	29.4	41.2	27.2	2.1

The percentage of examinees within each category were comparable across the selected states when treated as part of the national sample or individually. The differences between the percentages for the respective categories were very small. The largest difference between percentages for the Below Basic category was 0.3 (Oklahoma); for the Basic category, the largest difference was 0.4 (Texas); while for the Proficient and Advanced categories, the largest differences were 0.3 (Oklahoma) and 0.2 (Texas), respectively. Therefore, it appeared that any lack of score equity was minimal when the states were treated separately versus as part of the national sample.

To verify the consistent classification rates, we compared the proficiency distribution for each state. For instance, Figure 1 illustrates the smoothed proficiency distribution for California for the Reading assessment.

Figure 1. A comparison of the smoothed proficiency distribution for Texas based on the state and national item parameter estimates for the Reading assessment.



The solid line represents the proficiency distribution for California when the national item parameter estimates were used while the dashed line is the proficiency distribution for California when the state item parameter estimates were used. The two distributions are nearly identical,

supporting the conclusion that any effect due to a lack of invariance was minimal. The proficiency distributions for the other states were also similar for the Math and Reading assessments.

Because the percentages were comparable for each state determined separately versus as part of the national sample, it appeared that any lack of equity in the NAEP equating procedures for the 2005, Grade 8 Math and Reading assessments with respect to the five studied states was minimal. Furthermore, the alignment of a state's assessment program to NAEP did not appear to have a meaningful impact on a lack of score equity with respect to the achievement level results for these five states. For example, the Math results for Florida, which was highly aligned with NAEP, was comparable to North Carolina, which was not highly aligned with the NAEP Math assessment.

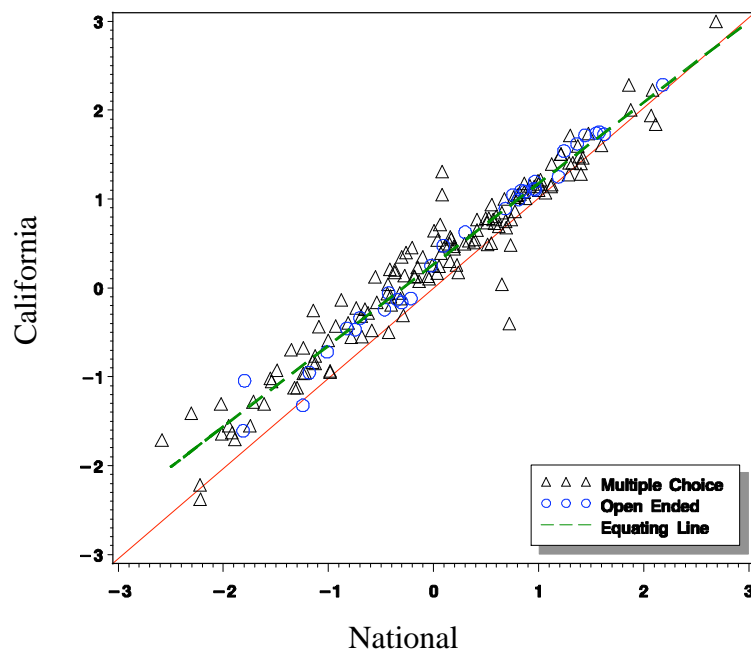
This page left intentionally blank

Study 2: Item Parameter Invariance Across Selected States

Because a lack of invariance in the item parameter values may have a deleterious effect on scale stability and state results, we examined whether items were functioning differently in the five selected states in the Math and Reading assessment, especially compared to the national sample. Item parameters were calibrated for the national sample and for the five respective states separately using the commercial version of PARSCALE. The sampling weights were used in the calibration; the subscale structure was preserved and simultaneously calibrated; the default priors in PARSCALE were used; and the starting values, based on the field test, were implemented. The 3PLM was used for the dichotomously scored multiple-choice items, while the 2PLM and GPCM were used for open-ended responses that were scored either dichotomously or polytomously, respectively. Six sets of item parameter estimates were produced for each of the Grade 8 Math and Reading assessments for 2005.

The item parameter estimates for the discrimination (a) and difficulty (b) parameters for each of the five states, as well as the national sample, were compared, resulting in fifteen pair-wise comparisons for each parameter (30 total pair-wise comparisons) for each test. The mean-sigma procedure was used to create an equating line between the two groups. Figure 2 displays the scatter-plot for the b -parameter estimates between California and the national sample (NATL) for the Math assessment. Appendix A and B reports the scatter-plots for each pair-wise comparison for the b - and a -parameters for the Math and Reading assessment, respectively.

Figure 2. b -parameter estimates for California and the national sample.



The dashed line represents the equating line determined by the mean-sigma method. The b -parameter estimates that fall far from the equating line represent items that appear to function differently in the two groups. The triangular and circular shapes represent the coordinates for b -parameter estimates for multiple-choice and open-ended items, respectively.

An examination of the a - and b -plots showed that a few of the items may be functioning differently across states, especially when compared to the national sample. Variation in the item parameter estimates across pairs of groups appears substantially greater for the a -parameter estimates than for the b -parameter estimates but this is common because the a -parameter estimates, relatively speaking, tend to have larger standard errors. In addition to the a - and b -

plots we used a statistical criterion that incorporated the standard error to identify potential items that may be functioning differently between the two samples. We flagged item parameter estimates that differed by more than three standard errors as potentially behaving differently in the two groups of interest. The standard errors of the difference between item parameter estimates were computed as follows:

$$S_{b_{i,gp1}-b_{i,gp2}} = \sqrt{\text{VAR}(b_{i,gp1}) + \text{VAR}(b_{i,gp2})} \text{ and } S_{a_{i,gp1}-a_{i,gp2}} = \sqrt{\text{VAR}(a_{i,gp1}) + \text{VAR}(a_{i,gp2})} \quad (1)$$

Gp1 and gp2 refer to the two groups being compared; i.e., either two states or a state with the national sample. The variances (VAR) for the respective estimate were reported by PARSCALE.

It is important to note that a potential problem with using the formulas shown in (1) to compute the standard errors is that because the states are a subgroup of the national sample the standard errors may be underestimated. Therefore, the proportion of items being flagged may be slightly inflated.

Although the state pair-wise comparisons are interesting (e.g., Florida versus Massachusetts), the most relevant and useful information pertains to the comparisons between the states and national sample. Tables 6 and 7 report the proportion of detected items for the 2005 Grade 8 Math and Reading assessments for the *b*- and *a*-parameters. The proportion of common items between 2003 and 2005 that were flagged are reported in parentheses.

Table 6. Proportion of 2005 Grade 8 Math items exhibiting item parameter estimate differences beyond three standard errors.

State	NATL		
	<i>b</i> -parameter	<i>a</i> -parameter	<i>b</i> - or <i>a</i> -parameter
Florida	.07 (.07)	.04 (.05)	.11 (.11)
Massachusetts	.04 (.04)	.04 (.05)	.07 (.07)
California	.18 (.17)	.08 (.09)	.23 (.23)
North Carolina	.06 (.07)	.04 (.05)	.08 (.10)
Oklahoma	.04 (.05)	.02 (.02)	.06 (.07)

Proportions of flagged equating items between 2005 and 2003 are reported in parentheses.

Table 7. Proportion of 2005 Grade 8 Reading items exhibiting item parameter estimate differences beyond three standard errors.

State	National		
	<i>b</i> -parameter	<i>a</i> -parameter	<i>b</i> - or <i>a</i> -parameter
California	.08 (.10)	.06 (.08)	.11 (.14)
New York	.06 (.09)	.02 (.02)	.08 (.11)
North Carolina	.07 (.05)	.06 (.04)	.11 (.09)
Oklahoma	.06 (.06)	.05 (.06)	.08 (.10)
Texas	.11 (.10)	.04 (.05)	.14 (.15)

Proportions of flagged equating items between 2005 and 2003 are reported in parentheses.

There appears to be a small but nontrivial number of items that are performing differently when calibrated with the entire national sample compared to the states separately. For the Math assessment, the largest proportion of items was observed for California (0.23, overall) while the smallest proportion was observed for Oklahoma (0.06, overall). For the Reading assessment, the largest proportion was observed for Texas (0.14, overall) while the smallest proportion was observed for New York and Oklahoma, (0.08, overall). While sample size certainly influences the number of items detected (e.g., California was the largest sample while Oklahoma was the smallest), the large proportion of items in which the *a*- or *b*-parameter estimates were greater than three standard errors across the five states could affect score equity.

The proportion of anchor items that were flagged between 2003 and 2005 was comparable to the overall number of items. Interestingly, there was little overlap between items detected for the *a*- and *b*-parameter; i.e., very few items were detected as differing in both the *a*- and *b*-parameter. We suspect this finding is simply due to the way item parameter estimation is carried out. If one parameter estimate is far off the mark, the estimation algorithms compensate by trying to improve the fit of the model to the data by adjusting the second model parameter. Because the IRT models generally provide good fits to the student item response data, it is not likely that both model parameter estimates could be far out of line with item parameter estimates obtained from another examinee sample unless there is a major lack of item parameter invariance across the two samples.

The proportion of items detected is not the only factor that will influence score equity; the magnitude and direction of the differences between the item parameter estimates will also play an important role in affecting the scale. Therefore, the mean difference between the item parameter estimates flagged was computed to determine how much the estimates differed overall. Tables 8 and 9 report the means and standard deviations of the differences between the *a*- and *b*-parameter for flagged items for the Math and Reading assessments.

Table 8. Mean difference between a- and b-parameter estimates identified to function differentially for Grade 8 Math, 2005.

State	National	
	<i>b</i> -parameter	<i>a</i> -parameter
Florida	-0.30 (0.50)	-0.00 (0.20)
Massachusetts	-0.22 (1.09)	-0.09 (0.16)
California	-0.18 (0.54)	-0.08 (0.25)
North Carolina	-0.24 (0.83)	-0.00 (0.21)
Oklahoma	-0.70 (1.25)	-0.32 (0.19)

Standard deviations reported in parentheses.

Table 9. Mean difference between a- and b-parameter estimates identified to function differentially for Grade 8 Reading, 2005.

State	National	
	<i>b</i> -parameter	<i>a</i> -parameter
California	-0.13 (0.19)	-0.08 (0.29)
New York	0.20 (0.44)	-0.20 (0.16)
North Carolina	0.16 (0.46)	-0.01 (0.31)
Oklahoma	-0.07 (0.78)	-0.23 (0.22)
Texas	0.10 (0.40)	0.00 (0.30)

Standard deviations reported in parentheses.

The average difference for each comparison was relatively small, even though in some comparisons a large proportion of items were flagged. For example, the average difference between the *b*-parameter estimates for the National and California comparison was only -0.13, yet 18 percent of the items were flagged as being three standard errors apart. It appears that, even though there may be a non-ignorable proportion of items functioning differently, the differences between the estimates are minimal when magnitude and direction are considered. As a result, the impact due to a lack of item parameter invariance may be minimal on score equity.

The item parameter estimates for 2003 were also examined using the same procedures described previously as a cross-validation. Tables 10, 11, 12, and 13 report the proportion of items flagged and mean differences for items that differed by more than three standard errors, respectively.

Table 10. Proportion of 2003 Grade 8 Math items exhibiting item parameter estimate differences beyond three standard errors.

State	National		
	<i>b</i> -parameter	<i>a</i> -parameter	<i>b</i> - or <i>a</i> -parameter
Florida	.09 (.11)	.07 (.11)	.15 (.16)
Massachusetts	.05 (.07)	.07 (.07)	.10 (.12)
California	.30 (.30)	.18 (.29)	.38 (.38)
North Carolina	.07 (.07)	.05 (.07)	.11 (.12)
Oklahoma	.03 (.02)	.02 (.02)	.04 (.05)

Proportions of flagged equating items between 2005 and 2003 are reported in parentheses.

Table 11. Proportion of 2003 Grade 8 Reading items exhibiting item parameter estimate differences beyond three standard errors.

State	National		
	<i>b</i> -parameter	<i>a</i> -parameter	<i>b</i> - or <i>a</i> -parameter
California	.14 (.15)	.10 (.09)	.20 (.19)
New York	.08 (.09)	.02 (.02)	.10 (.10)
North Carolina	.11 (.10)	.04 (.00)	.14 (.10)
Oklahoma	.09 (.09)	.06 (.05)	.13 (.13)
Texas	.13 (.08)	.12 (.08)	.20 (.14)

Proportions of flagged equating items between 2005 and 2003 are reported in parentheses.

Table 12. Mean difference between *a*- and *b*-parameter estimates identified to function differentially for Grade 8 Math, 2003.

State	National	
	<i>b</i> -parameter	<i>a</i> -parameter
Florida	0.02 (0.33)	-0.20 (0.23)
Massachusetts	-0.31 (0.81)	-0.11 (0.33)
California	-0.08 (0.76)	0.27 (0.74)
North Carolina	0.04 (0.42)	-0.16 (0.30)
Oklahoma	0.17 (2.48)	-0.13 (0.23)

Standard deviations reported in parentheses.

Table 13. Mean difference between *a*- and *b*-parameter estimates identified to function differentially for Grade 8 Reading, 2003.

State	National	
	<i>b</i> -parameter	<i>a</i> -parameter
California	0.04 (0.57)	0.16 (0.18)
New York	0.01 (0.67)	0.20 (0.10)
North Carolina	0.17 (0.46)	0.00 (0.30)
Oklahoma	0.43 (0.90)	-0.05 (0.37)
Texas	0.44 (0.74)	0.12 (0.31)

Standard deviations reported in parentheses.

The results for 2003 were comparable to 2005. A moderate to large proportion of items were flagged as behaving differently between the national sample and the five states. The mean difference between the item parameter estimates for flagged items was again relatively small.

Overall, there appear to be a small number of items that exhibited a lack of invariance with respect to the item parameter values across states. This is important because the item parameter values define the scale; therefore, any instability in the item parameter values may lead to a lack of score equity. However, because the average difference between the estimates for a particular comparison was not far from 0 for either parameter, the overall effect on score equity due to a lack of item parameter invariance is likely to be small as was observed in the comparison of the achievement level results described Study 1. This was an important finding in the context of the four studies that were carried out.

This page left intentionally blank

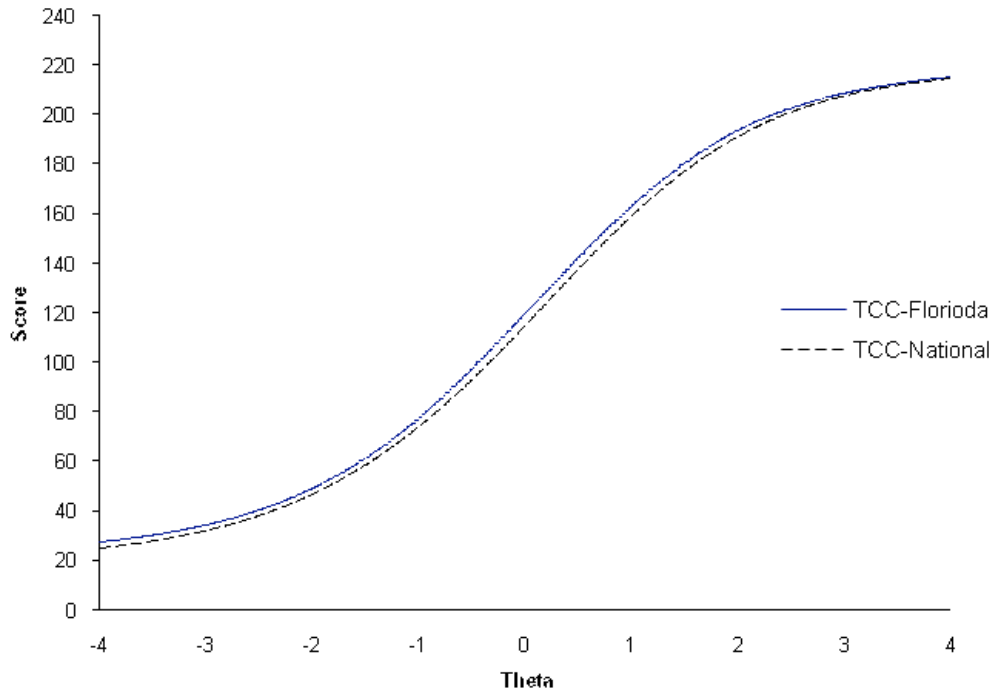
Study 3: Comparison of Test Characteristic Curves Within a Year

Study 2 compared the item parameter estimates for each state to each other and to the national sample. In study 3, we compared each state's test characteristic curve (TCC) for the Math and Reading assessments with the TCC from the national sample within 2005. This allowed us to examine whether the Reading and Math score report scales for each state were comparable whether derived from the national sample or the state sample. Lack of invariance would be present were the TCCs for each subject and grade to look different after adjustments are made for the nonequivalence of the state and national proficiency distributions.

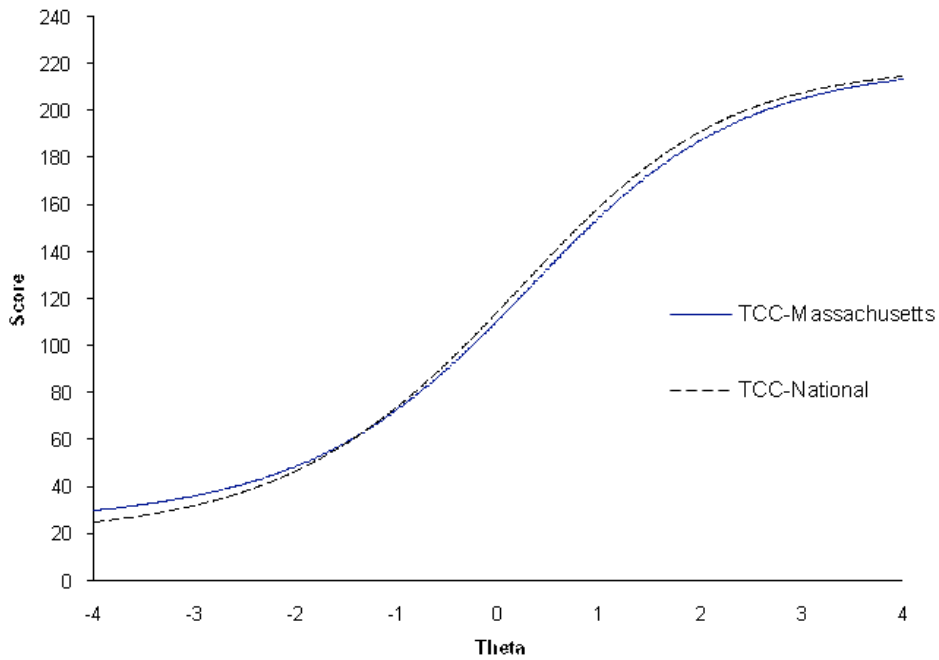
The item parameters were calibrated for the national sample and for the five respective states separately for 2005 using the commercial version of PARSCALE, resulting in six sets of item parameter estimates per assessment. The sampling weights were used in the calibration; the subscale structure was preserved and simultaneously calibrated; the default priors in PARSCALE were used; and the starting values, based on the field test, were implemented. To allow for a meaningful comparison of the TCCs between the states and the national sample, the item parameter estimates were placed onto the scale of the NAEP item parameter estimates, provided by ETS, using the mean-sigma method. The TCCs for the national sample and for each state were computed using the transformed item parameter estimates. The TCC for each state was compared to the TCC based on the national sample. The extent to which invariance holds over states can be represented by the difference between the TCCs; i.e., large differences indicate a lack of invariance or scale stability. Figure 3 (a) to (e) shows the comparison of TCCs for each state to the national sample for the Math assessment. Figure 4 (a) to (e) shows the comparison of TCCs for each state to the national sample for the Reading assessment.

Figure 3. Test characteristic curves for each state compared to the national sample for the Math assessment.

(a) Florida versus National



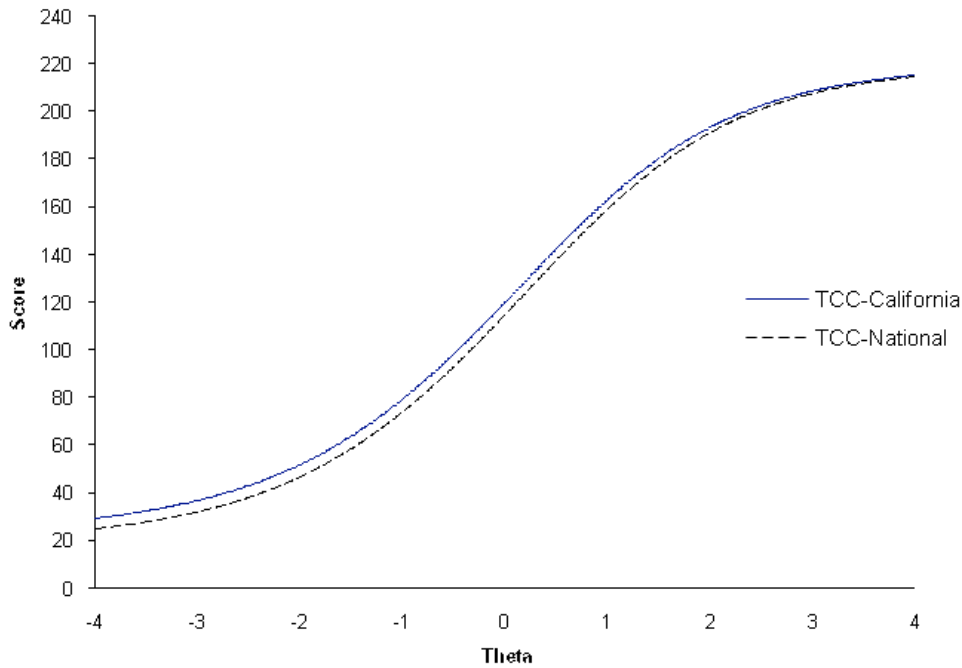
(b) Massachusetts versus National



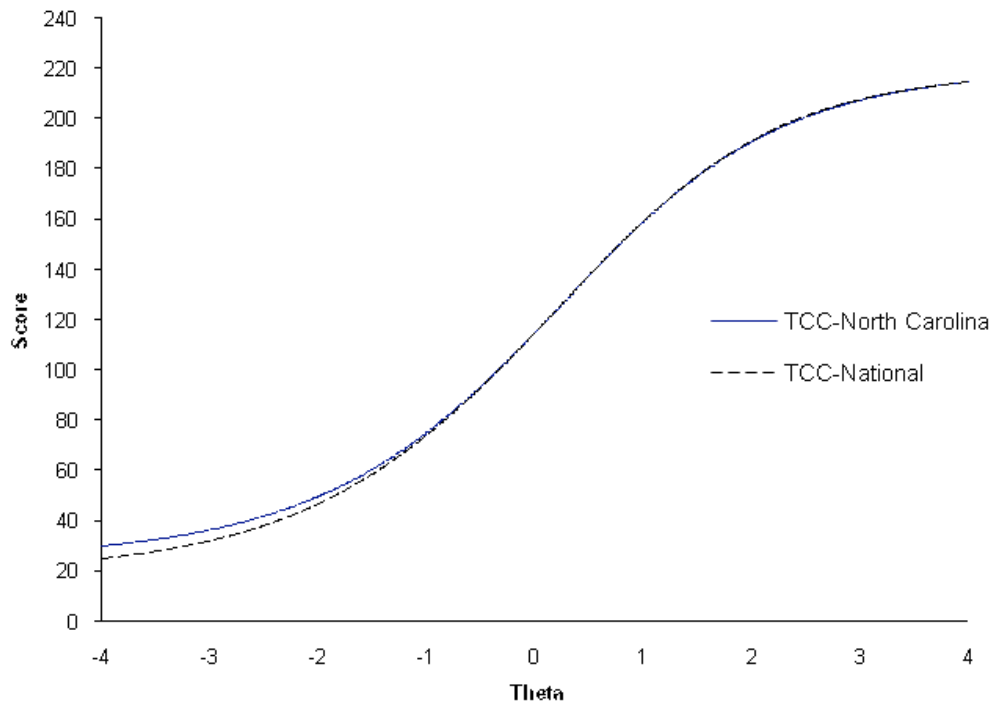
Continues next page

Figure 3. Test characteristic curves for each state compared to the national sample for the Math assessment (Continued)

(c) California versus National



(d) North Carolina versus National



Continues next page

Figure 3. Test characteristic curves for each state compared to the national sample for the Math assessment (Continued)

(e) Oklahoma versus National

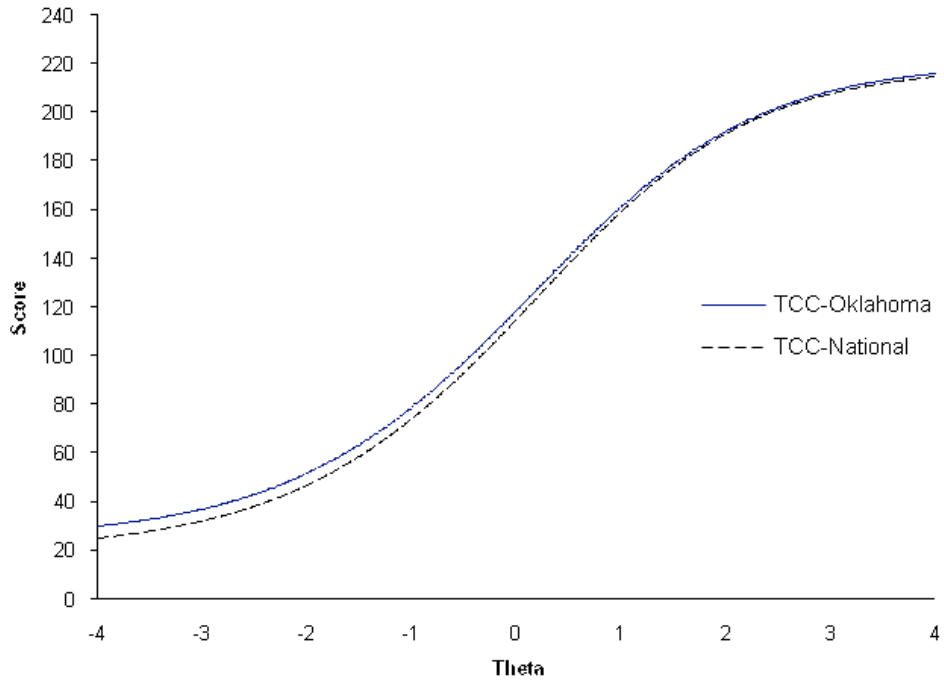
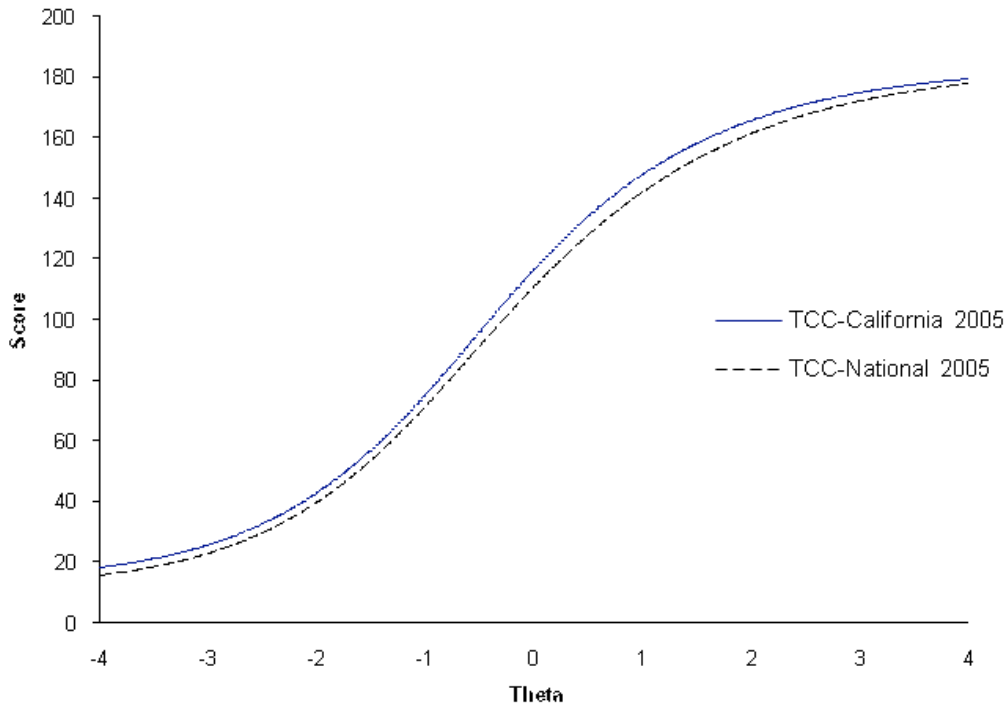
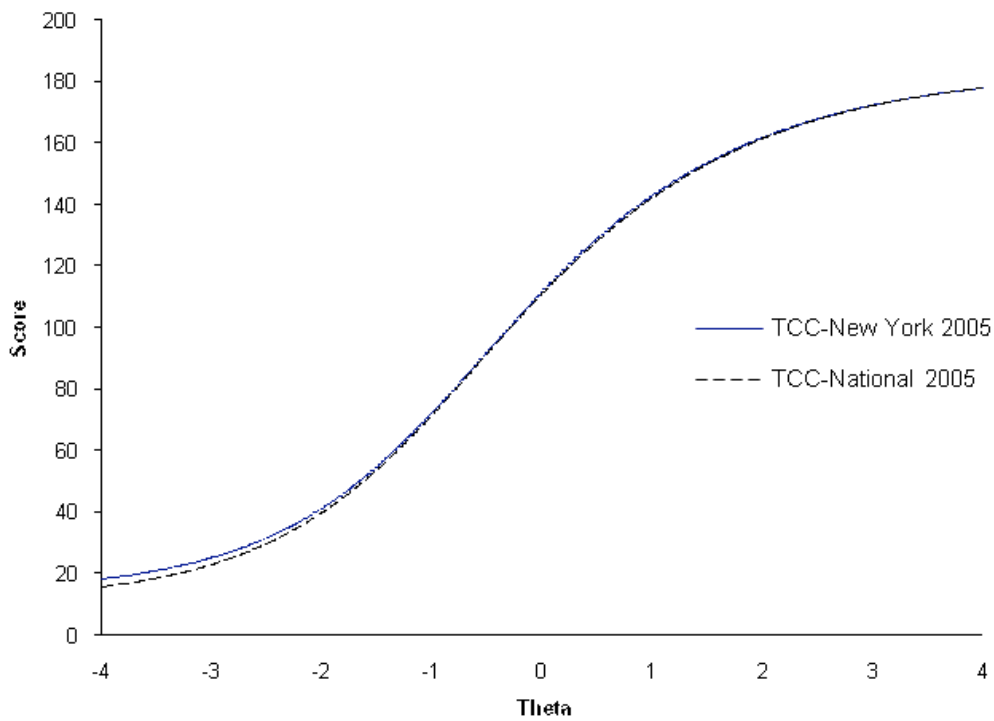


Figure 4. Test characteristic curves for each 2005 state sample compared to the 2005 national sample for the Reading assessment.

(a) California versus National



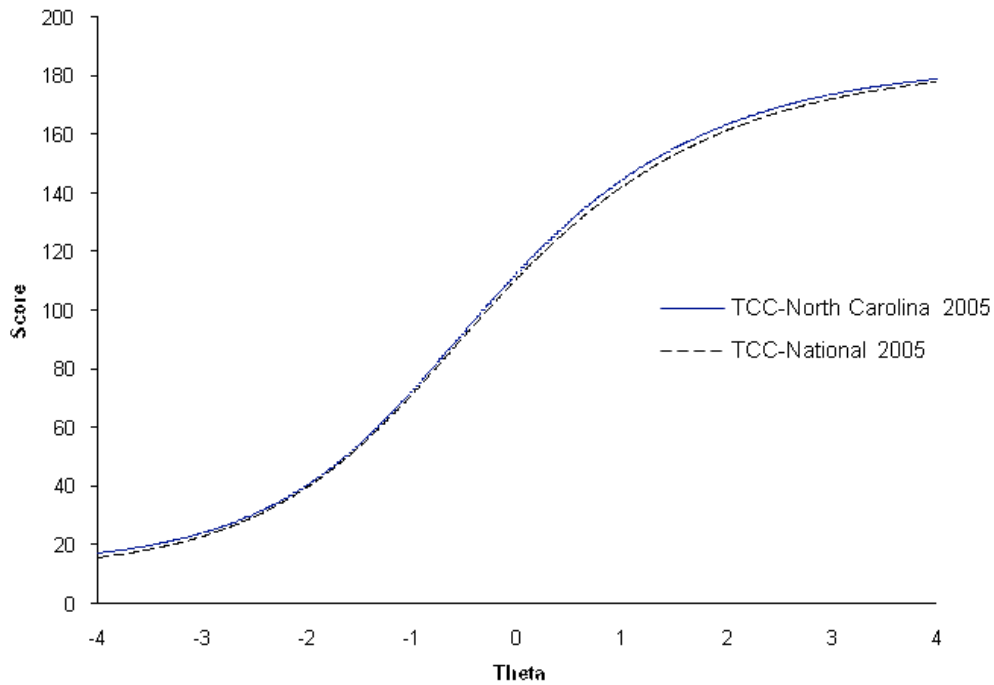
(b) New York versus National



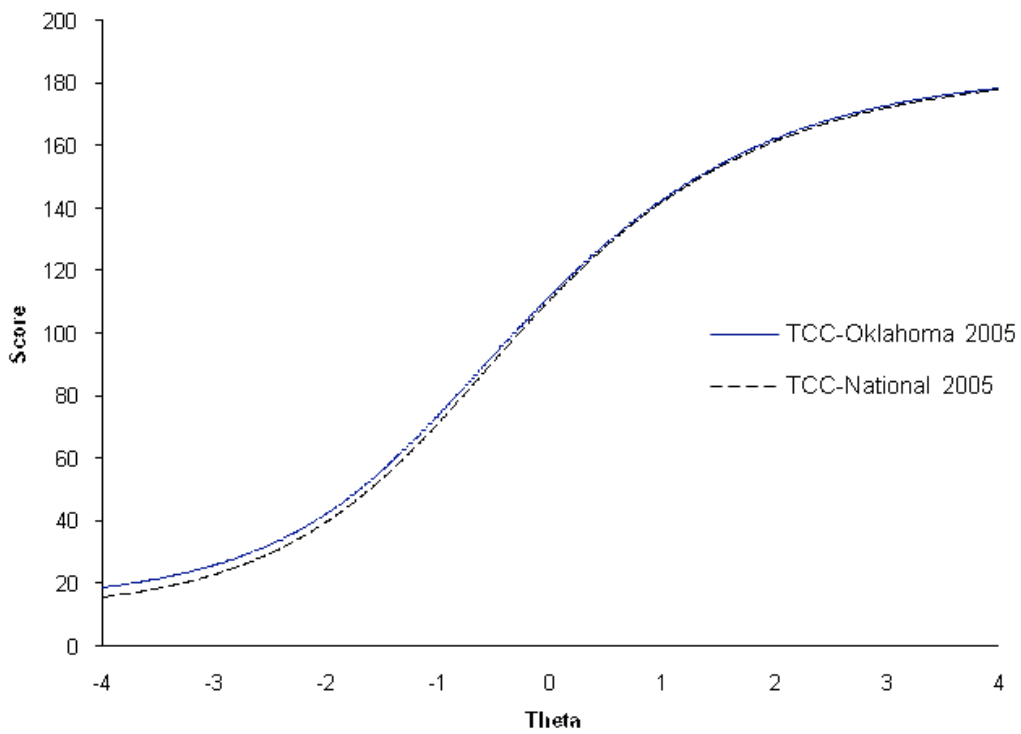
Continues next page

Figure 4. Test characteristic curves for each 2005 state sample compared to the 2005 national sample for the Reading assessment (Continued)

(c) North Carolina versus National



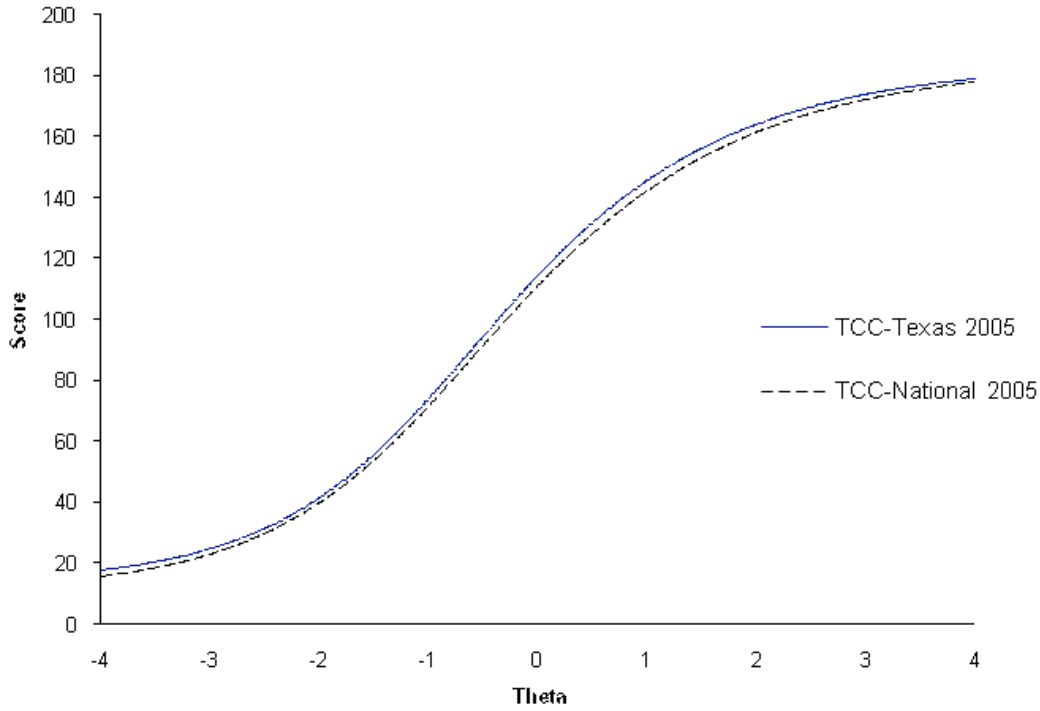
(d) Oklahoma versus National



Continues next page

Figure 4. Test characteristic curves for each 2005 state sample compared to the 2005 national sample for the Reading assessment (Continued)

(e) Texas versus National



The differences between the TCCs for each state comparison were summarized using a root mean squared difference (*RMSD*). The discrepancy between the TCCs for 31 θ -values, ranging from -3.0 to 3.0 in increments of 0.2, were used to compute the following *RMSD* for state *s*:

$$RMSD_s = \sqrt{\frac{\sum_{j=1}^{31} (TCC_{state}(\theta) - TCC_{NATL}(\theta))^2}{31}}. \quad (2)$$

The *RMSD* for each comparison is reported in Tables 14 and 15.

Table 14. *RMSD* for each comparison between the TCC for each respective state and the national sample for the 2005 Math assessment.

Comparison	<i>RMSD</i>
Florida – National	3.38
Massachusetts – National	3.04
California – National	4.49
North Carolina – National	1.69
Oklahoma – National	3.70

Table 15. *RMSD* for each comparison between the TCC for each respective state and the national sample for the 2005 Reading assessment.

Comparison	<i>RMSD</i>
California – National	4.31
New York – National	0.88
North Carolina – National	1.73
Oklahoma – National	1.65
Texas – National	2.49

Unfortunately, it is not possible to determine if the *RMSD* values are large indicating a lack of scale stability without a baseline for interpreting the *RMSD* statistic, though they appear small from a review of the corresponding TCCs. To address the lack of a baseline, a nonparametric bootstrapping procedure was used to obtain an empirical distribution for *RMSD* under the condition that invariance holds (i.e., the scale is stable across the state and national sample). If the observed *RMSD* values (shown in Tables 14 and 15) are large relative to the empirical distribution, then there is evidence that the national and state scales are not invariant.

Empirical Distribution for *RMSD*

The purpose of the nonparametric bootstrapping procedure was to obtain an empirical distribution for the *RMSD* statistic under the condition that the scale is stable across states. A homogeneous subpopulation for the NAEP 2005 data was selected for the Math and Reading assessments to obtain a condition in which score equity would likely hold. This reference subpopulation was defined by the largest ethnic group (white), excluding students with limited English proficiency or individualized educational plans. This process resulted in a pool of 78,500 examinees out of 168,141 (49.7 percent) for the Math assessment and 75,406 out of 168,782 (44.7 percent) for the Reading assessment, hereafter referred to as the full-group in both assessments. Smaller samples were drawn from the full-group, without replacement, that represented the states. Since the five states differed in size, and sample size may have an impact on the *RMSD* statistic, three smaller sample sizes were examined for the Math assessment (2,800, 4,000, and 11,000) while four sample sizes were used in the Reading assessment (2,560, 4,000, 7,520, and 10,000). One thousand replications were performed for each of the smaller sample sizes. A constraint was placed on the sampling to require the sparseness of the data matrix for the smaller groups to mimic that of the full-group.

For each replication, the item parameter estimates were calibrated for the full-group and smaller group using the commercial version of PARSCALE. The sampling weights were used in the calibration; the subscale structure was preserved; the default priors in PARSCALE were used; and the starting values, based on the field test, were implemented. The convergence criterion was also relaxed to 0.01 to improve the proportion of replications that converge. Using these criteria, 98 percent of the replications converged for the Math assessment while 99 percent of the replications converged for the Reading assessment. Of those datasets that did not converge, it was typically one of the subscale tests that failed. The following results are based on the converged datasets. The calibrated datasets were equated back to the NAEP item parameter estimates, provided by ETS, using the mean-sigma method.

Once the item parameter estimates for the full-group and smaller-group were placed onto a common scale, the TCCs were constructed for each group. The *RMSD* was computed for each replication to obtain the empirical distribution for each sample size condition. Appendices C and D show the empirical distribution across the three sample sizes for the smaller group.

The *RMSD* for each state was compared to the appropriate empirical distribution based on the state's sample size. The percentile of the observed *RMSD* and sample size of each state is reported in Tables 16 and 17.

Table 16. Percentile of the observed RMSD for each state relative to the appropriate empirical distribution for the Math assessment.

Comparison	Percentile	Sample Size
Florida – National	89.3	4,258
Massachusetts – National	85.2	3,581
California – National	99.9	10,638
North Carolina – National	33.1	4,085
Oklahoma – National	88.2	2,608

Table 17. Percentile of the observed RMSD for each state relative to the appropriate empirical distribution for the Reading assessment.

Comparison	Percentile	Sample Size
California – National	91.2	9,754
New York – National	1.3	4,162
North Carolina – National	12.2	3,907
Oklahoma – National	19.4	2,492
Texas – National	31.6	7,759

The only state that produced an *RMSD* that may be considered significant (i.e., above the 95th percentile) was California (percentile=99.9) on the Math assessment. A review of the corresponding TCCs between California and the National (see Figure 3-c) does appear to suggest a small difference of the order of several score points.

Because the TCCs were mainly comparable between each state and the national sample, it appears that any lack of invariance that may be present was small resulting in a small impact on the score equity across states. These results are consistent with Study 2 that found the item parameter invariance between the states and national sample was small. Still, there is what appears to be a small but significant difference in the state and national TCCs for math and possibly reading too in California.

This page left intentionally blank

Study 4: Comparison of TCCs Between 2005 and 2003

Part of the NAEP operational procedure is to equate 2005 to 2003 to place the values onto the NAEP reporting scale and examine trends. Therefore, it is informative to examine whether the transformation at the item parameter level is comparable across the states and for the national sample. Conversion tables between 2005 and 2003 were constructed and compared for each state and the national sample to assess whether the scale shifted differentially between the states, indicating a lack of score equity.

The item parameters for the Math and Reading assessments were calibrated for the national sample and for the five respective states separately for 2005 and 2003 using the commercial version of PARSCALE, resulting in six pairs of item parameter estimates. The sampling weights were used in the calibrations; the subscale structure was preserved and simultaneously calibrated; the default priors in PARSCALE were used; and the starting values, based on the field test, were implemented in parameter estimation. To allow for a meaningful comparison of the conversion tables across states and the national sample, the 2003 item parameter estimates from each of the five states were placed onto the national scale using the mean-sigma method; hereafter, the transformed 2003 item parameter estimates will be referred to as 03*. Next, for each state and the national sample, the 2005 item parameter estimates were placed onto the 03* scale for the respective group (i.e., state and national sample) via the mean-sigma method using the common items between 2005 and 2003. For each state and the national sample, a test characteristic curve (TCC) was constructed for 2005 and 2003 based on all the respective items. The TCCs, converted to proportions to control for a difference in score points between 2005 and 2003, for the Math and Reading assessments are shown in Appendices C and D, respectively.

The discrepancy between the 2005 and 2003 TCCs, reported as proportions, for 31 θ - values, ranging from -3.0 to 3.0 in increments of 0.2, were used to compute the following *RMSD* and unsigned average difference (*UAD*):

$$RMSD_s = \sqrt{\frac{\sum_{j=1}^{31} (\text{Prop}_{05}(\theta) - \text{Prop}_{03}(\theta))^2}{31}} \quad (3)$$

and

$$UAD_s = \frac{\sum_{j=1}^{31} (\text{Prop}_{05}(\theta) - \text{Prop}_{03}(\theta))}{31} .$$

Table 18 reports the *RMSD* and *UAD* values for the national sample and for each state on the Math and Reading assessments.

Table 18. RMSD and UAD values summarizing the difference between raw scores in the conversion table for the national sample and for each state.

(a) Math assessment.

Group	Math	
	<i>RMSD</i>	<i>UAD</i>
National	0.016	0.015
Florida	0.018	0.000
Massachusetts	0.019	0.000
California	0.026	0.001
North Carolina	0.013	0.000
Oklahoma	0.014	0.000
Mean	0.018	0.003
Std Dev.	0.005	0.006

(b) Reading assessment

Group	Reading	
	<i>RMSD</i>	<i>UAD</i>
National	0.035	0.027
California	0.015	-0.014
New York	0.021	-0.019
North Carolina	0.028	-0.022
Oklahoma	0.021	-0.019
Texas	0.023	-0.021
Mean	0.024	-0.011
Std Dev.	0.007	0.019

The *RMSD* and *UAD* values were comparable for each state as well as the national sample indicating that the differences between the TCCs were comparable. This appears to be consistent with the results from the *a*- and *b*-plots (Study 2); i.e., because the mean differences in the *a*- and *b*-parameter estimates were small, we would not expect a large effect due to a lack of item parameter invariance.

Conclusion

The current evaluation examined the score equity of the 2005 Grade 8 NAEP Math and Reading assessment across selected states that varied regarding the alignment of the state assessment with NAEP. For the Math assessment, Florida and Massachusetts were judged to be “highly aligned” while North Carolina and Oklahoma were “not highly aligned” and California was considered to be somewhere in the middle. Score equity is an important concept that examines whether a test measures the same construct in a similar manner across distinct subgroups. In this evaluation, score equity was assessed by examining the scale for the states treated separately versus as part of the entire national sample.

Study 1 examined the consistency of the achievement level results across the five selected states when they were treated individually versus as part of the national sample. The achievement level proportions were comparable regardless of whether the states were treated individually or as part of the national sample. Furthermore, the distribution of plausible values for a particularly state were very similar when the item parameter estimates based on the state versus national sample were used.

In Study 2, we examined whether the item parameter values (i.e., difficulty and discrimination) differed across the five states when calibrated separately versus as part of the national sample. Although a moderate proportion of items were identified as having different parameter values when each state was compared to the national sample, the overall effect was expected to be minimal because the mean difference of the flagged estimates was close to 0 for both the *a*- and *b*-parameter estimates.

Study 3 compared the TCCs between each state and the national sample within 2005. An *RMSD* statistic was computed for each of the five comparisons to summarize the difference between the TCCs. In addition, a nonparametric bootstrapping procedure was used to construct an empirical distribution for the *RMSD* statistic for three different sample sizes under the null-hypothesis or assumption that the property of item parameter invariance was being met. Although California produced an *RMSD* value that may be considered large for the Math assessment, item parameter invariance seemed evident with the other states. Similar supportive patterns of invariance were observed for Reading, and again California appeared to be at slight variance to the findings in the other states.

Lastly, Study 4 examined the equating between 2005 and 2003 for each state and the national sample via differences in conversion tables constructed from TCCs. Because the conversion tables were similar across the states and the national sample, it was concluded that the scales exhibited minimal differences. Still small differences could be seen in the area of Reading. The pattern of TCCs for equating 2003 and 2005 reading scores, tended to be slightly different using national versus using state data.

In summary, all four studies support the conclusion that the score scales were comparable across the five selected states compared to the national sample. In other words, the scores and performance classifications for a particular state appeared to have about the same meaning regardless of whether they were obtained as part of the national sample or obtained from state data. This finding lends credence to the validity of the operational procedure implemented by ETS and the construction of the NAEP Math and Reading assessments for the Grade 8 student population. Furthermore, it appears unnecessary to perform separate scaling for states because it would not influence the interpretation of the findings meaningfully.

This page left intentionally blank

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Dorans, N.J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, 39, 59–84.
- Dorans, N.J., and Holland, P.W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281–306.
- Kingsbury, G. G., Olson, A., Cronin, J. C., Hauser, C., and Houser, R. L. (December, 2004). *The state of standards: Research investigating proficiency levels in fourteen states*. Portland, Oreg.: NWEA.
- Kolen, M.J. and Brennan, R.J. (2004). *Test equating, scaling, and linking: Methods and Practices* (2nd ed.). New York: Springer-Verlag.
- Liu, J., Cahn, M. F., and Dorans, N. J. (2006). An application of score equity assessment: Invariance of linkage of new SAT to old SAT across gender groups. *Journal of Educational Measurement*, 43, 113–130.
- Lord, F.M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, N.J.: Erlbaum.
- Smithson, J. (February, 2006). Alignment standards, assessments, and instruction using surveys of enacted curriculum. A paper presented at the SEC Collaborative Membership Meeting, Austin, Texas.

This page left intentionally blank

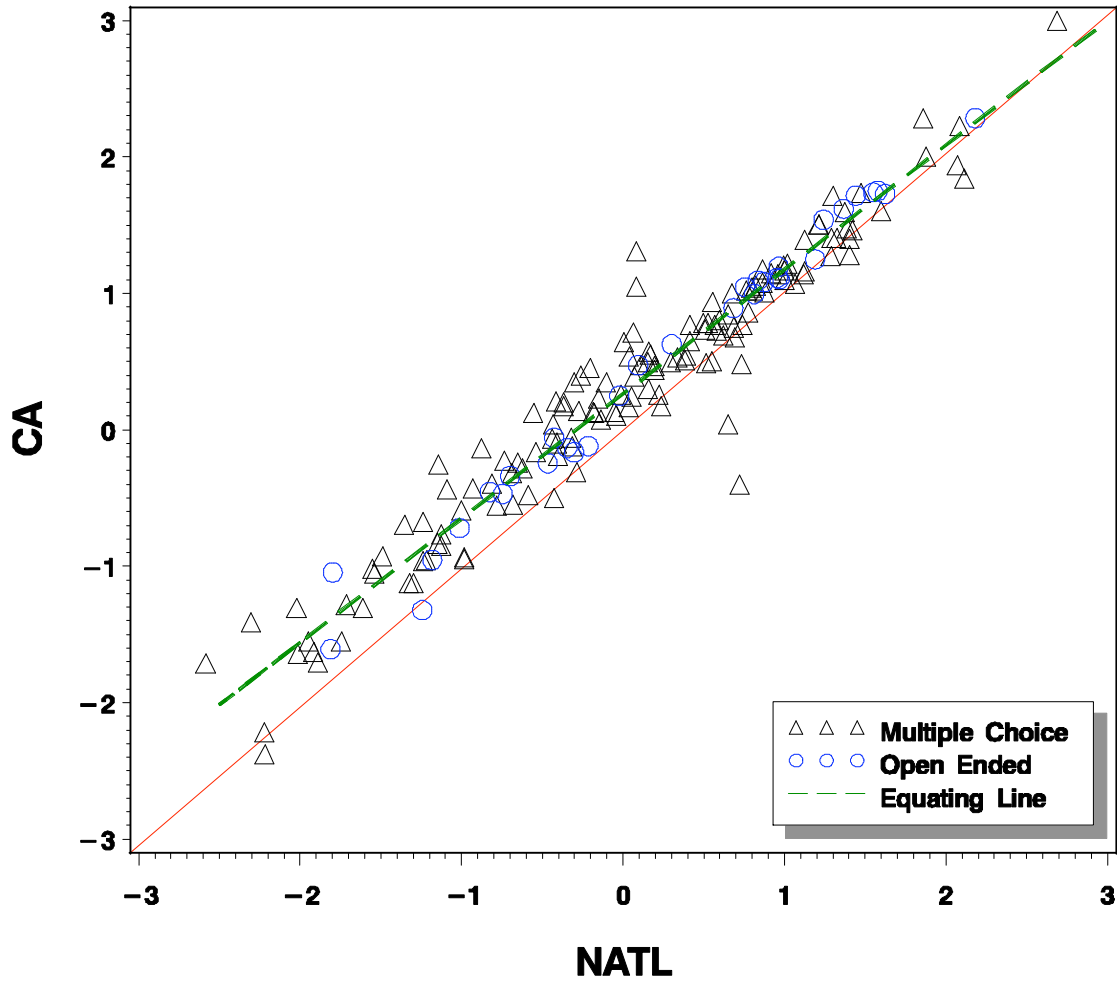
Appendixes

This page left intentionally blank

Appendix A: *a*- and *b*-plots for 2005, Grade 8, Math Assessment

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National

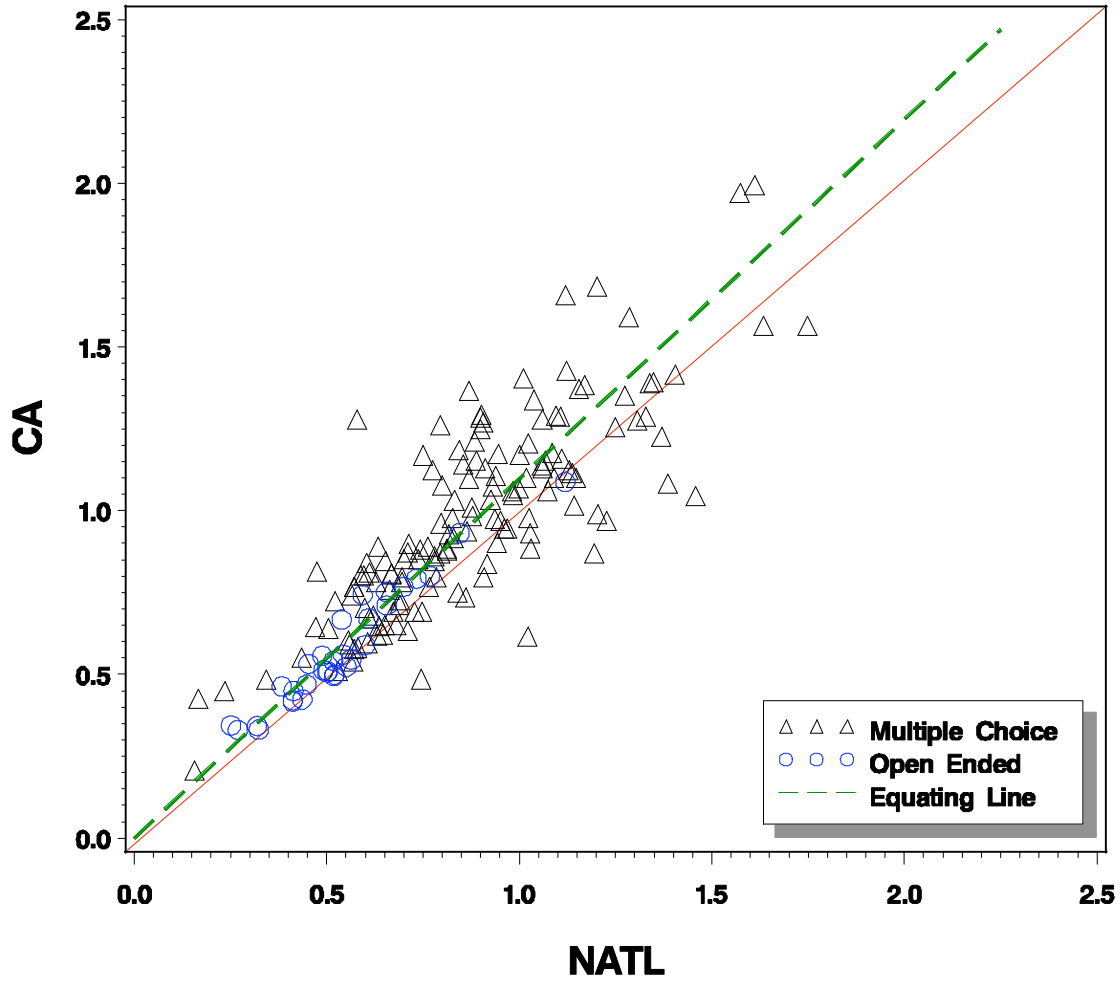
2005 NAEP Math Gr 8 b–plot: CA vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

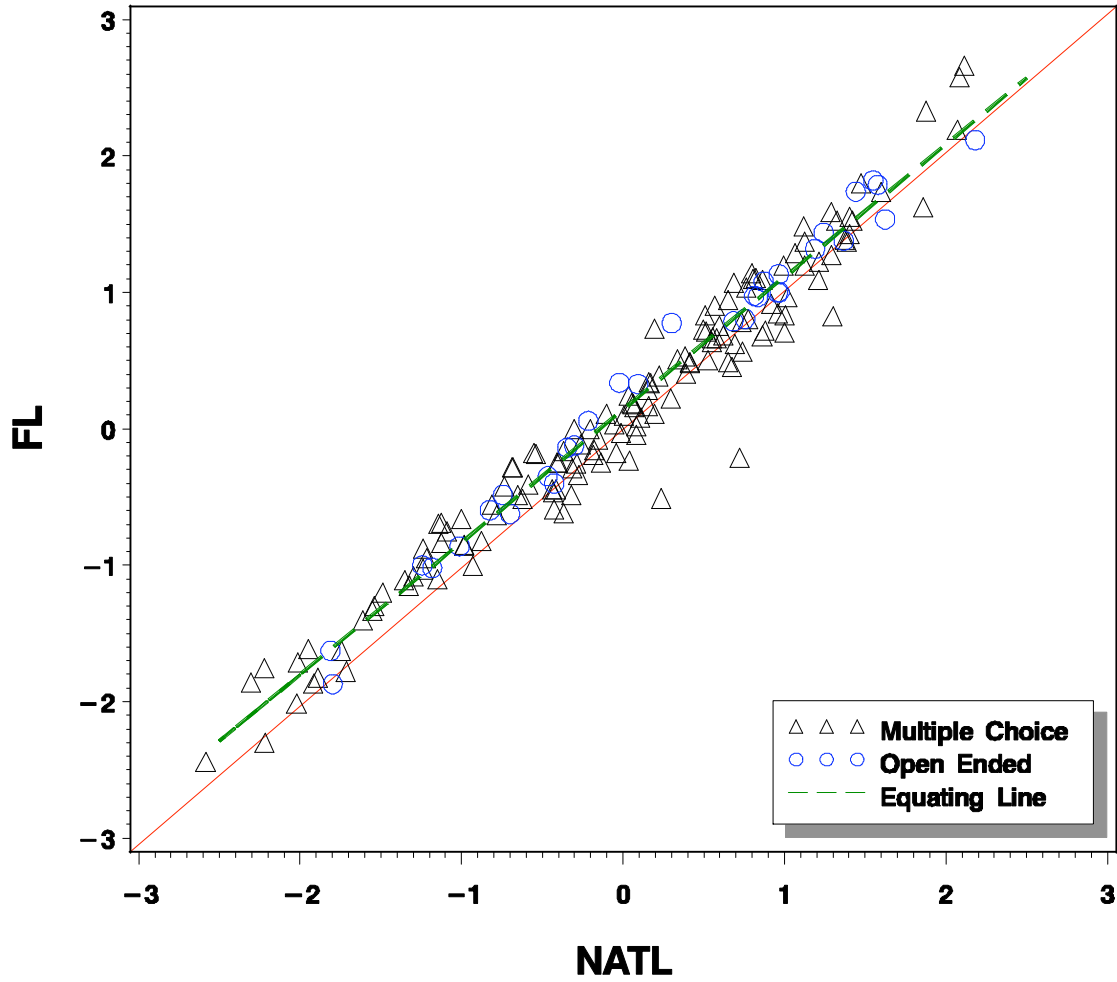
2005 NAEP Math Gr 8 a–plot: CA vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

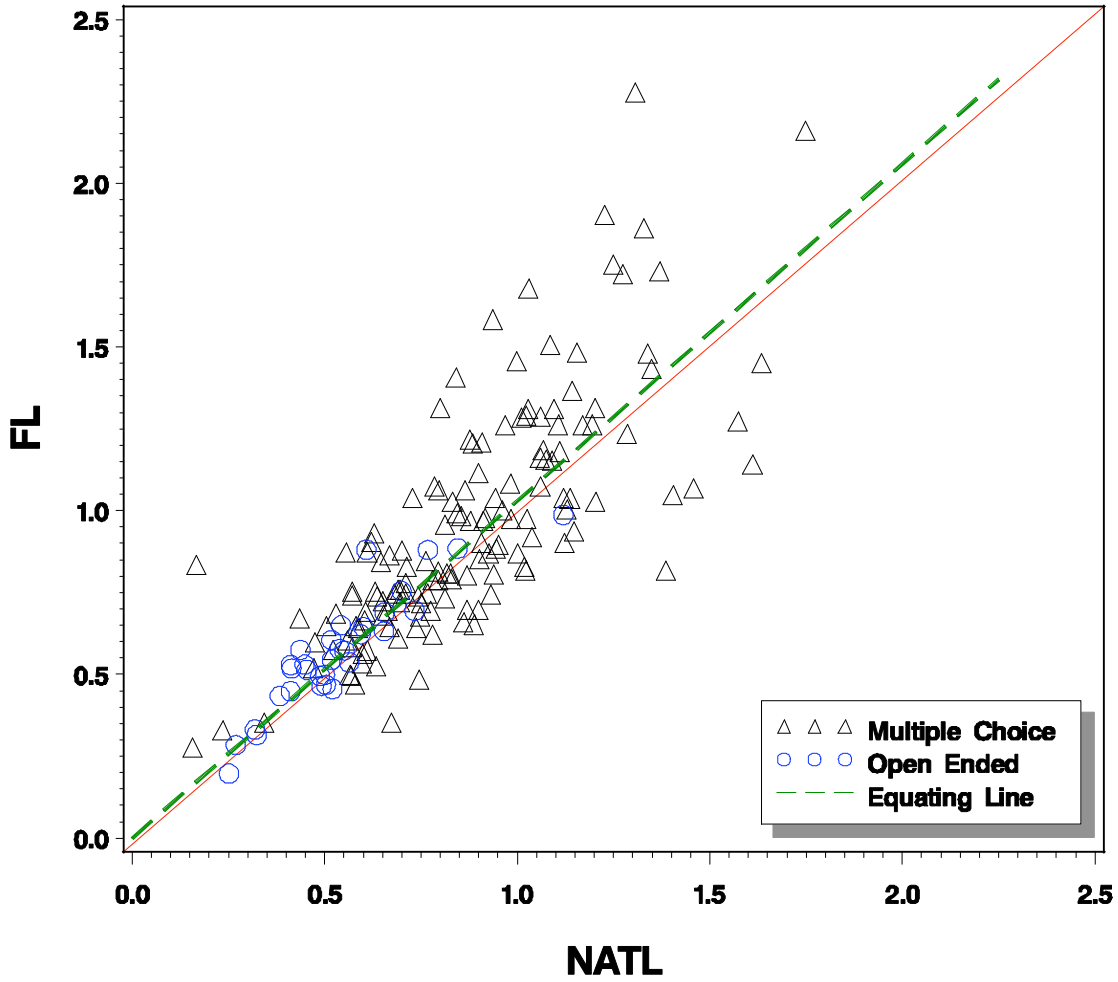
2005 NAEP Math Gr 8 b—plot: FL vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

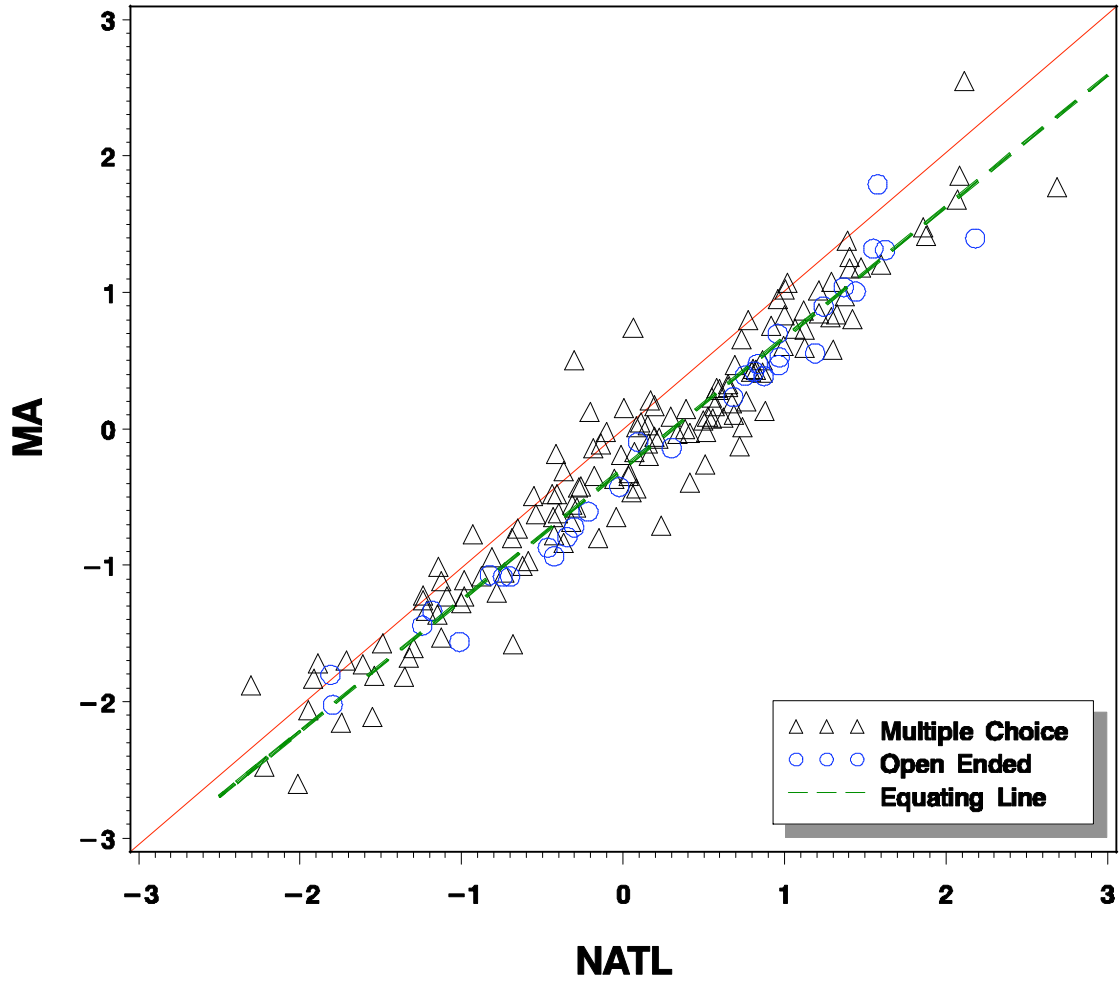
2005 NAEP Math Gr 8 a–plot: FL vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

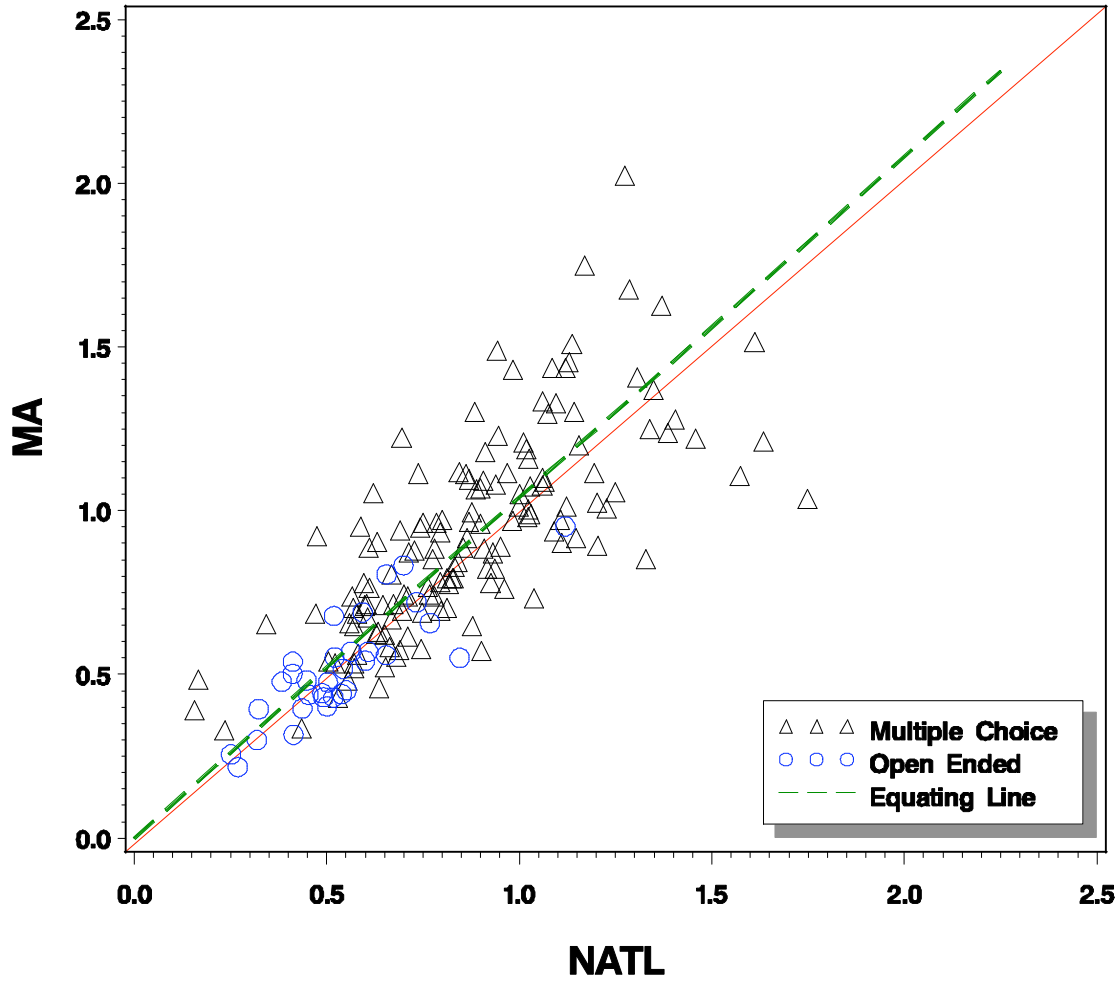
2005 NAEP Math Gr 8 b—plot: MA vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

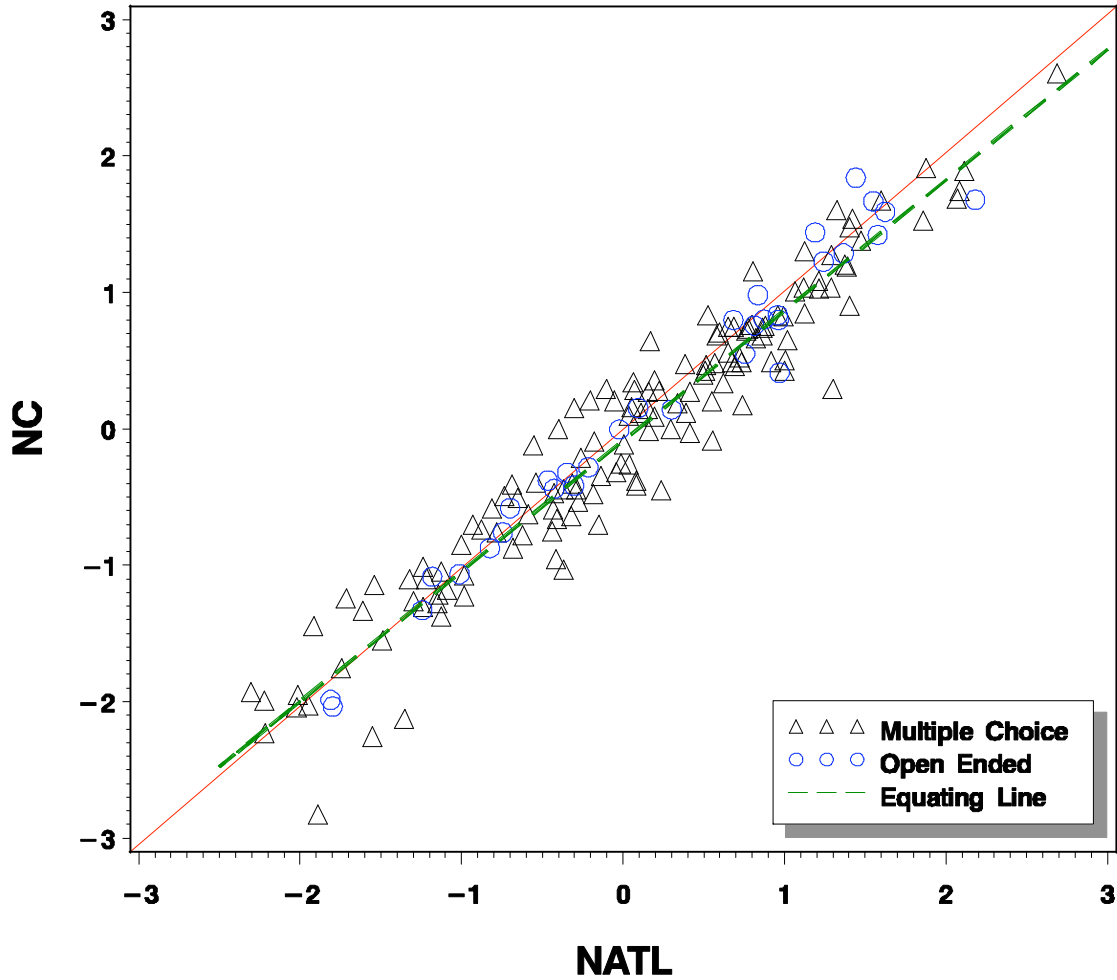
2005 NAEP Math Gr 8 a – plot: MA vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

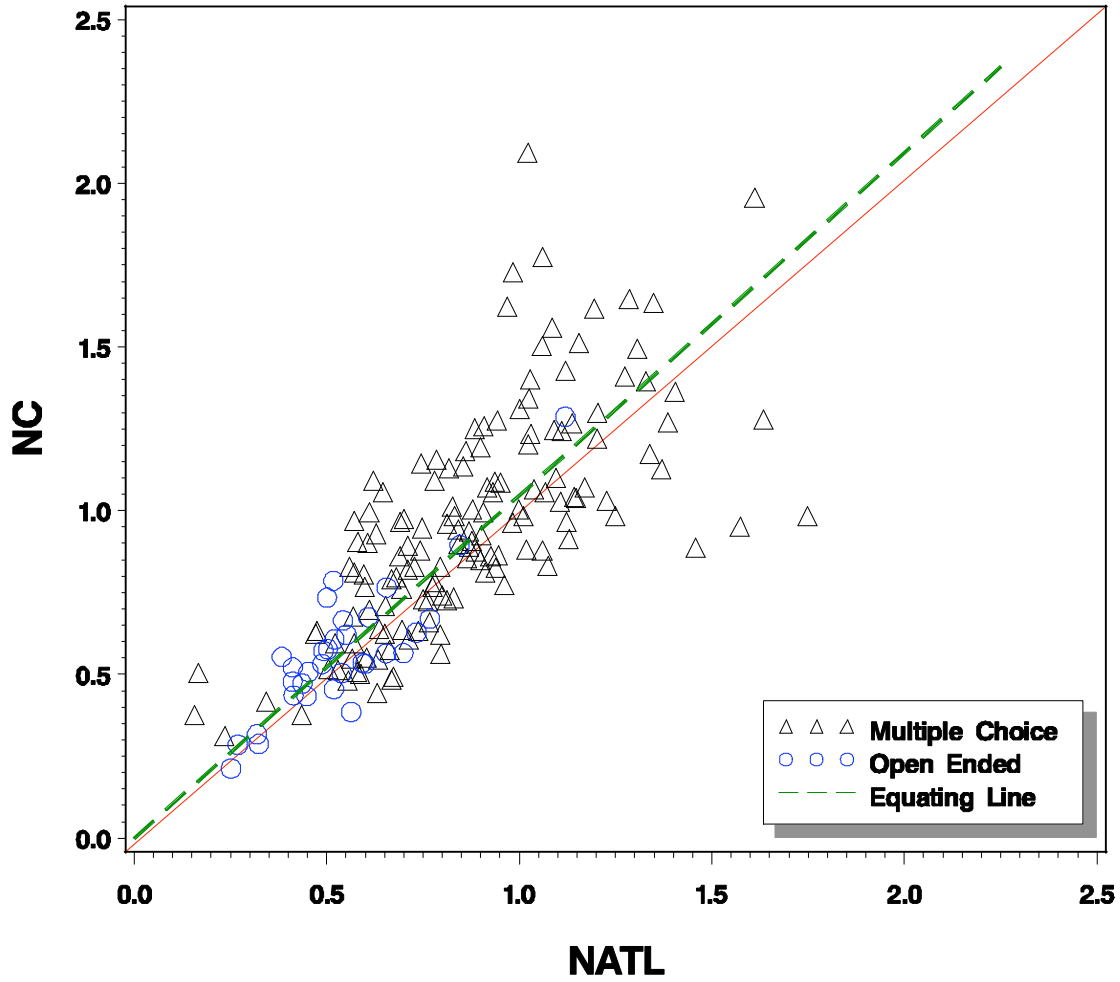
2005 NAEP Math Gr 8 b—plot: NC vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

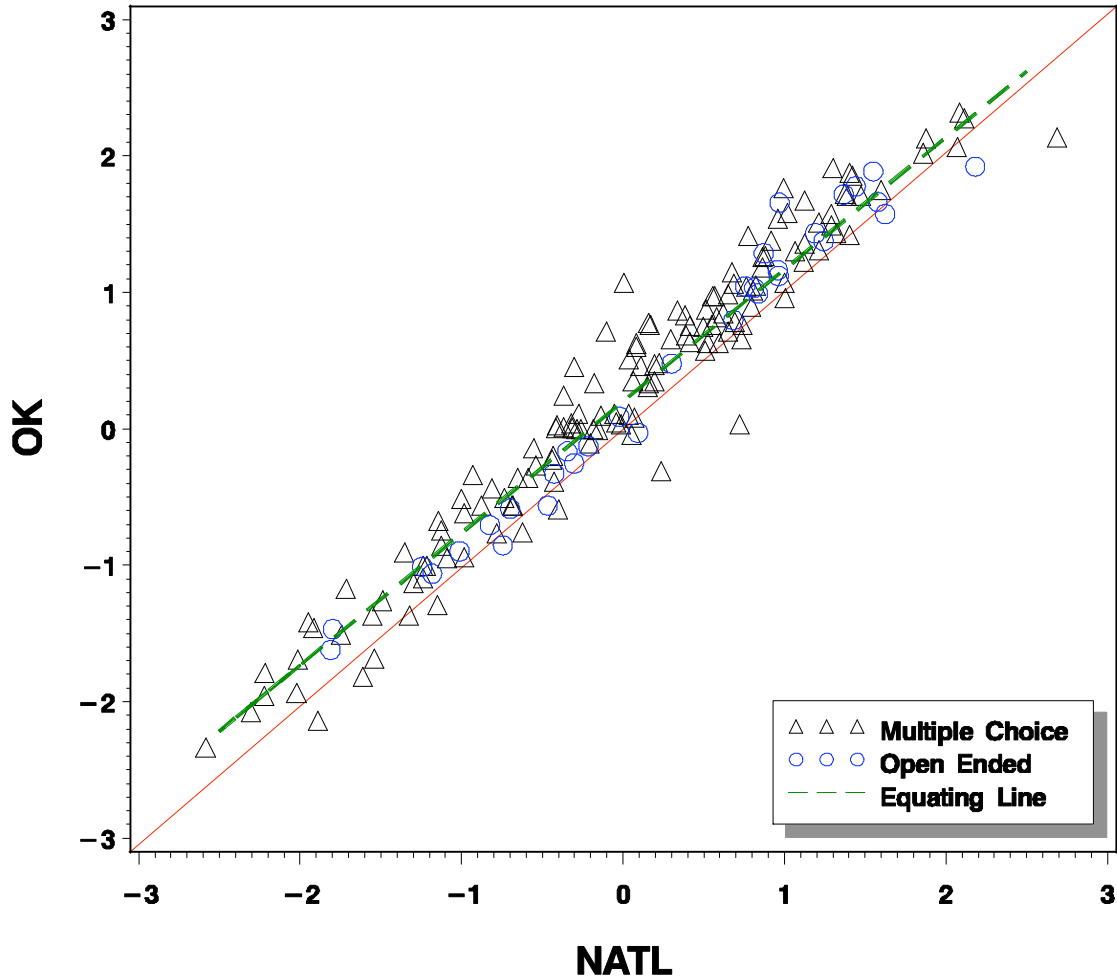
2005 NAEP Math Gr 8 a—plot: NC vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

2005 NAEP Math Gr 8 b—plot: OK vs NATL



Continues next page

Figure A-1. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs National (Continued)

2005 NAEP Math Gr 8 a – plot: OK vs NATL

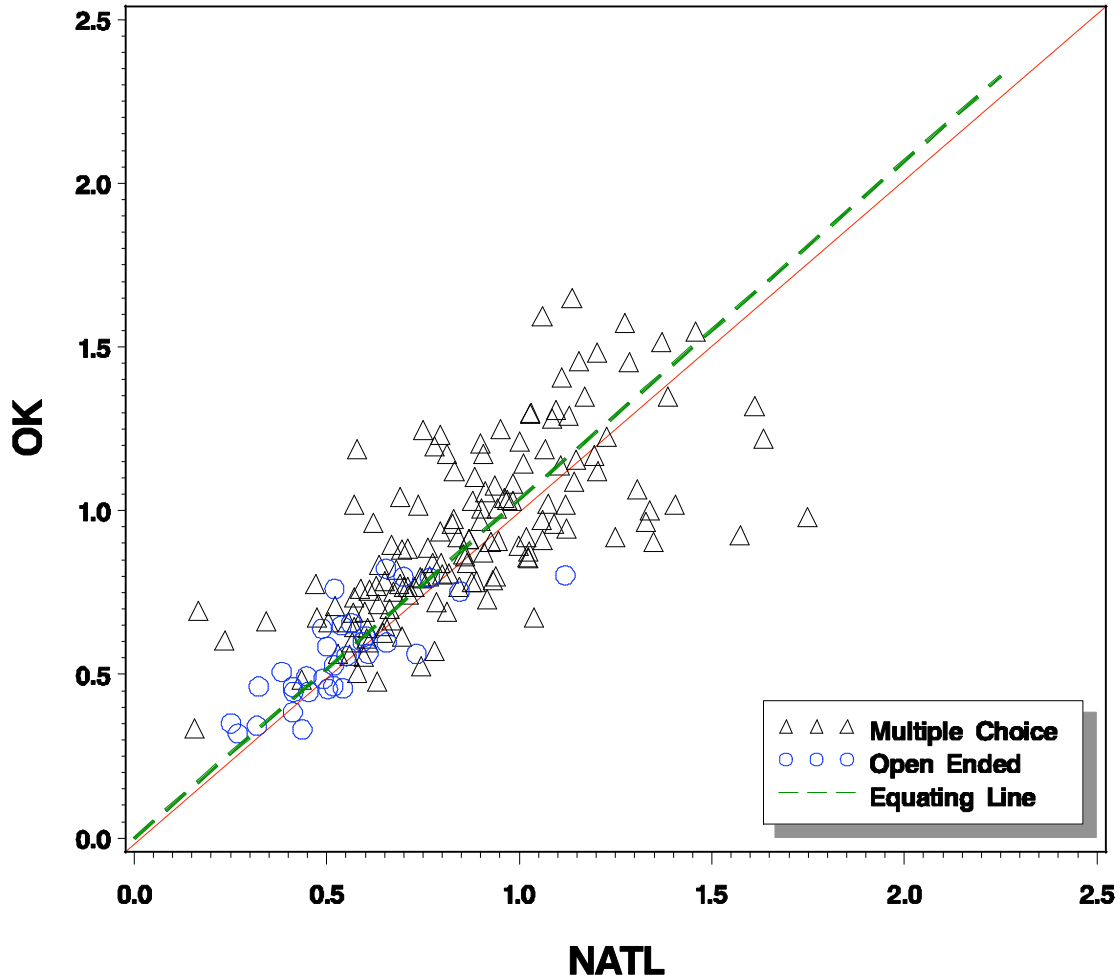
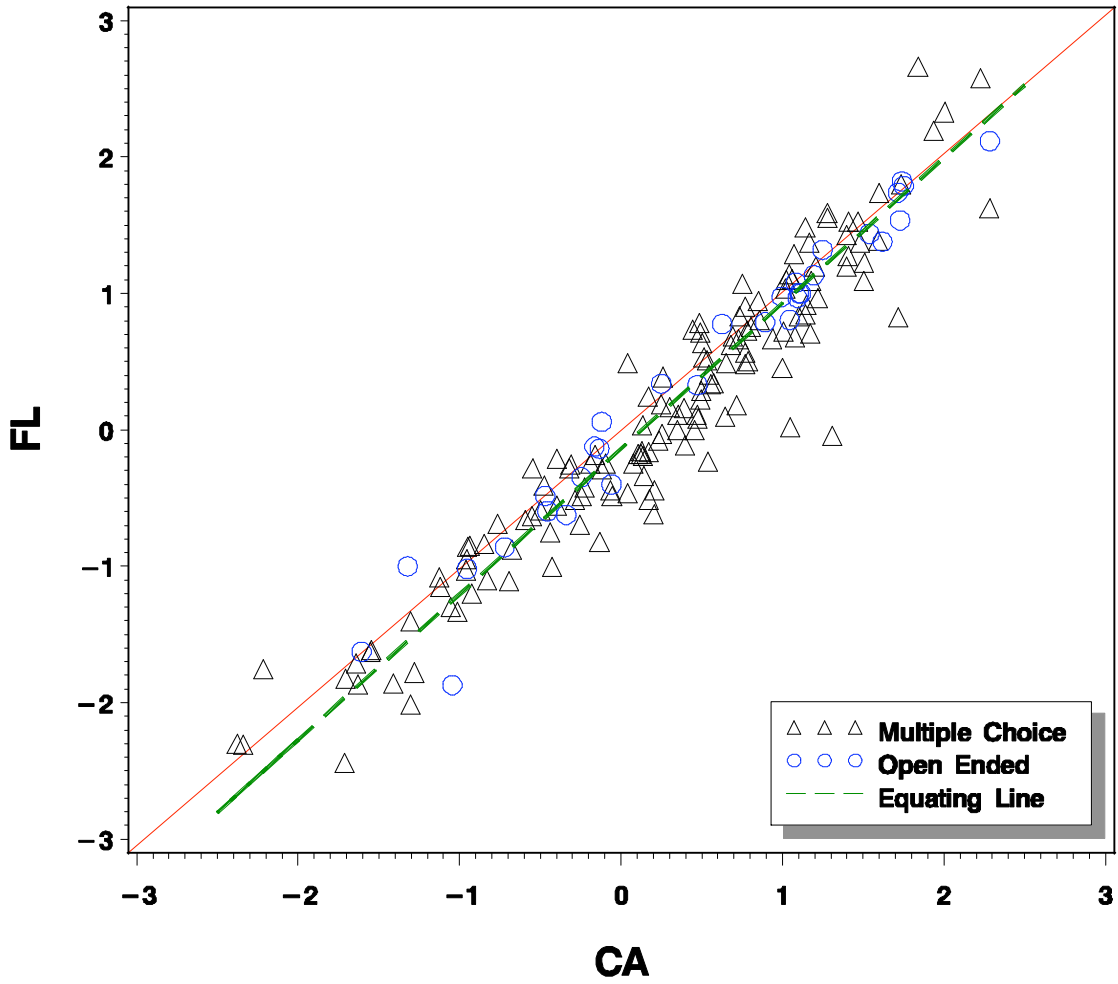


Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States

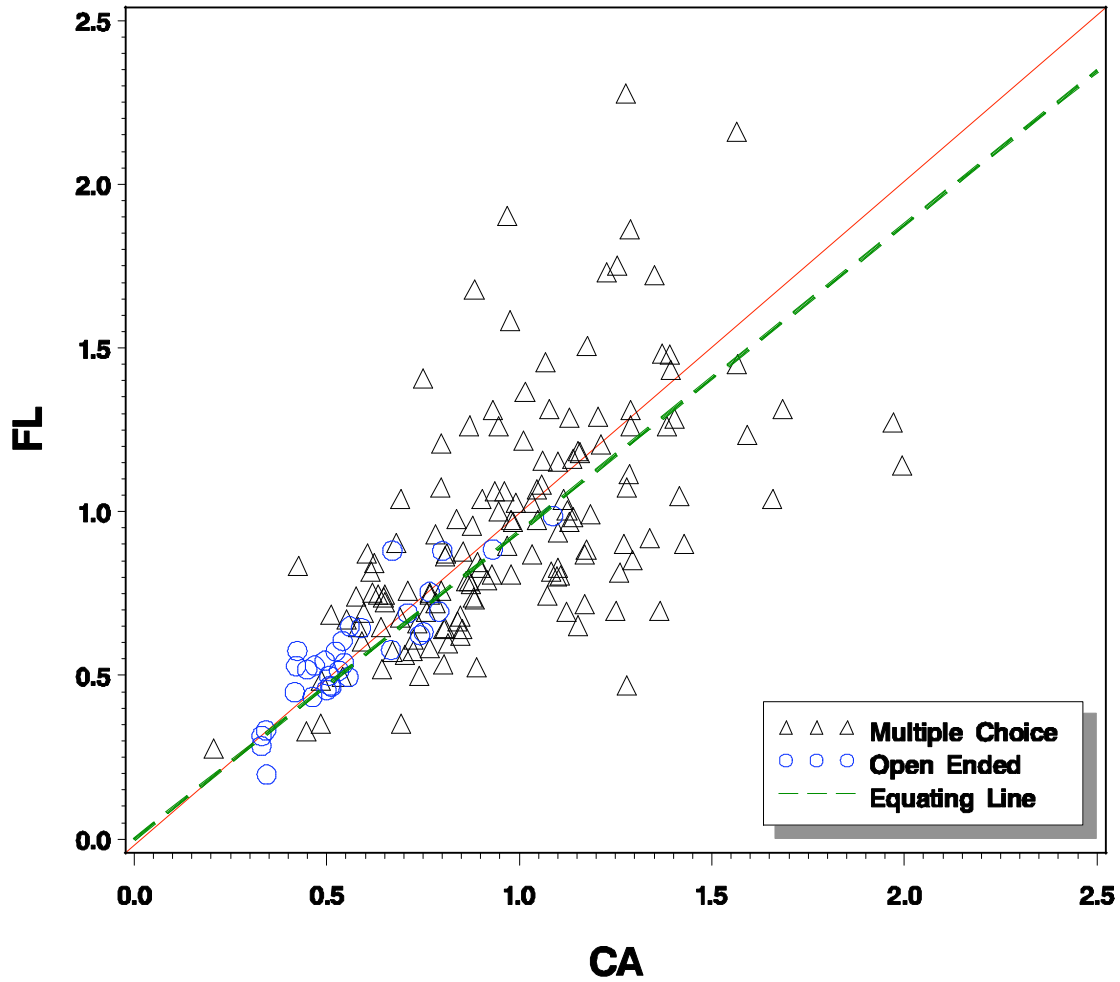
2005 NAEP Math Gr 8 b–plot: FL vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

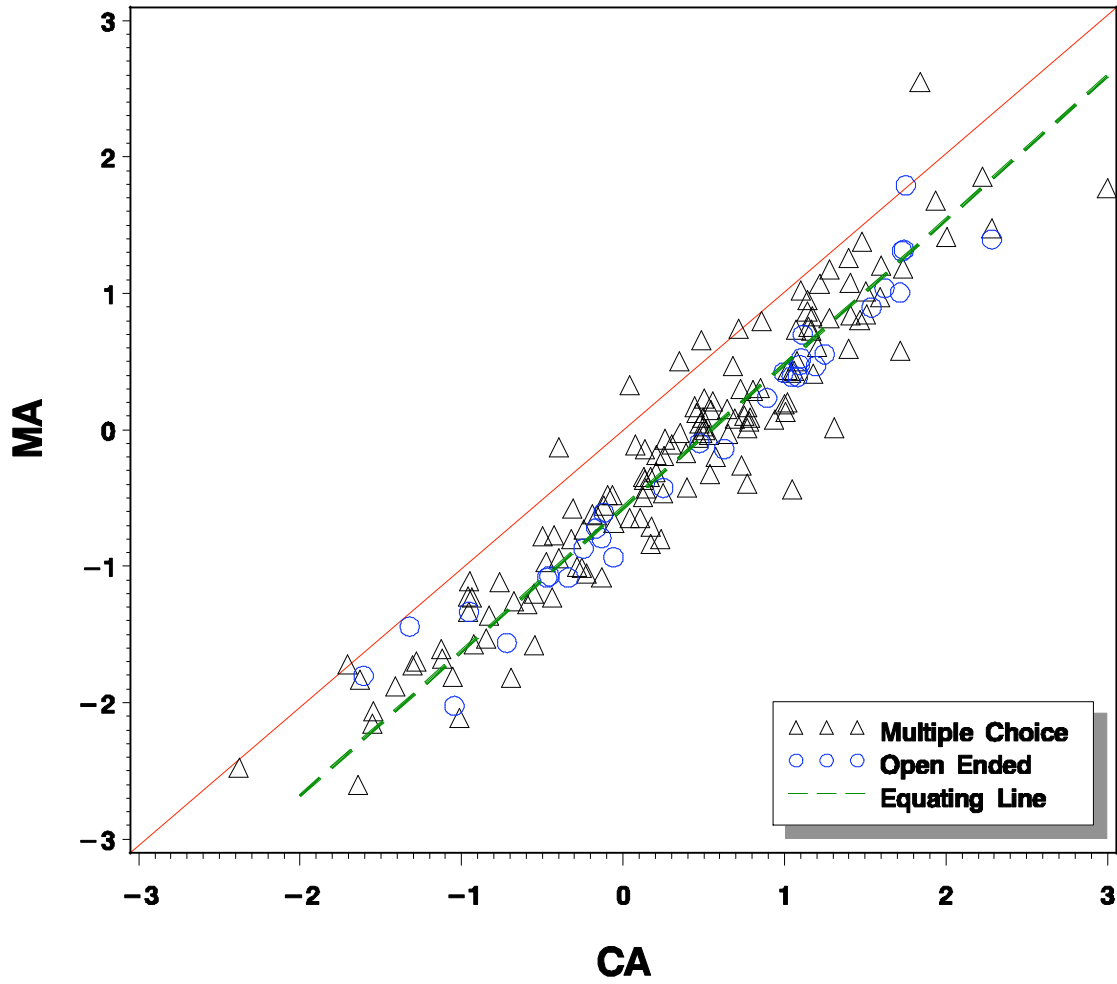
2005 NAEP Math Gr 8 a–plot: FL vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

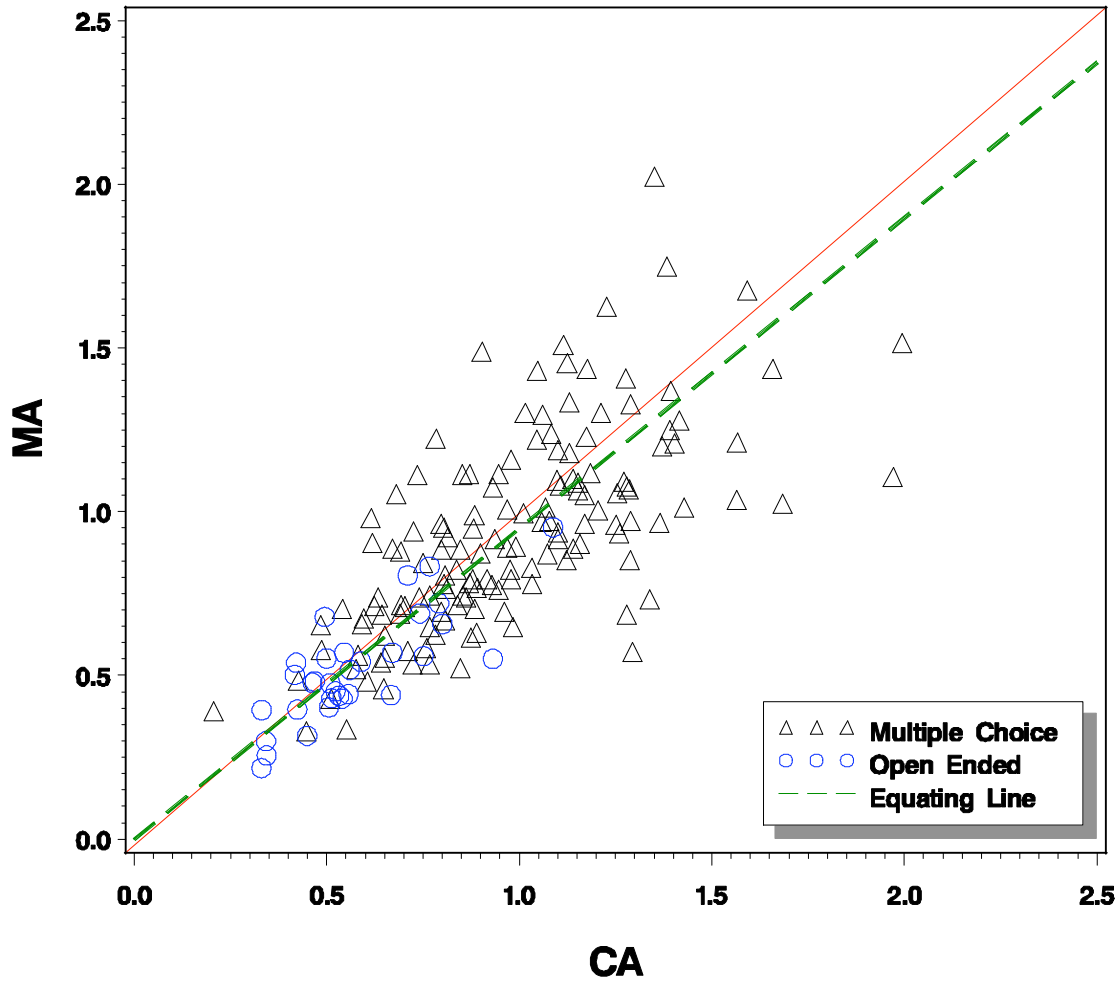
2005 NAEP Math Gr 8 b—plot: MA vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

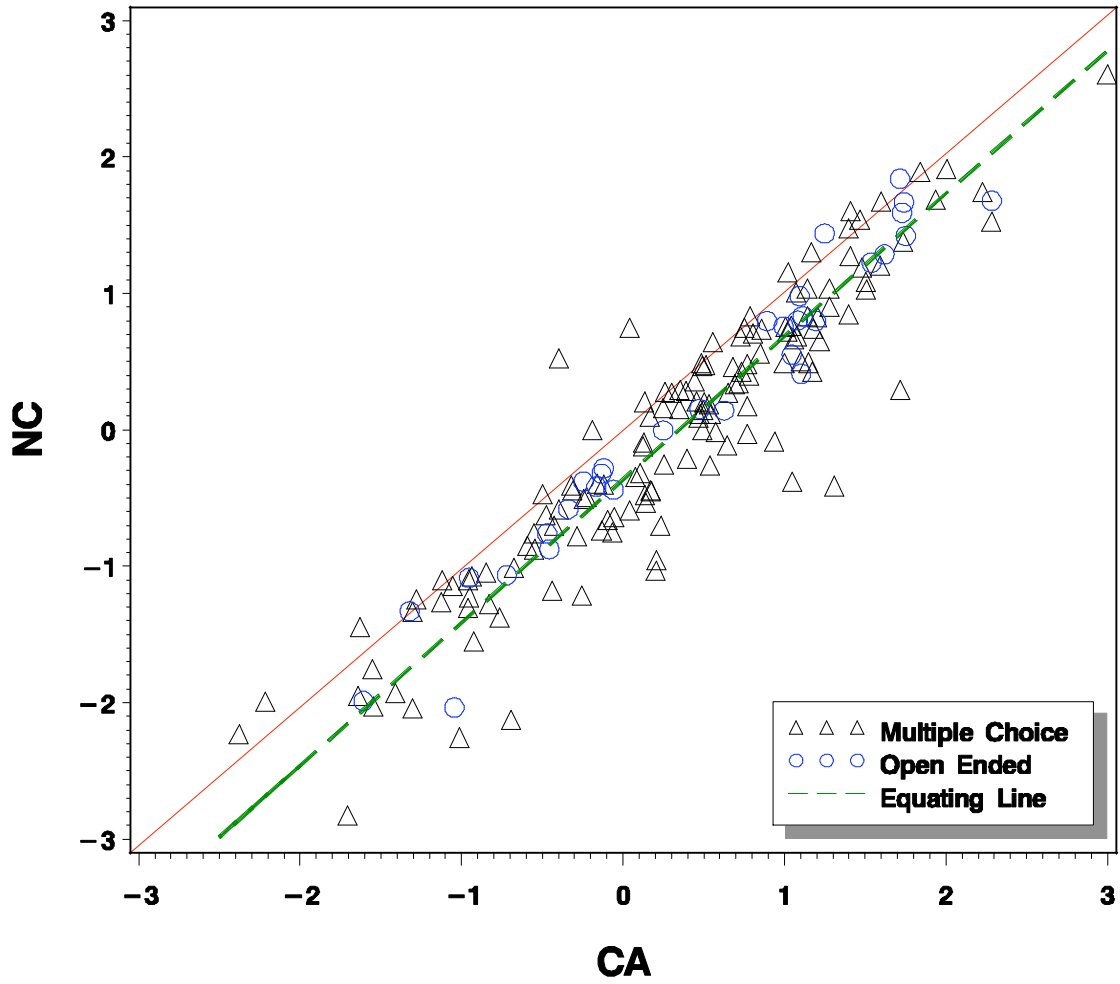
2005 NAEP Math Gr 8 a – plot: MA vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

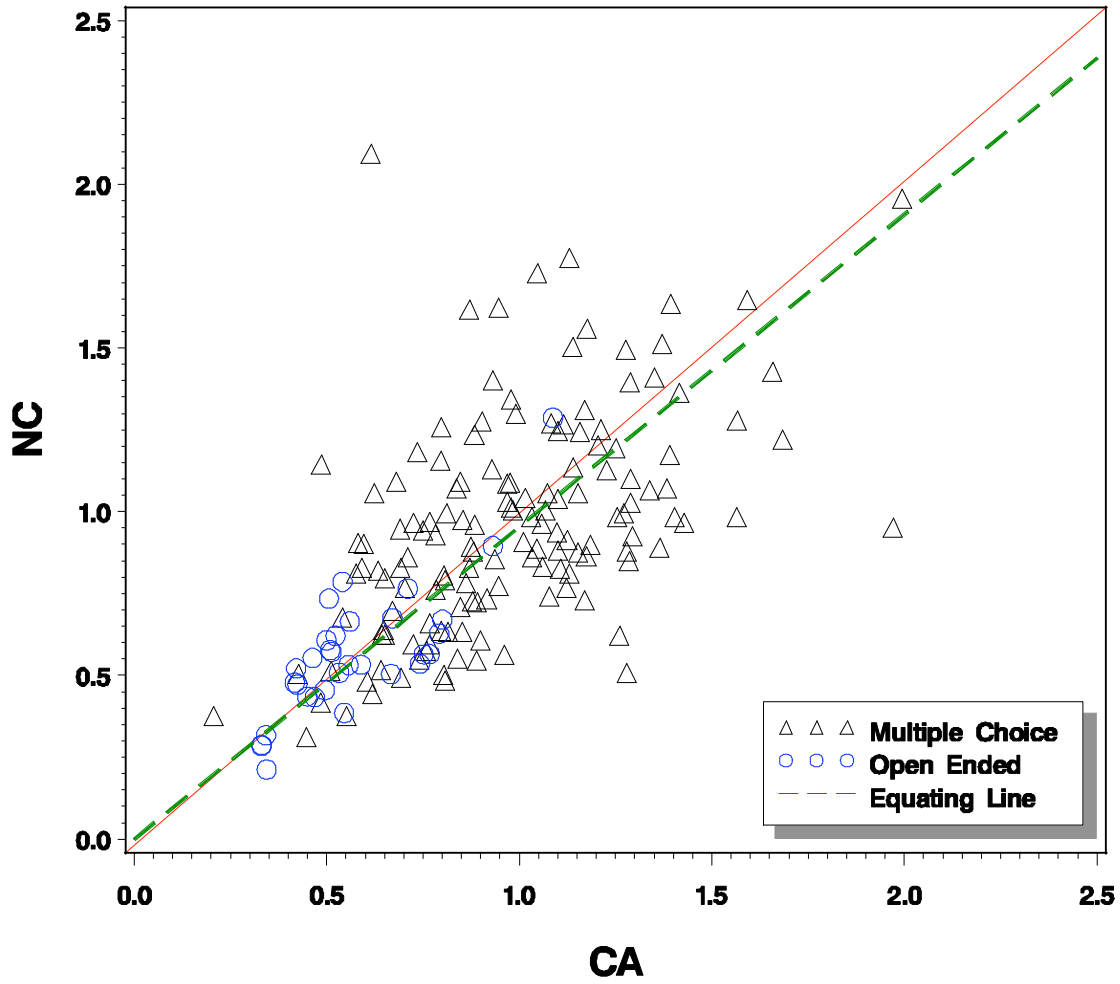
2005 NAEP Math Gr 8 b—plot: NC vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

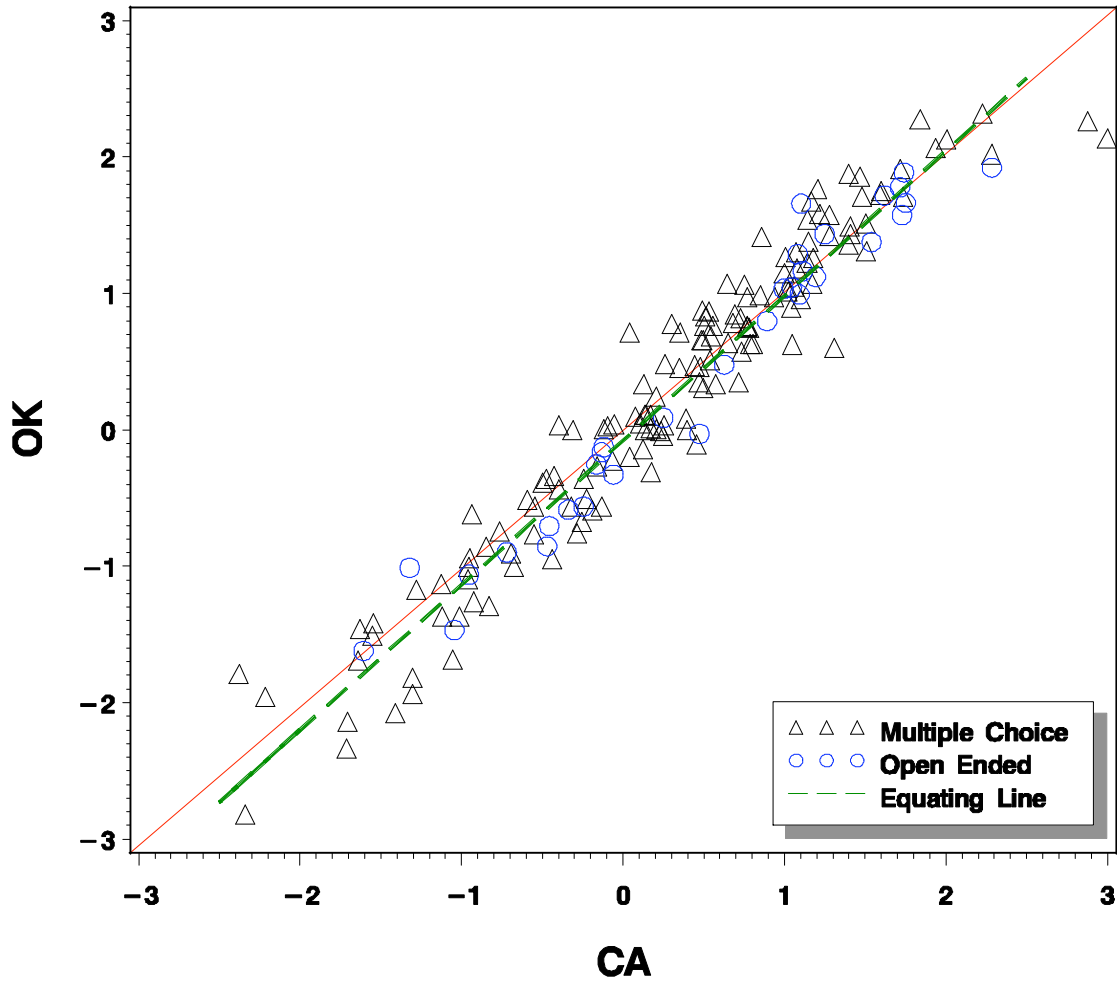
2005 NAEP Math Gr 8 a – plot: NC vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States (Continued)

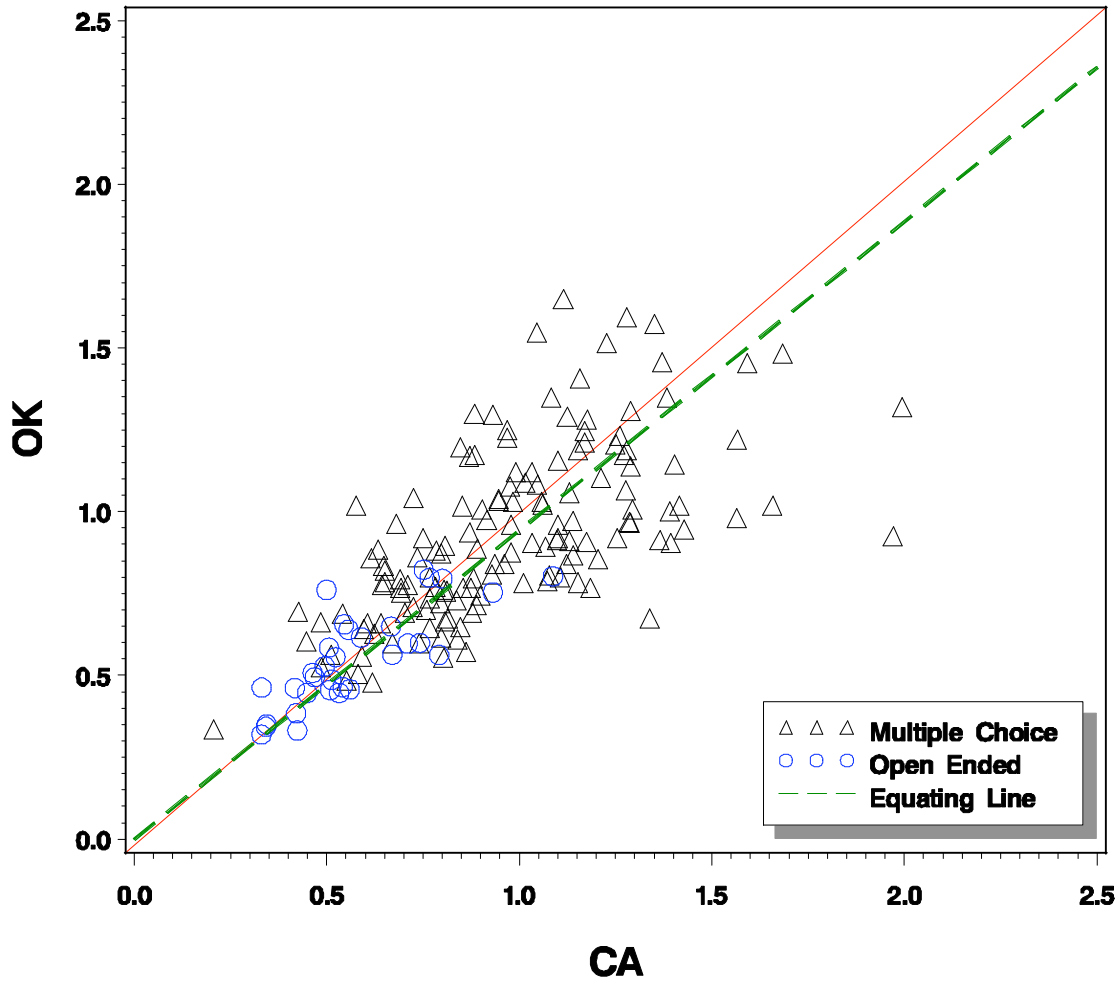
2005 NAEP Math Gr 8 b—plot: OK vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

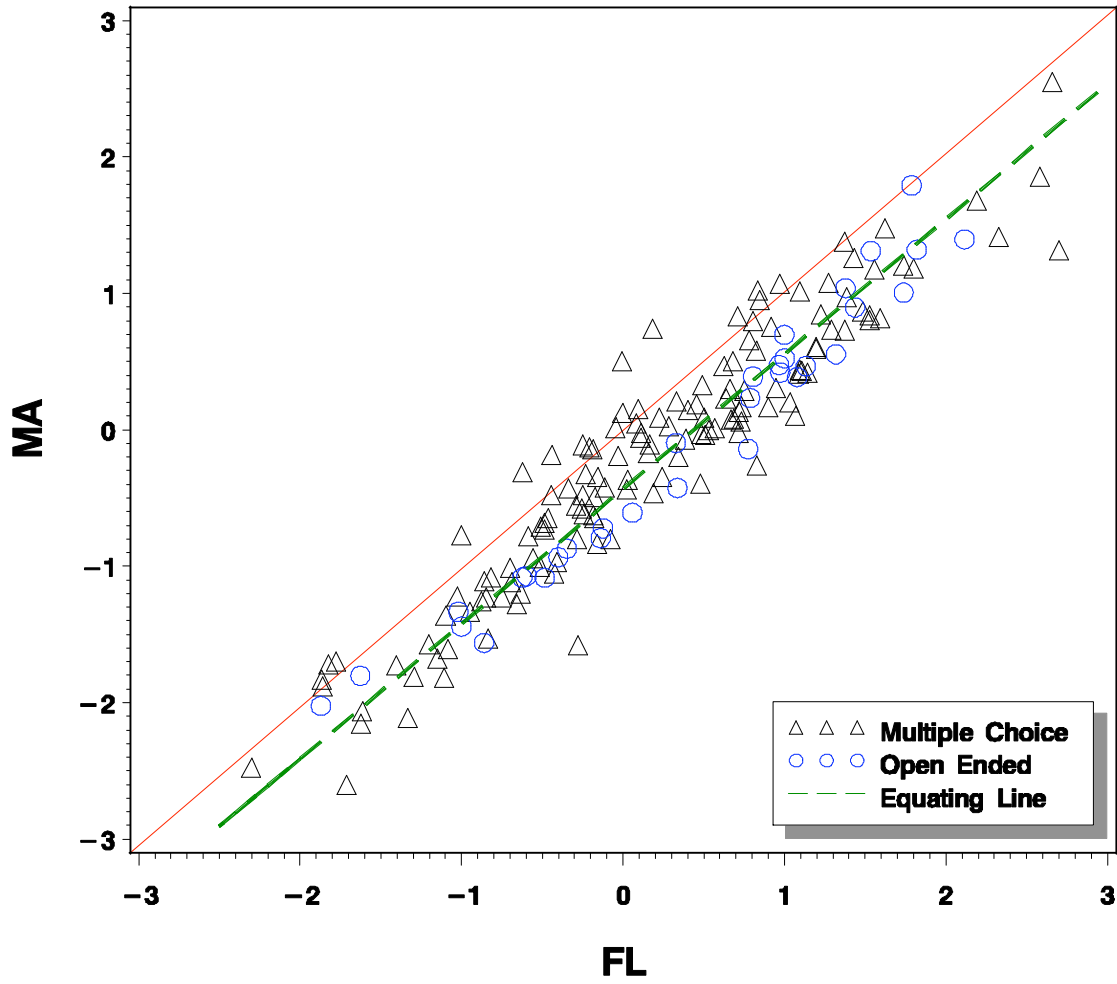
2005 NAEP Math Gr 8 a – plot: OK vs CA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

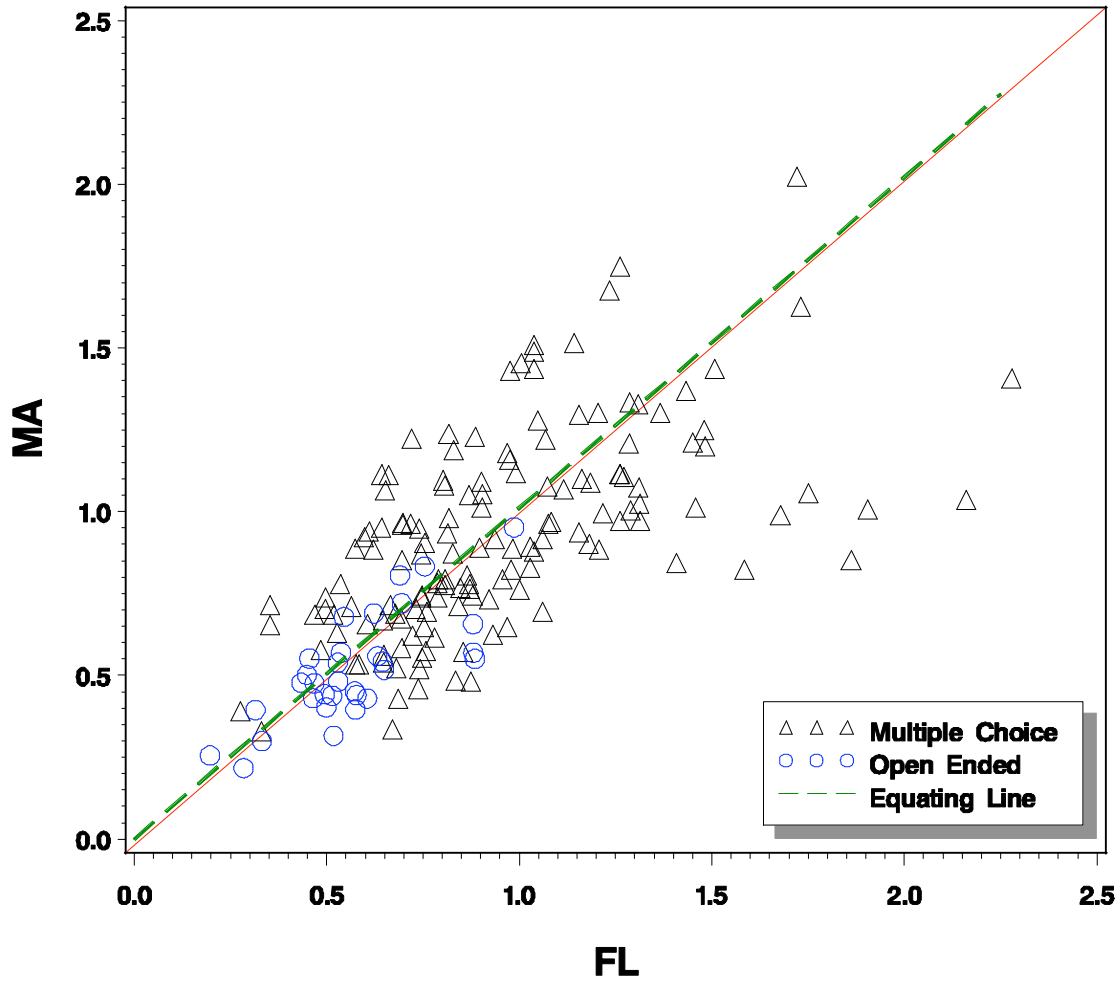
2005 NAEP Math Gr 8 b—plot: MA vs FL



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

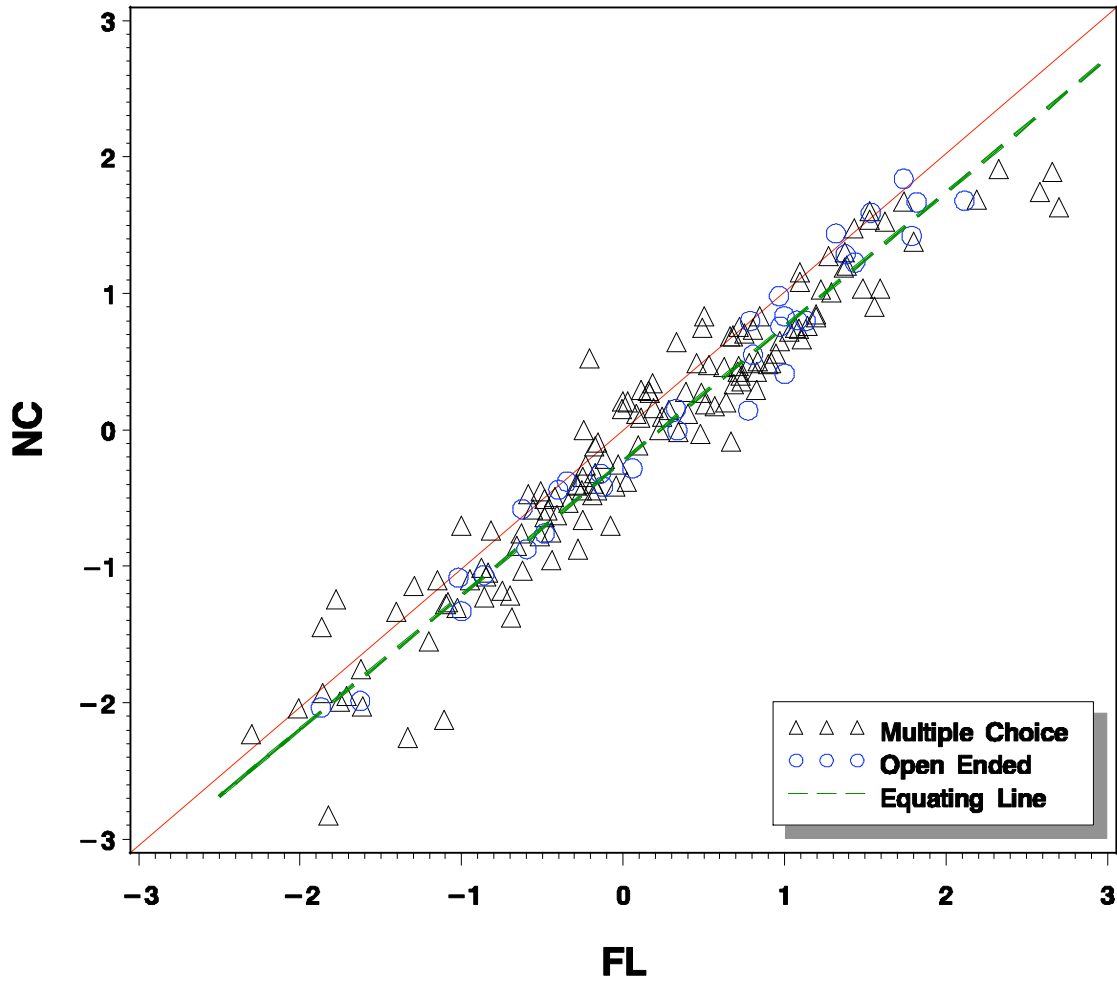
2005 NAEP Math Gr 8 a – plot: MA vs FL



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

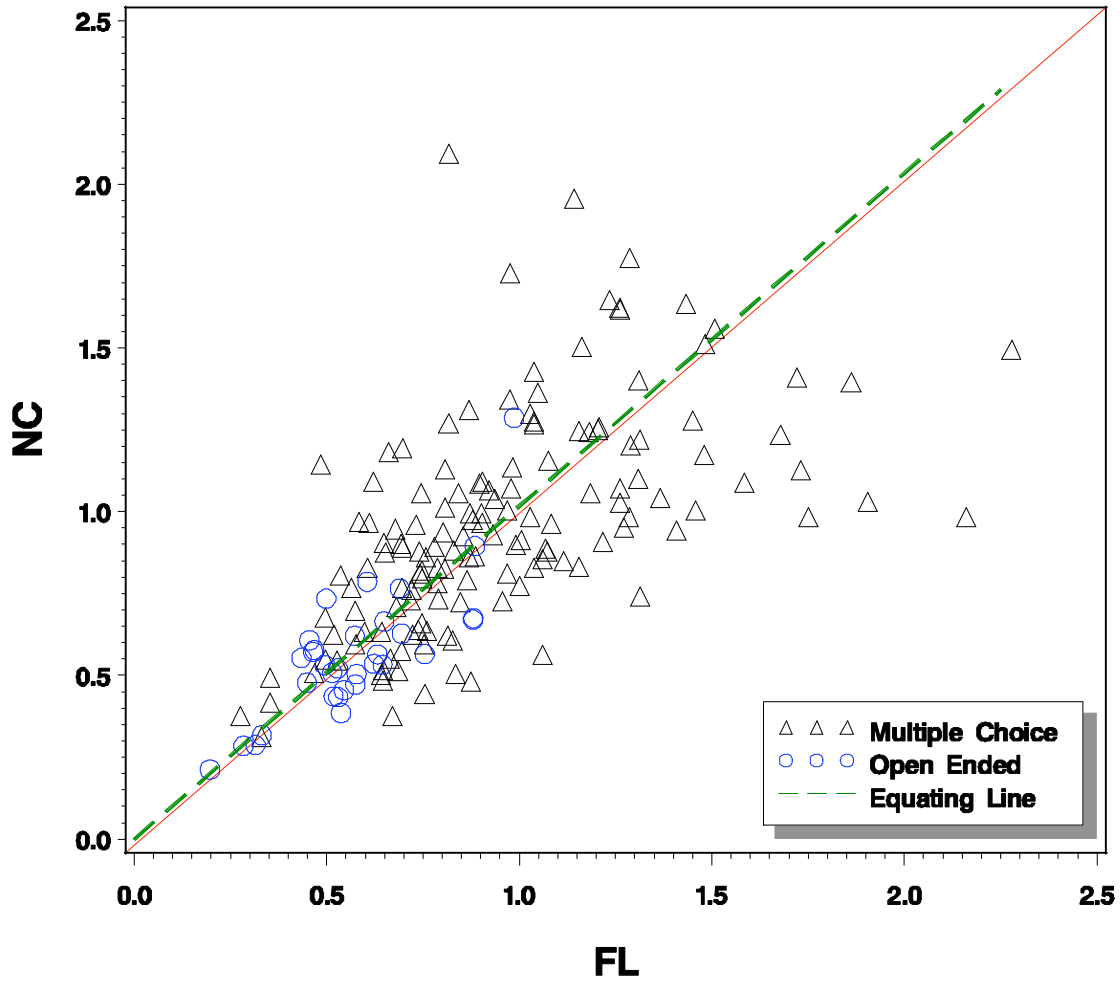
2005 NAEP Math Gr 8 b—plot: NC vs FL



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

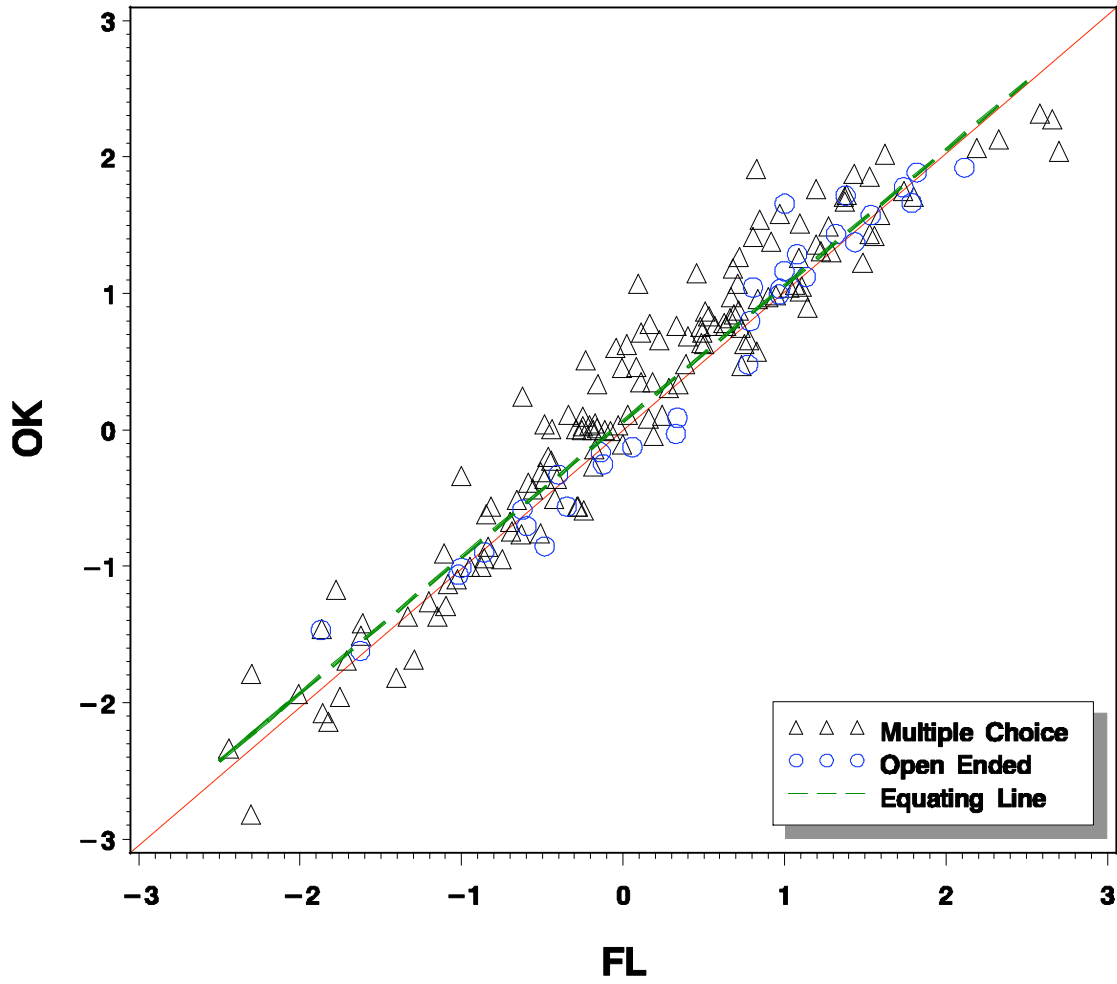
2005 NAEP Math Gr 8 a–plot: NC vs FL



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States (Continued)

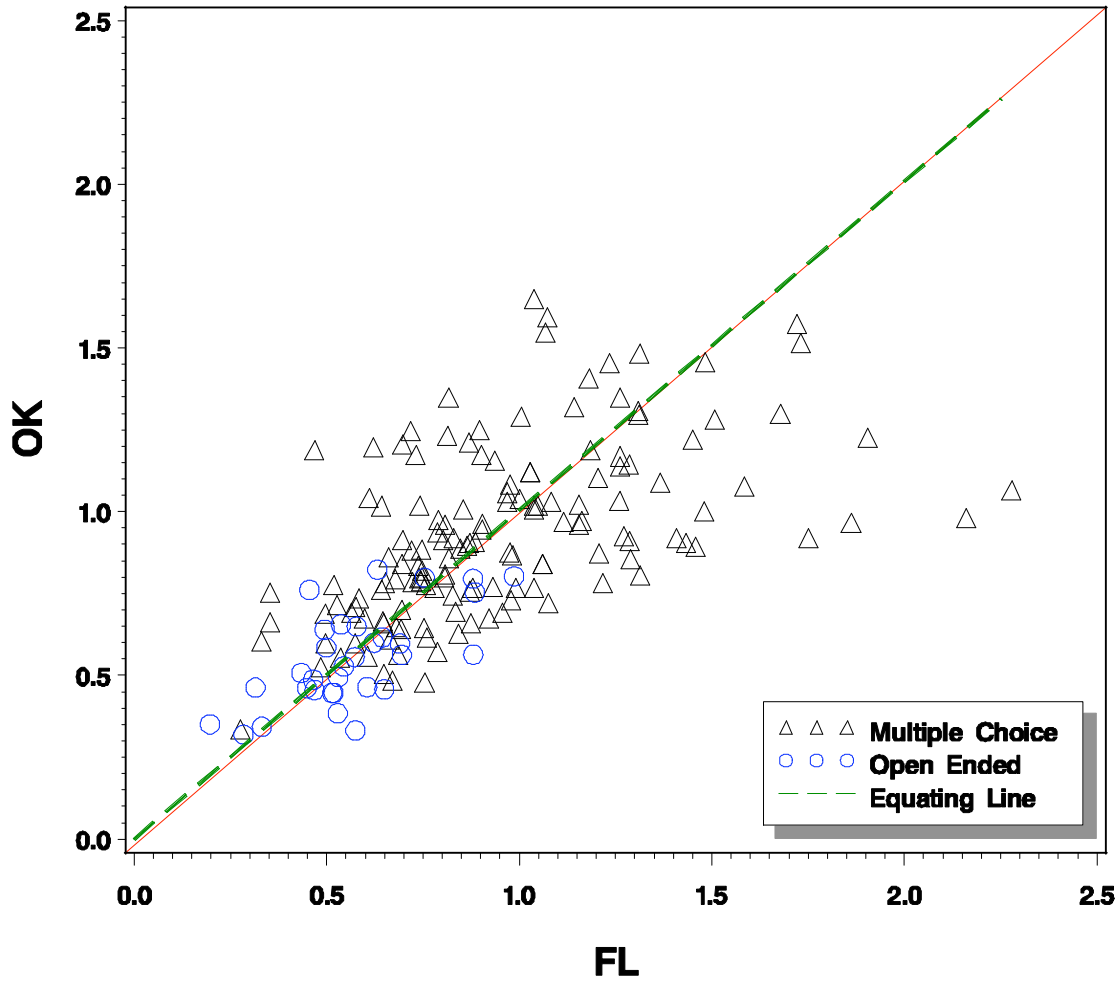
2005 NAEP Math Gr 8 b–plot: OK vs FL



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

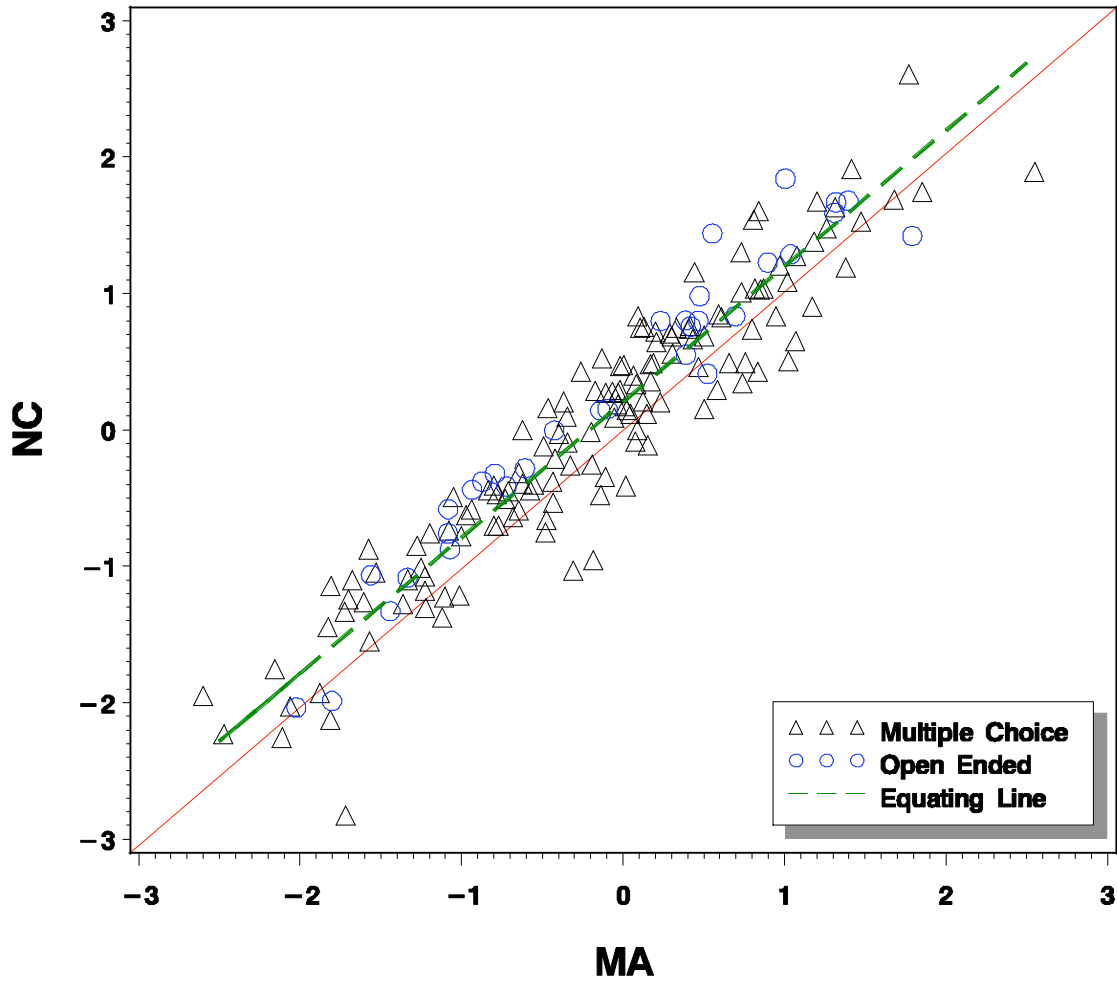
2005 NAEP Math Gr 8 a – plot: OK vs FL



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

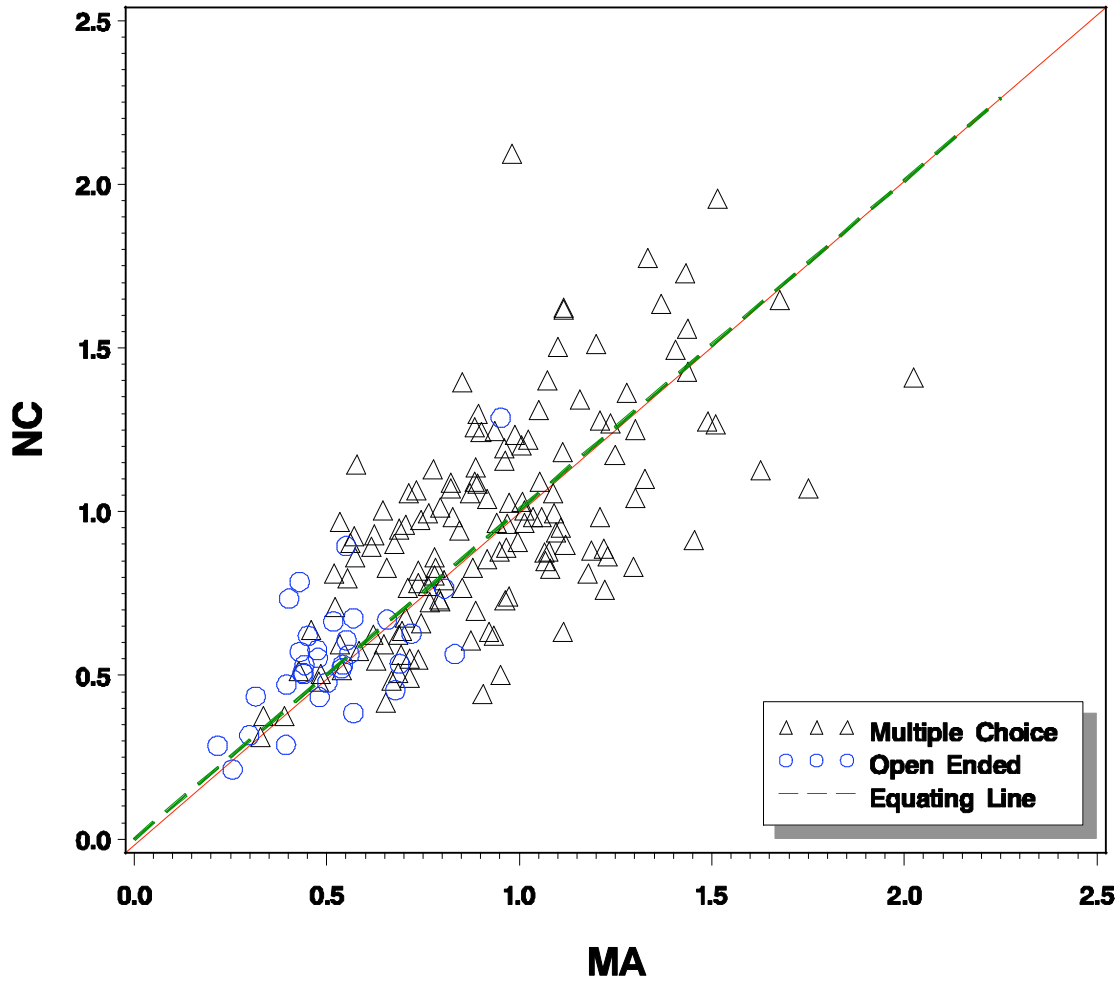
2005 NAEP Math Gr 8 b—plot: NC vs MA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

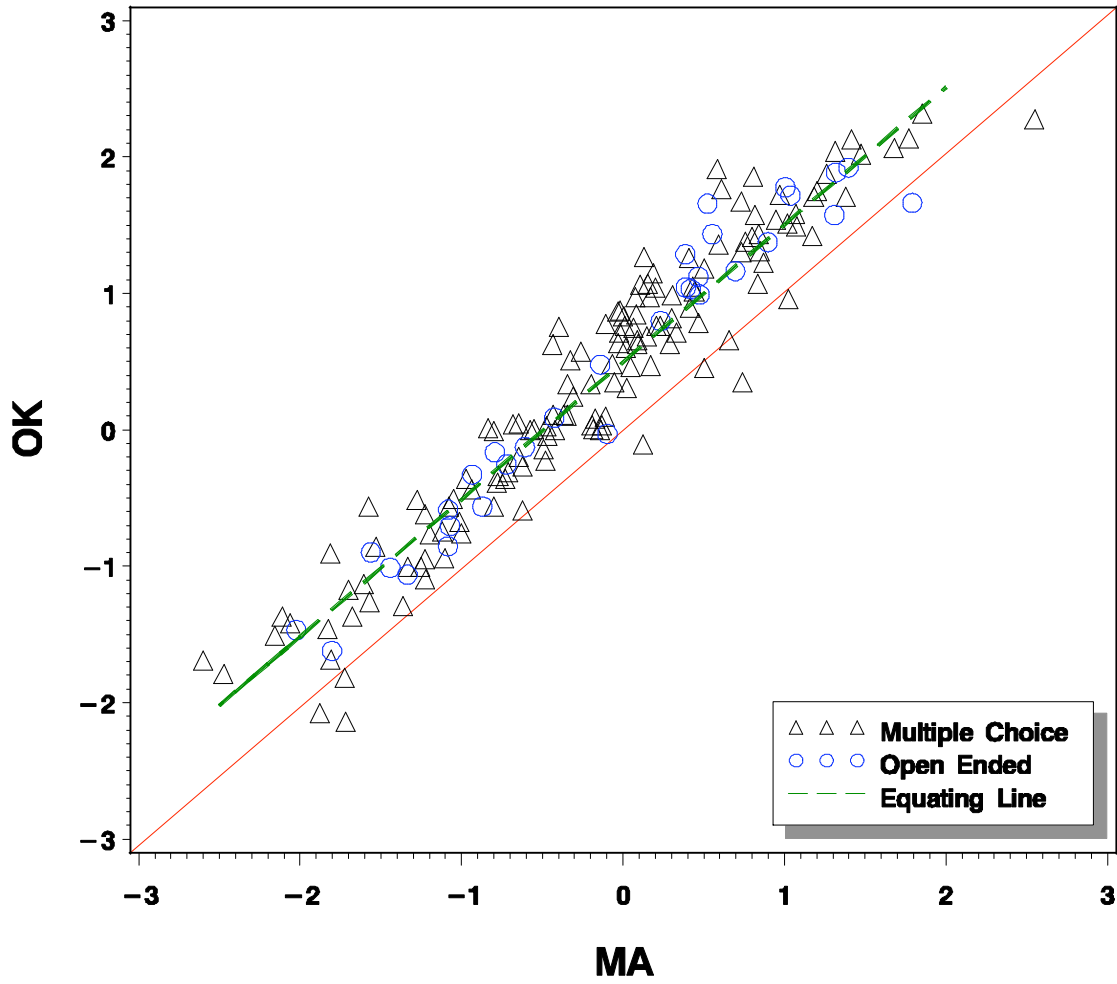
2005 NAEP Math Gr 8 a – plot: NC vs MA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States (Continued)

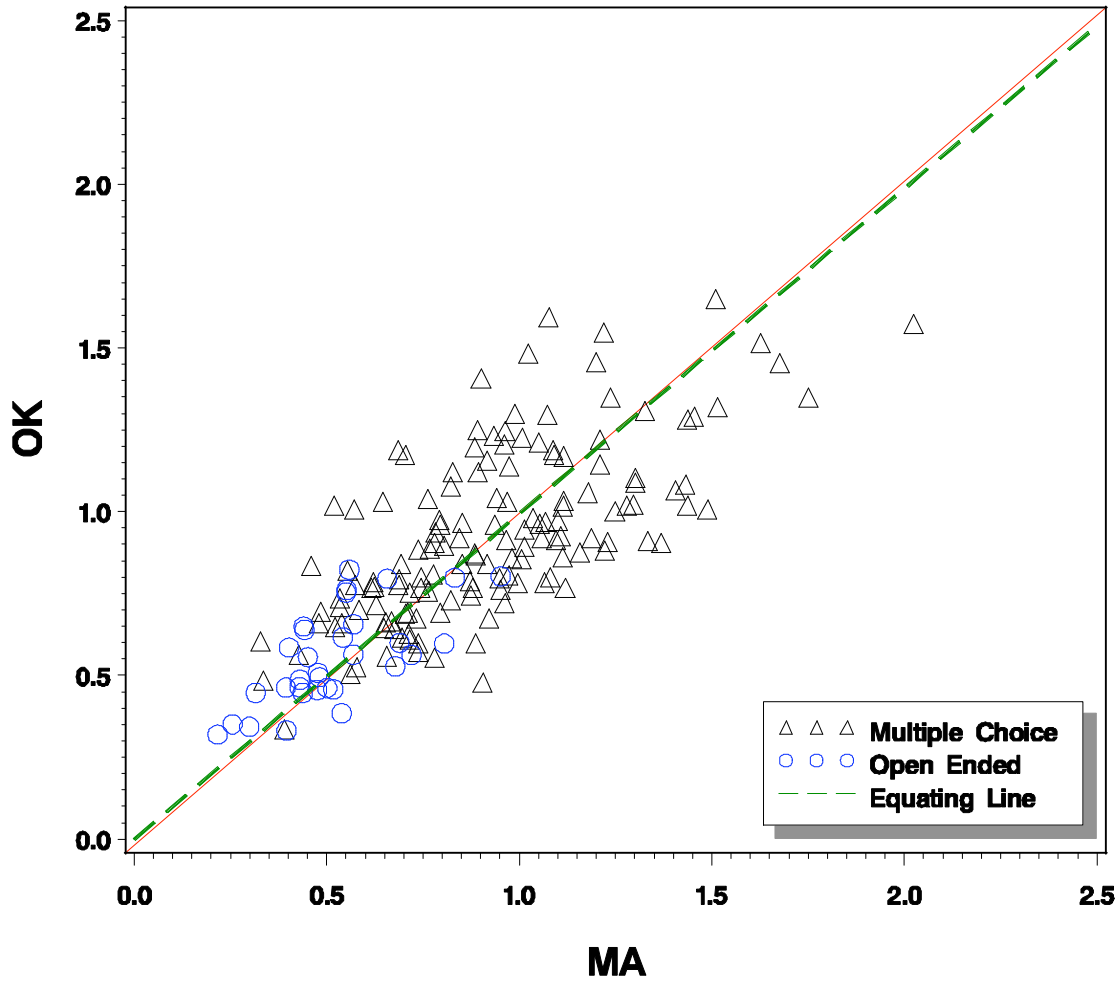
2005 NAEP Math Gr 8 b—plot: OK vs MA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

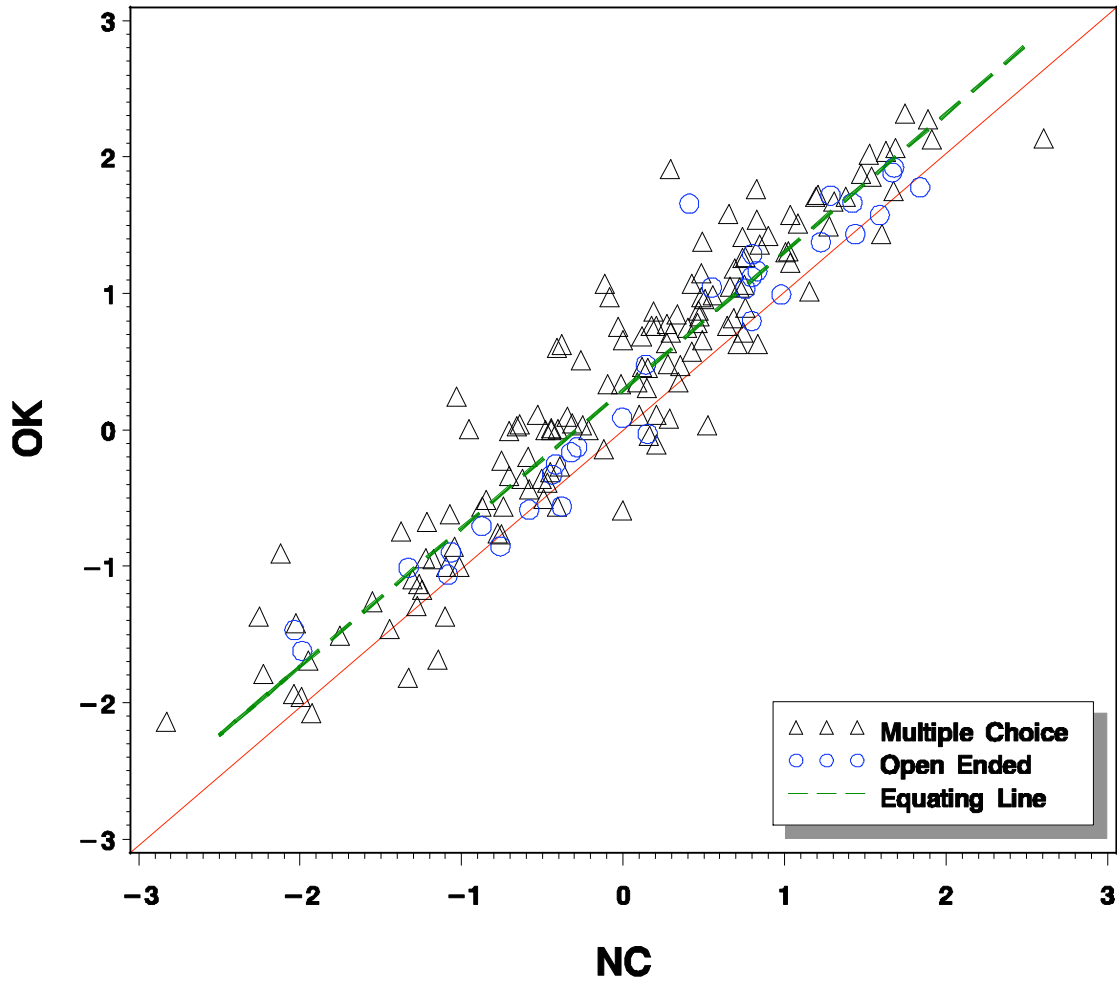
2005 NAEP Math Gr 8 a – plot: OK vs MA



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

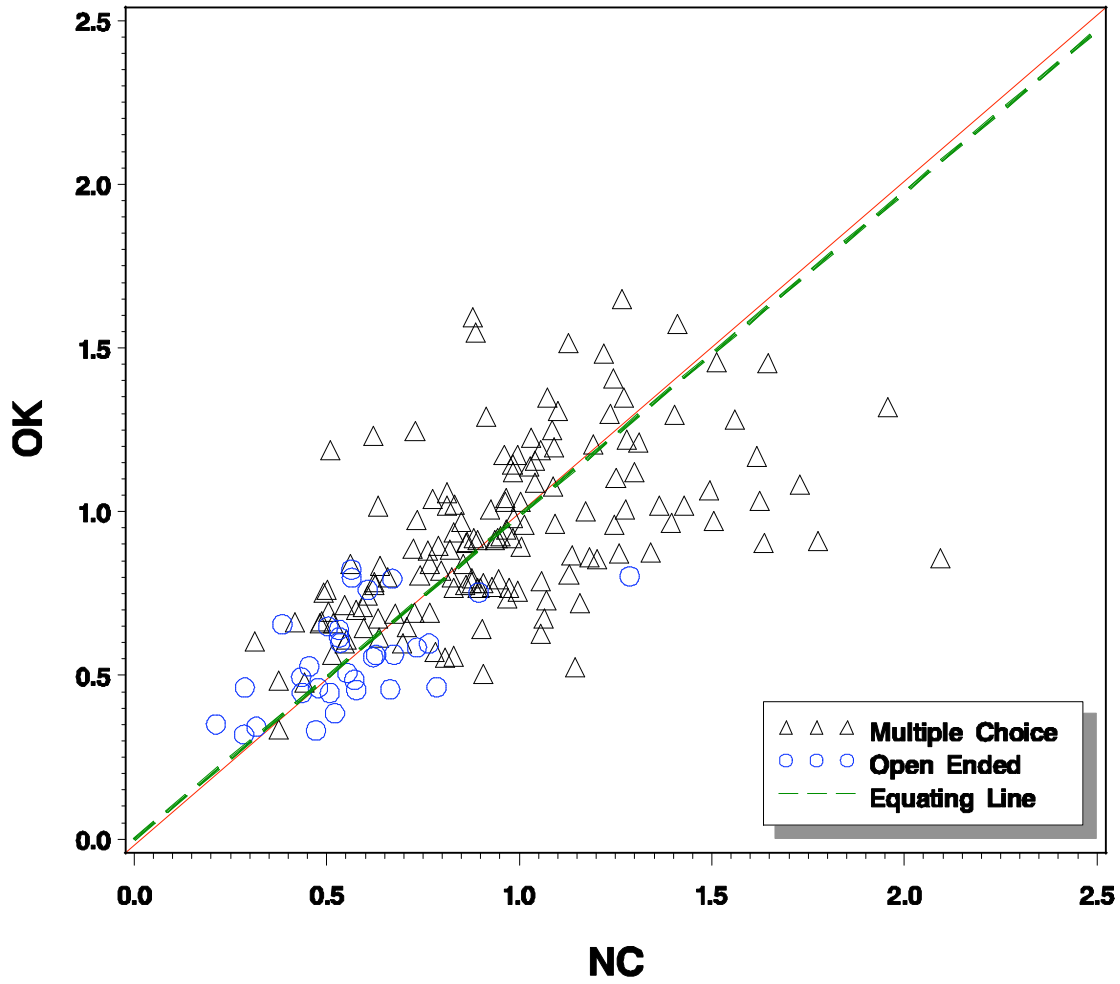
2005 NAEP Math Gr 8 b—plot: OK vs NC



Continues next page

Figure A-2. 2005 NAEP Math Gr 8 a- and b- plots: Selected States vs Selected States
(Continued)

2005 NAEP Math Gr 8 a – plot: OK vs NC

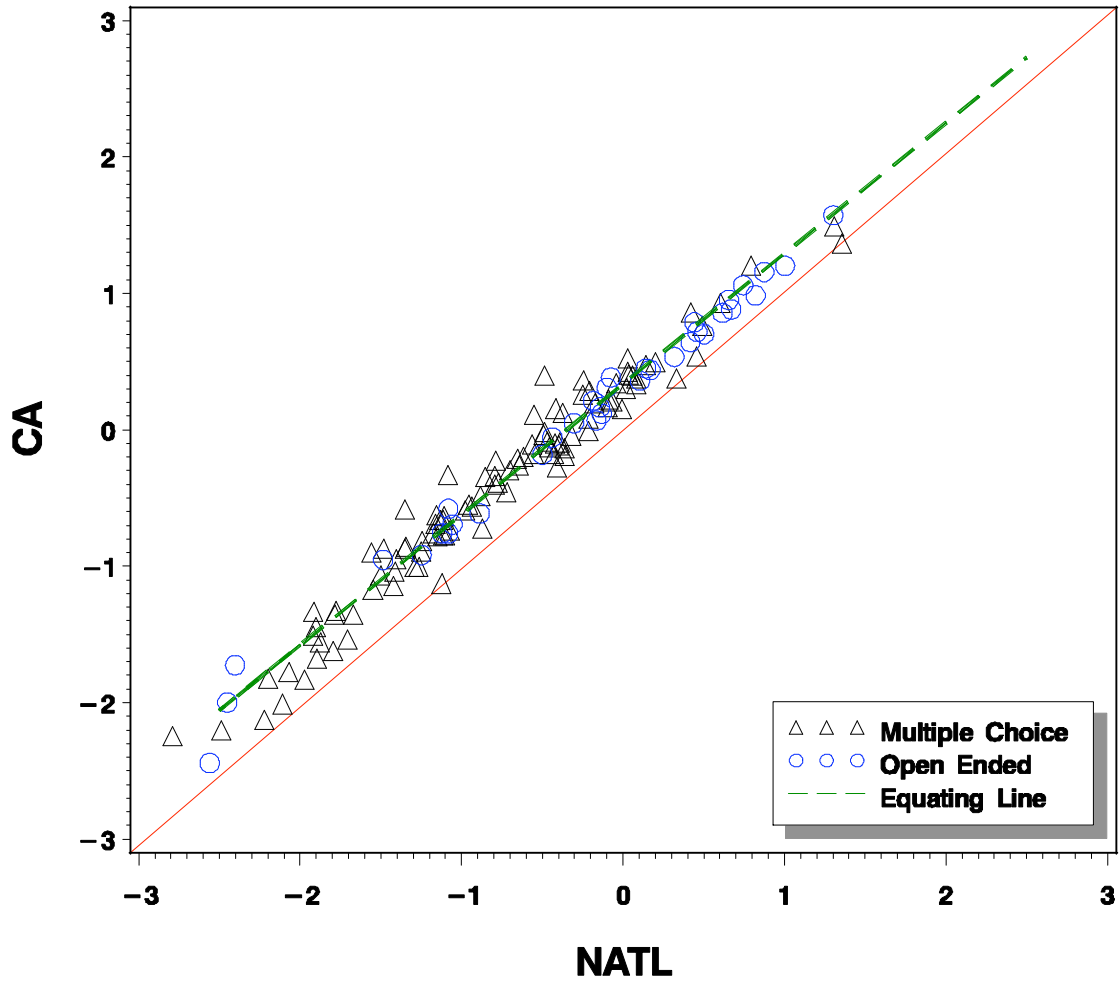


This page intentionally left blank

Appendix B: *a*- and *b*-plots for 2005 and 2003, Grade 8, Reading Assessment

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National

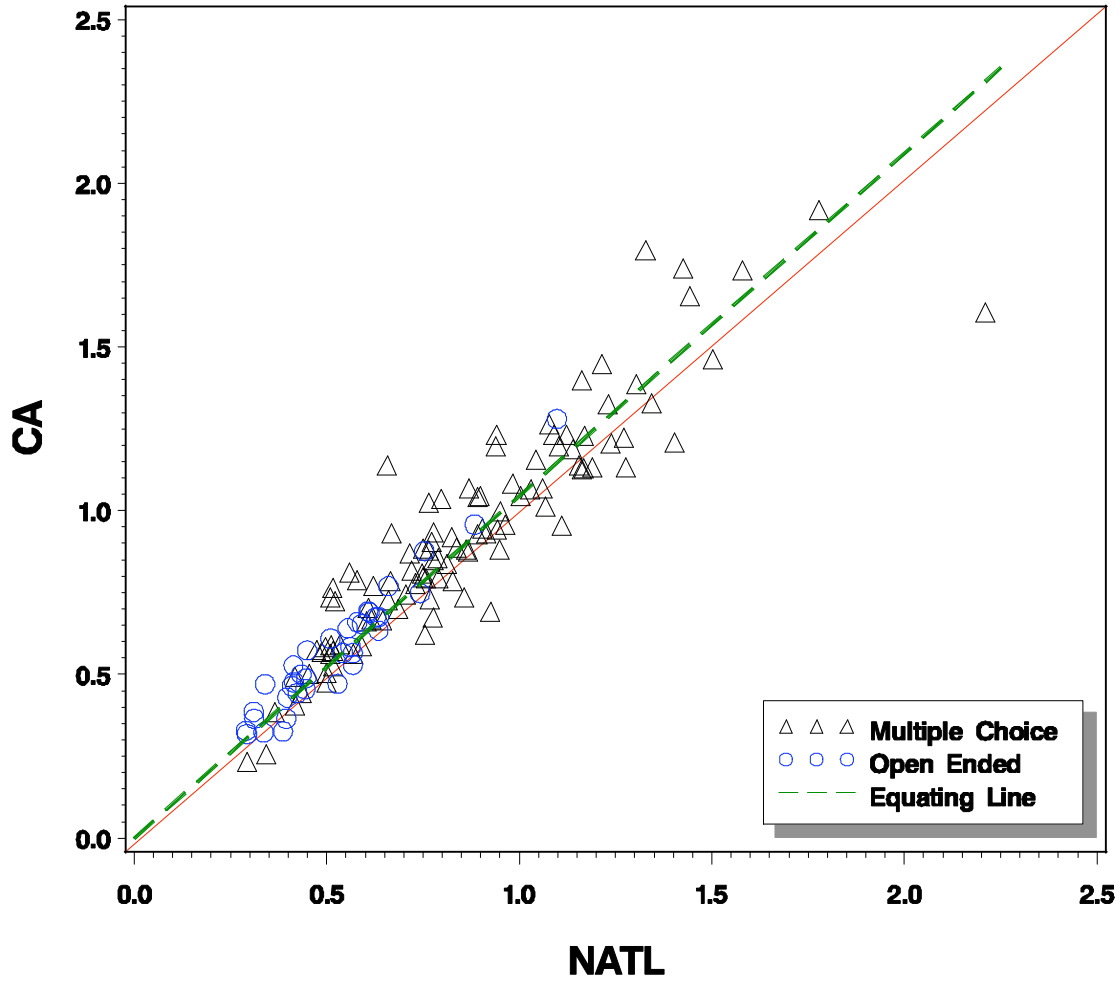
2005 NAEP Rdng Gr 8 b—plot: CA vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

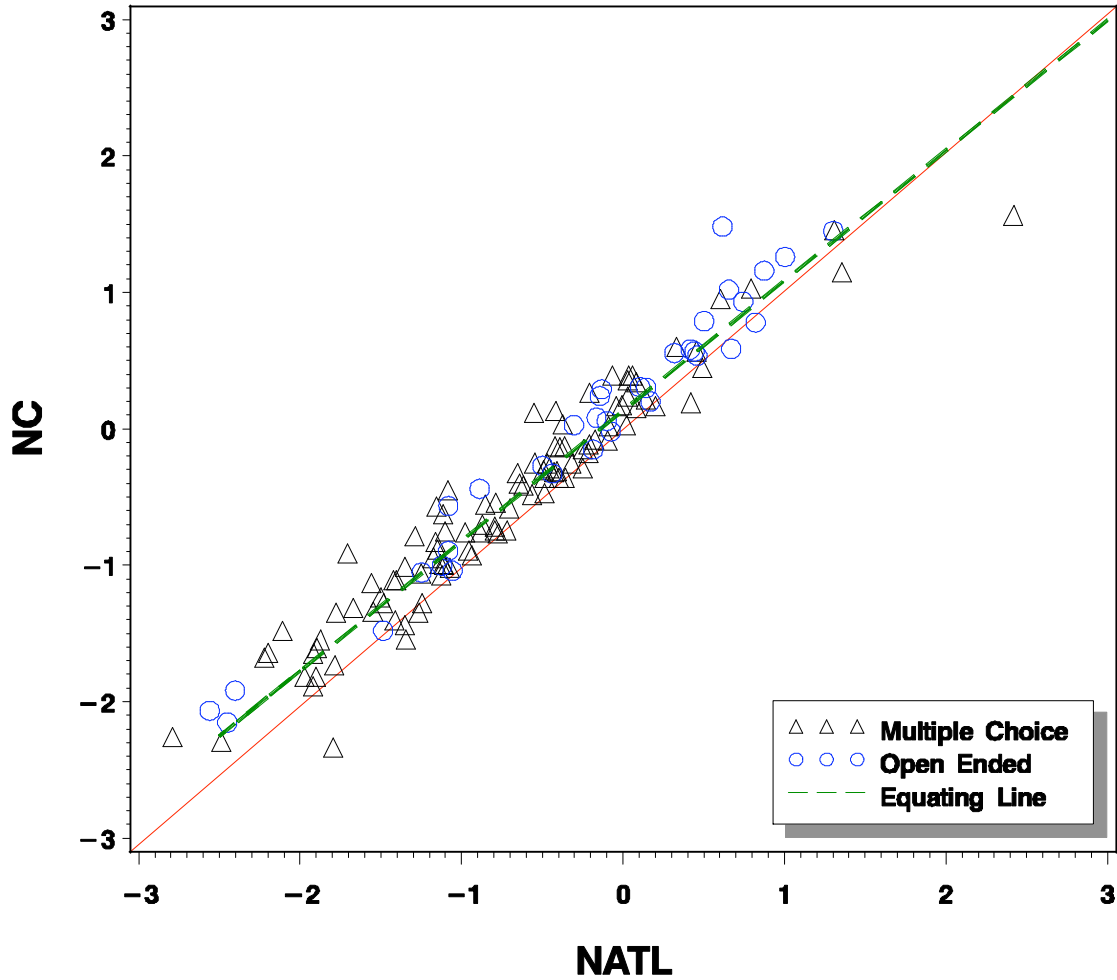
2005 NAEP Rdng Gr 8 a – plot: CA vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

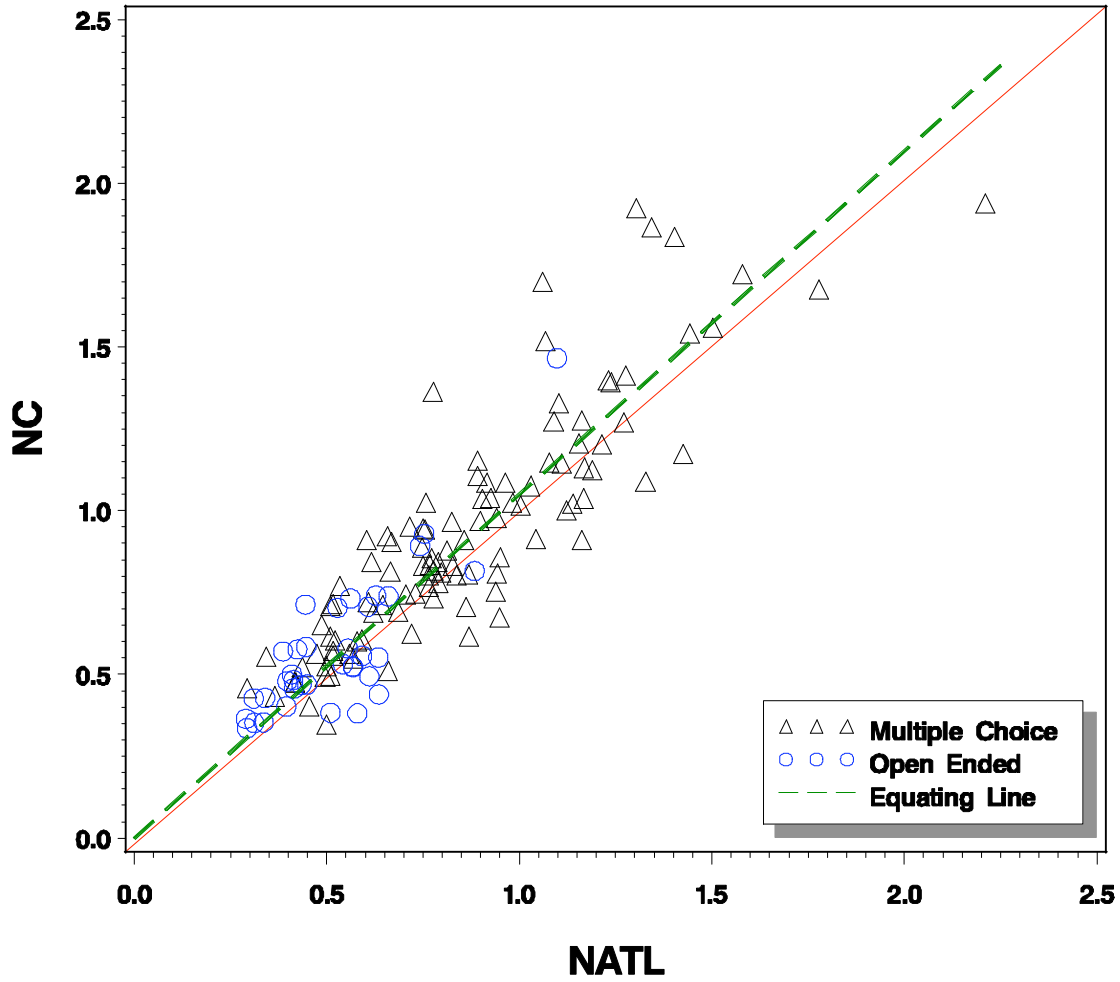
2005 NAEP Rdng Gr 8 b–plot: NC vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

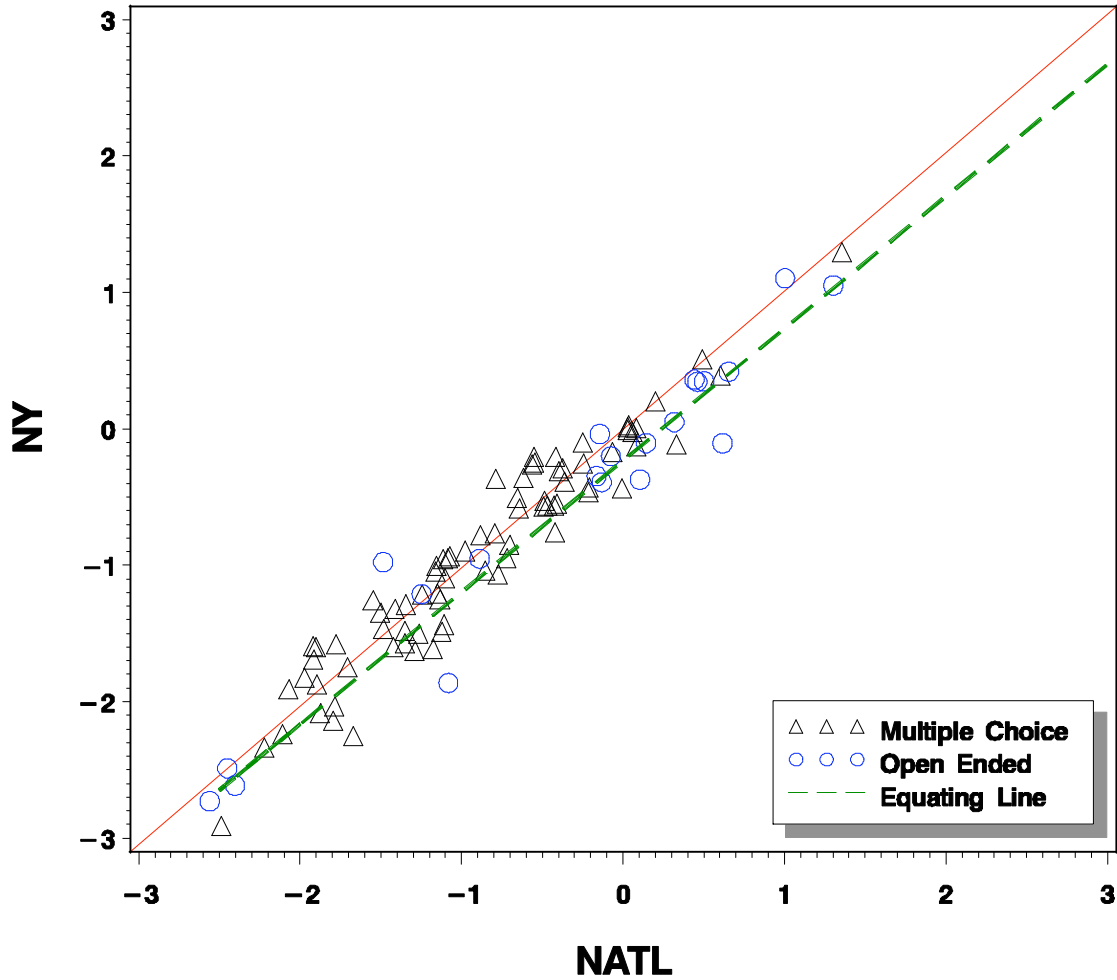
2005 NAEP Rdng Gr 8 a–plot: NC vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

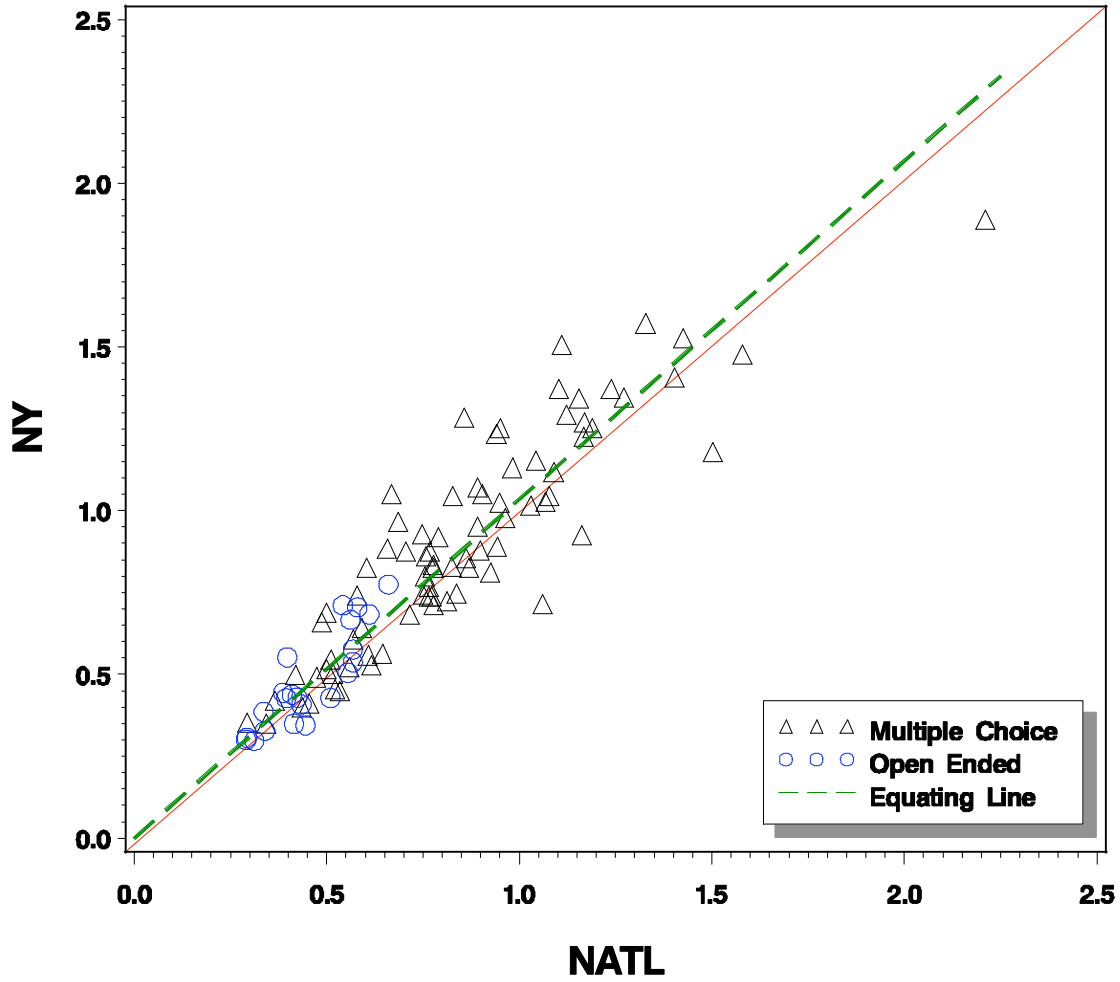
2005 NAEP Rdng Gr 8 b—plot: NY vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

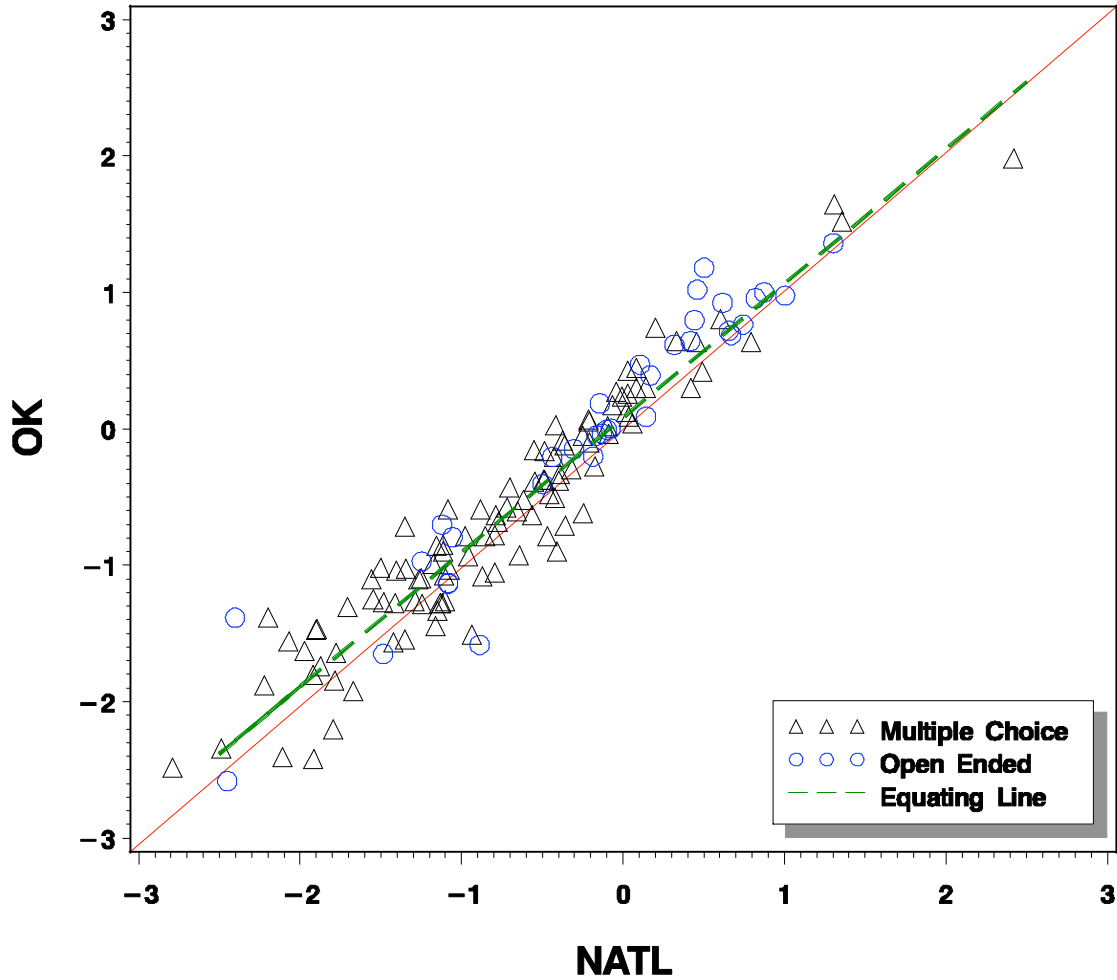
2005 NAEP Rdng Gr 8 a–plot: NY vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

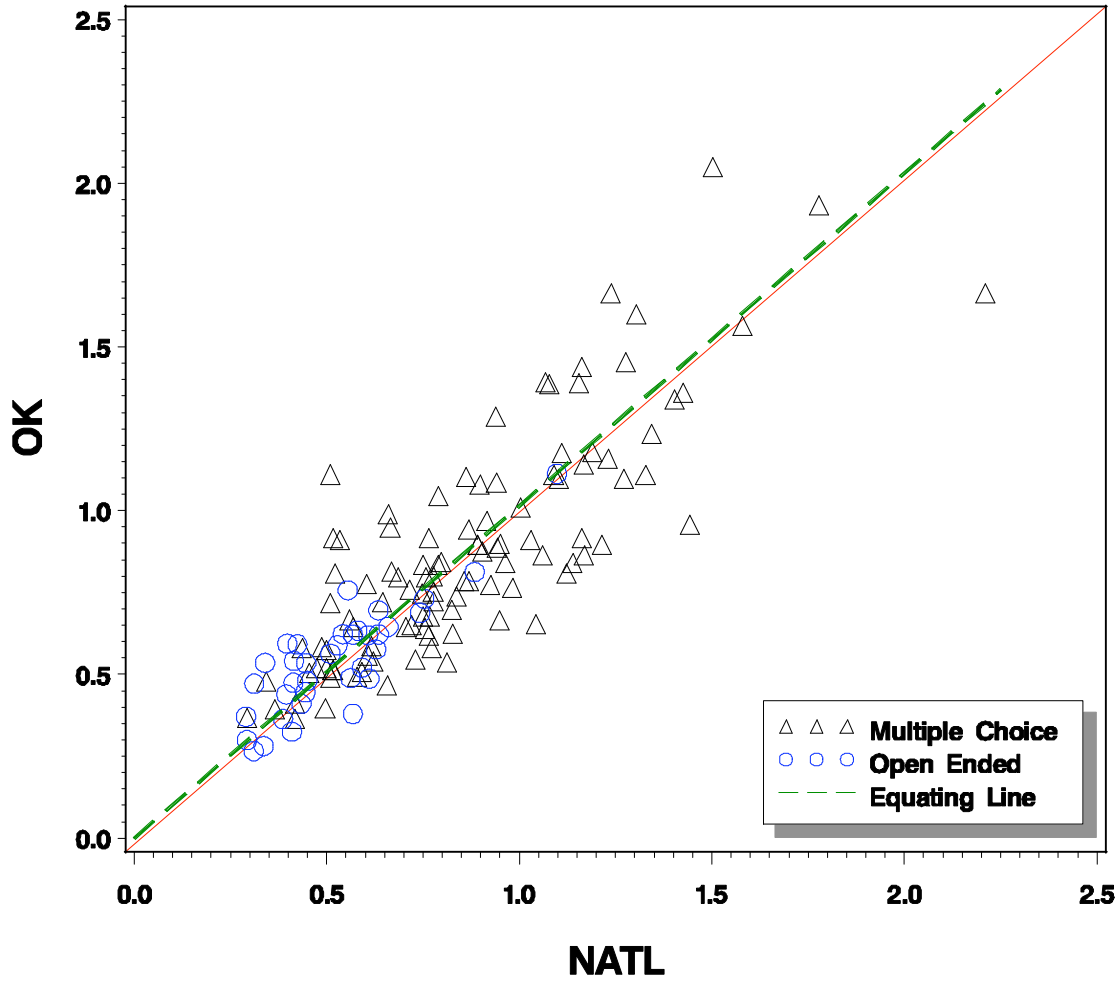
2005 NAEP Rdng Gr 8 b–plot: OK vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

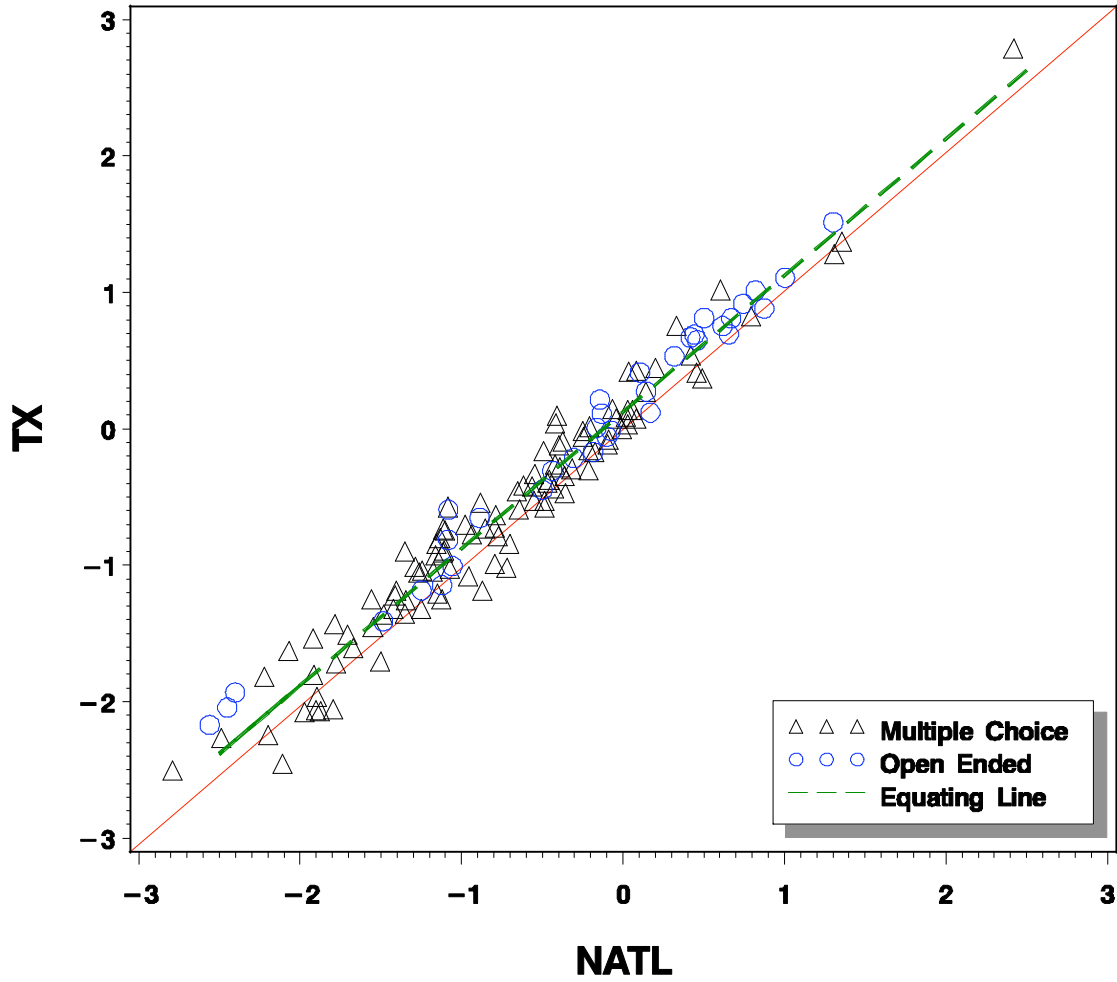
2005 NAEP Rdng Gr 8 a–plot: OK vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

2005 NAEP Rdng Gr 8 b–plot: TX vs NATL



Continues next page

Figure B-1. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

2005 NAEP Rdng Gr 8 a – plot: TX vs NATL

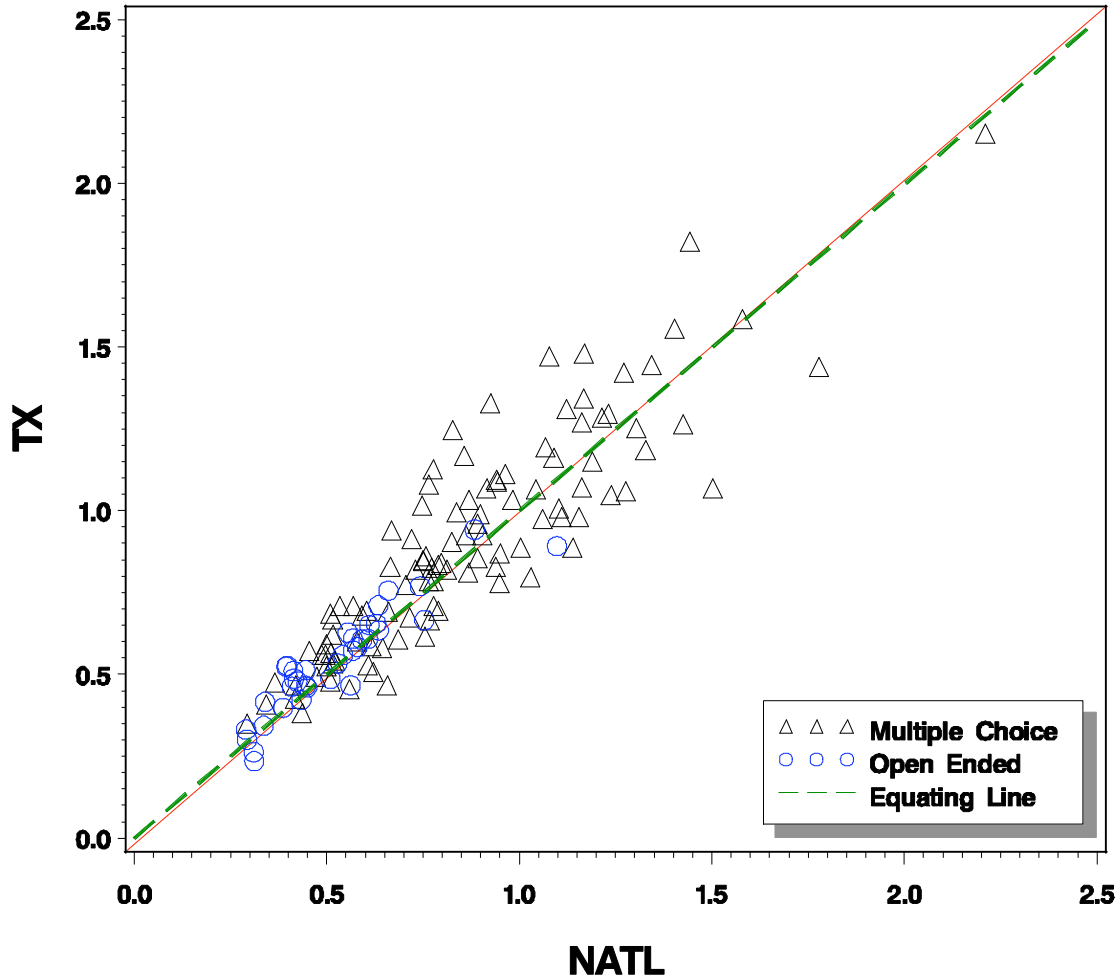
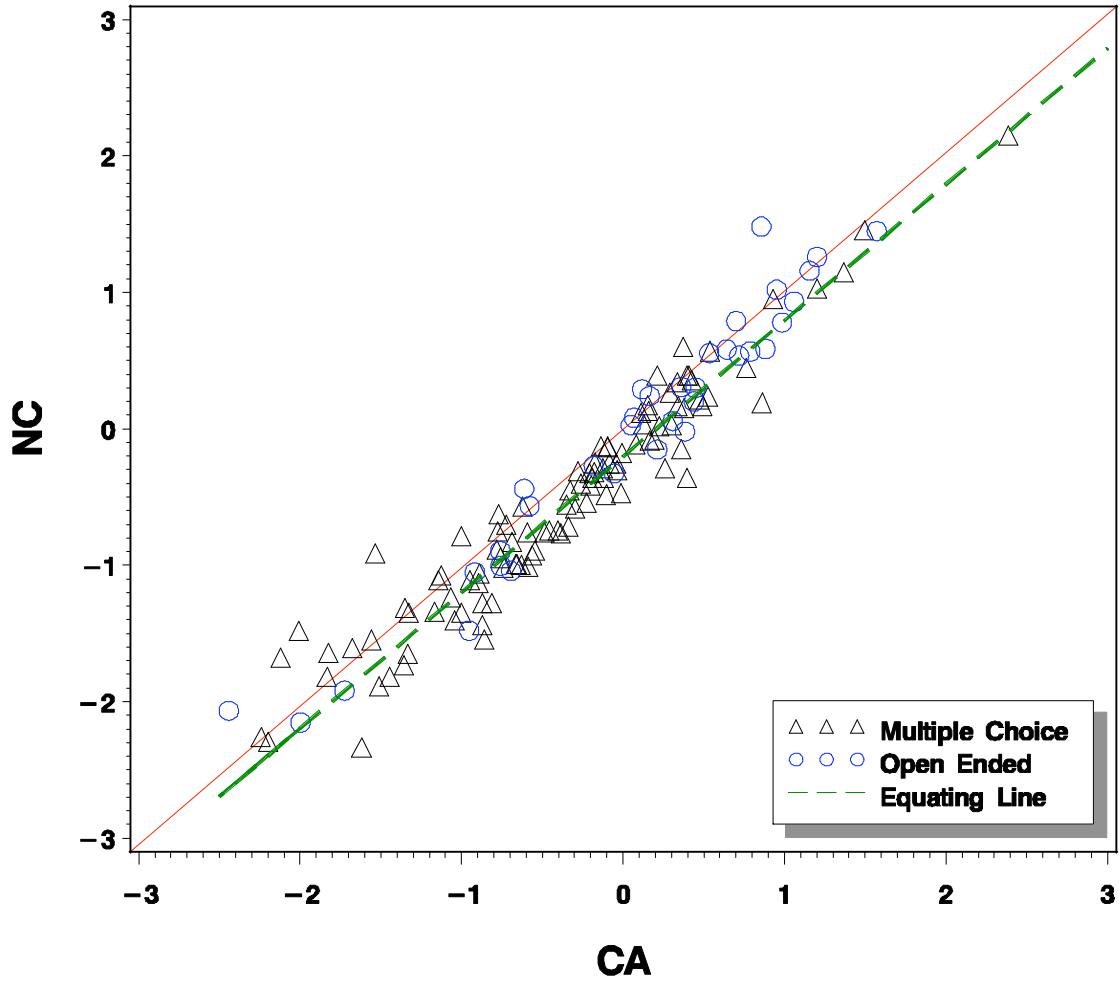


Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States

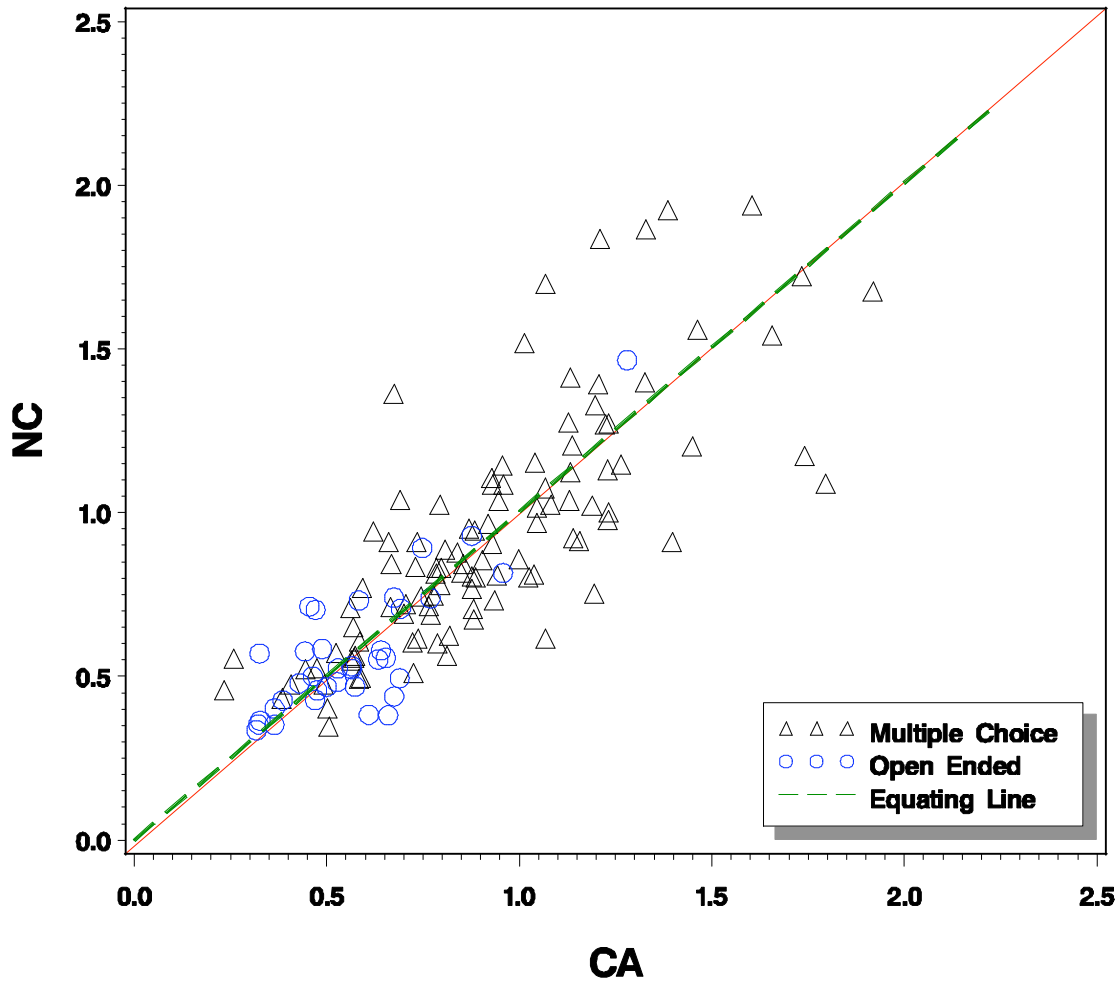
2005 NAEP Rdng Gr 8 b–plot: NC vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

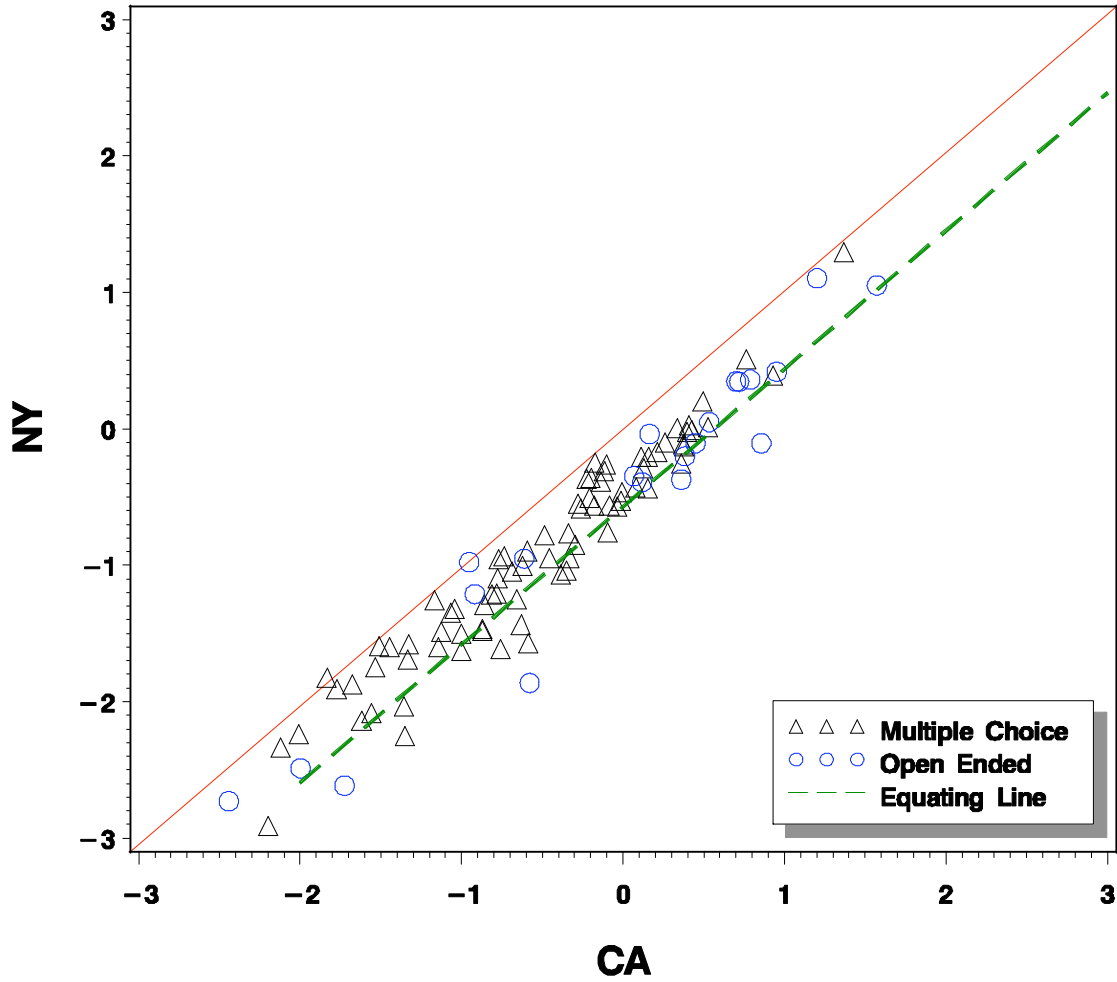
2005 NAEP Rdng Gr 8 a – plot: NC vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

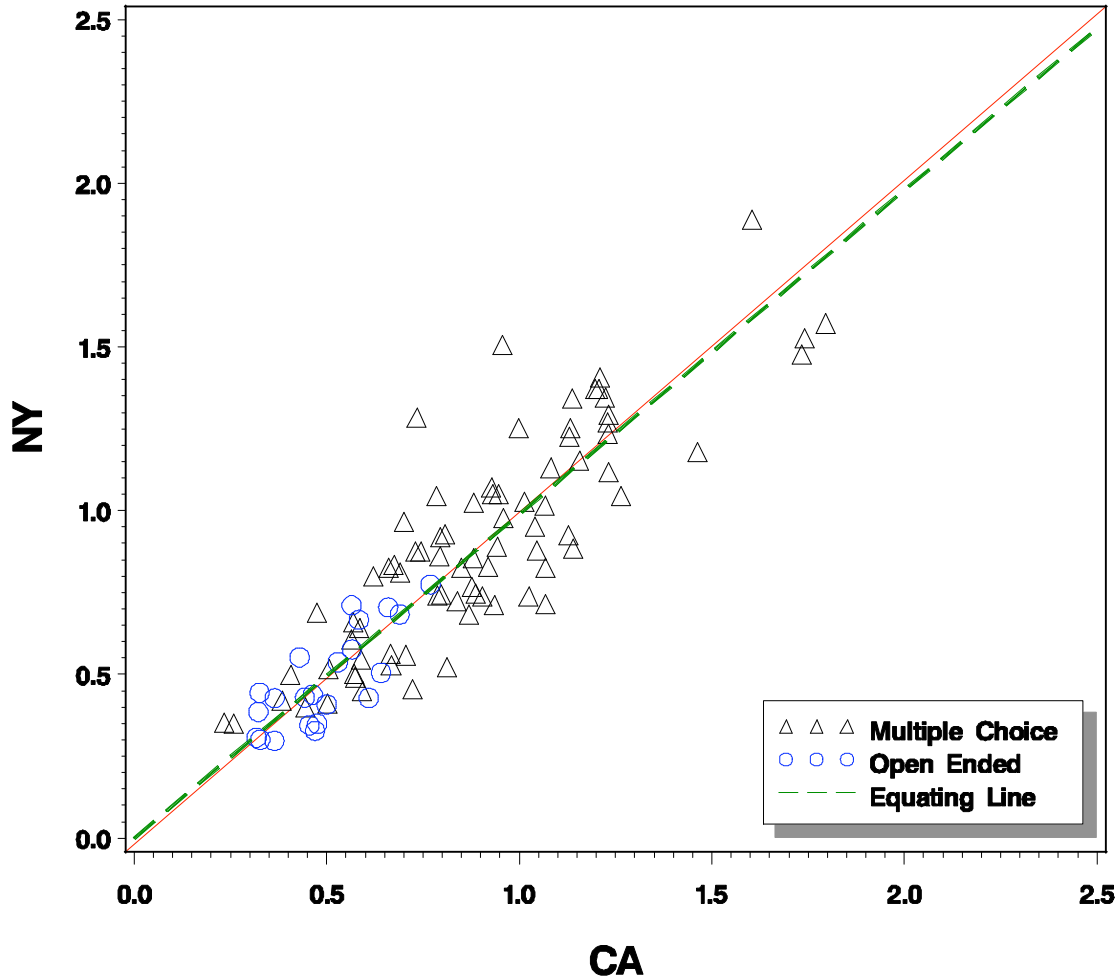
2005 NAEP Rdng Gr 8 b—plot: NY vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

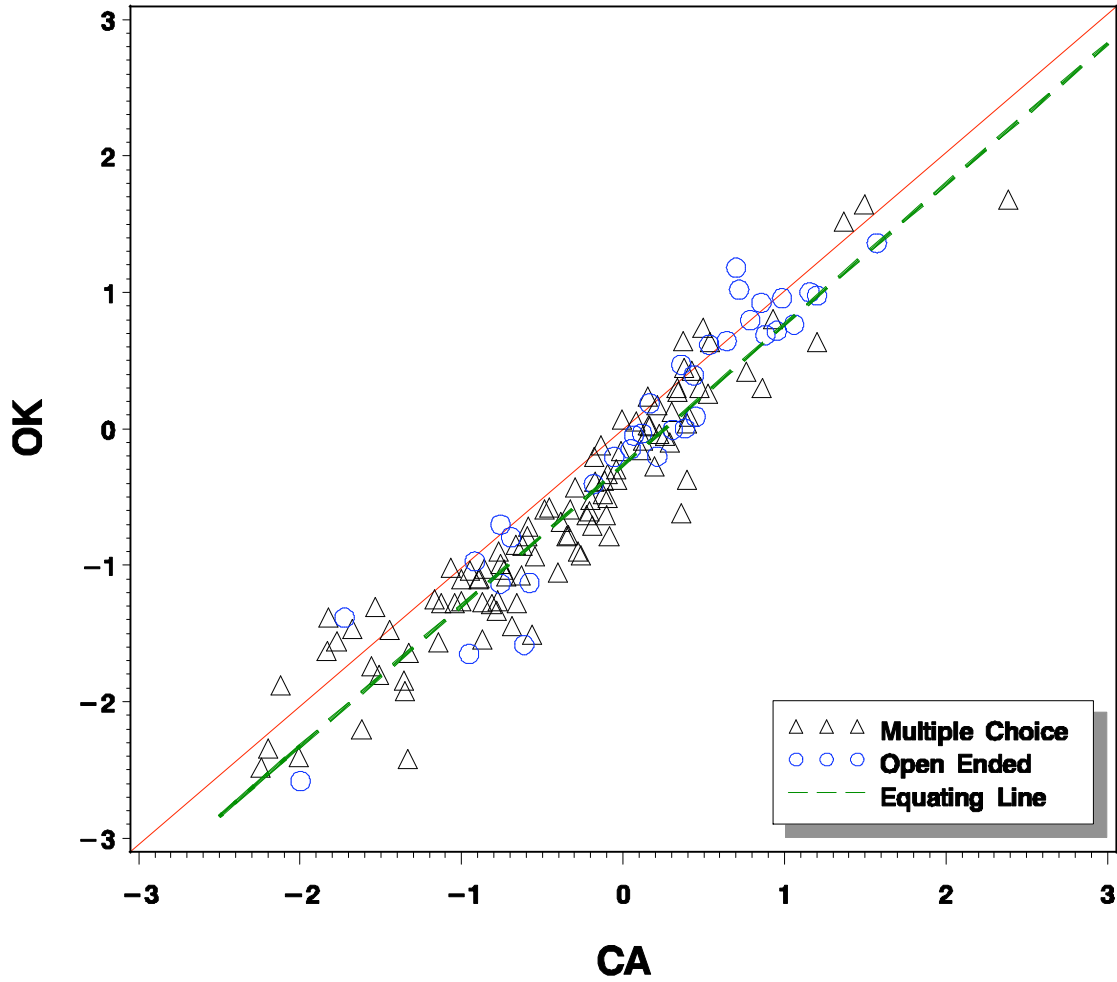
2005 NAEP Rdng Gr 8 a–plot: NY vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

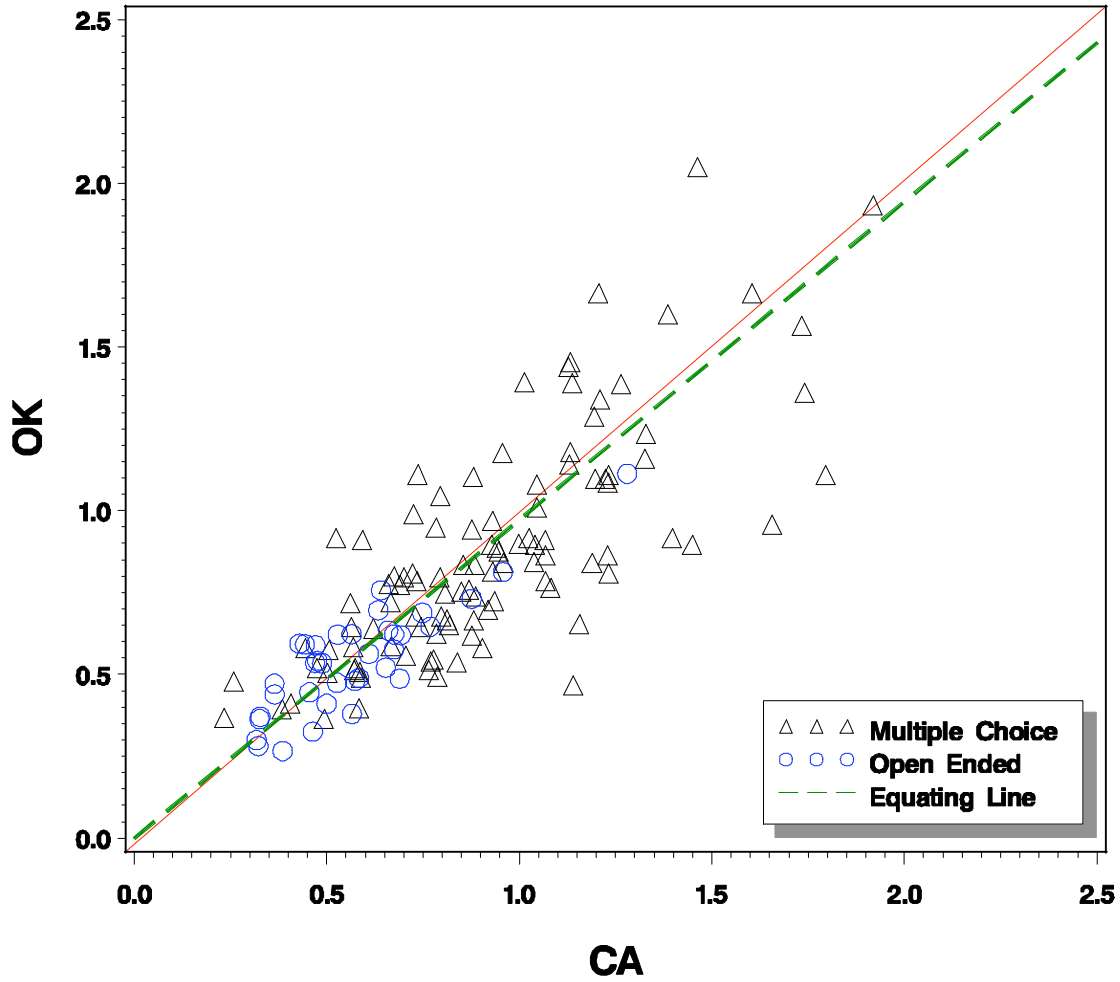
2005 NAEP Rdng Gr 8 b–plot: OK vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

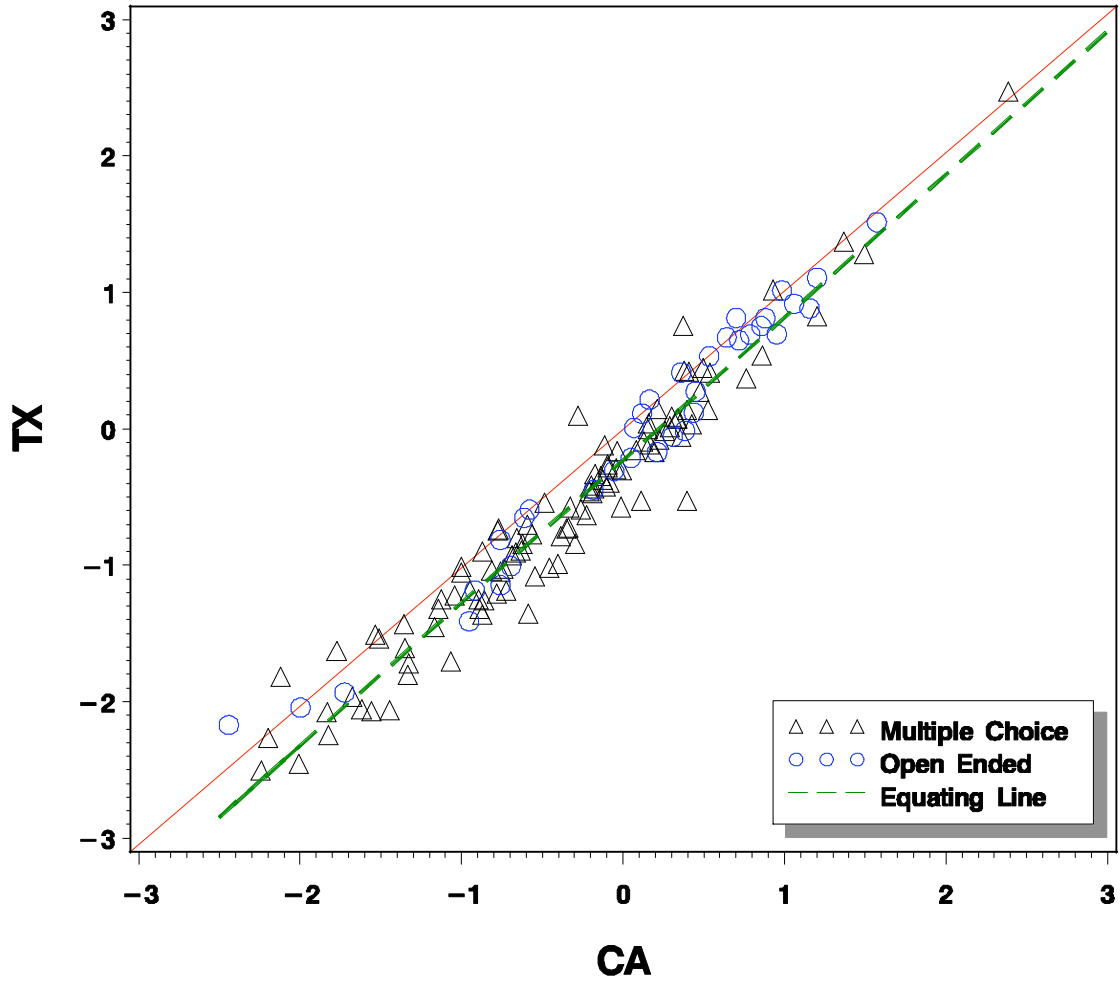
2005 NAEP Rdng Gr 8 a – plot: OK vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

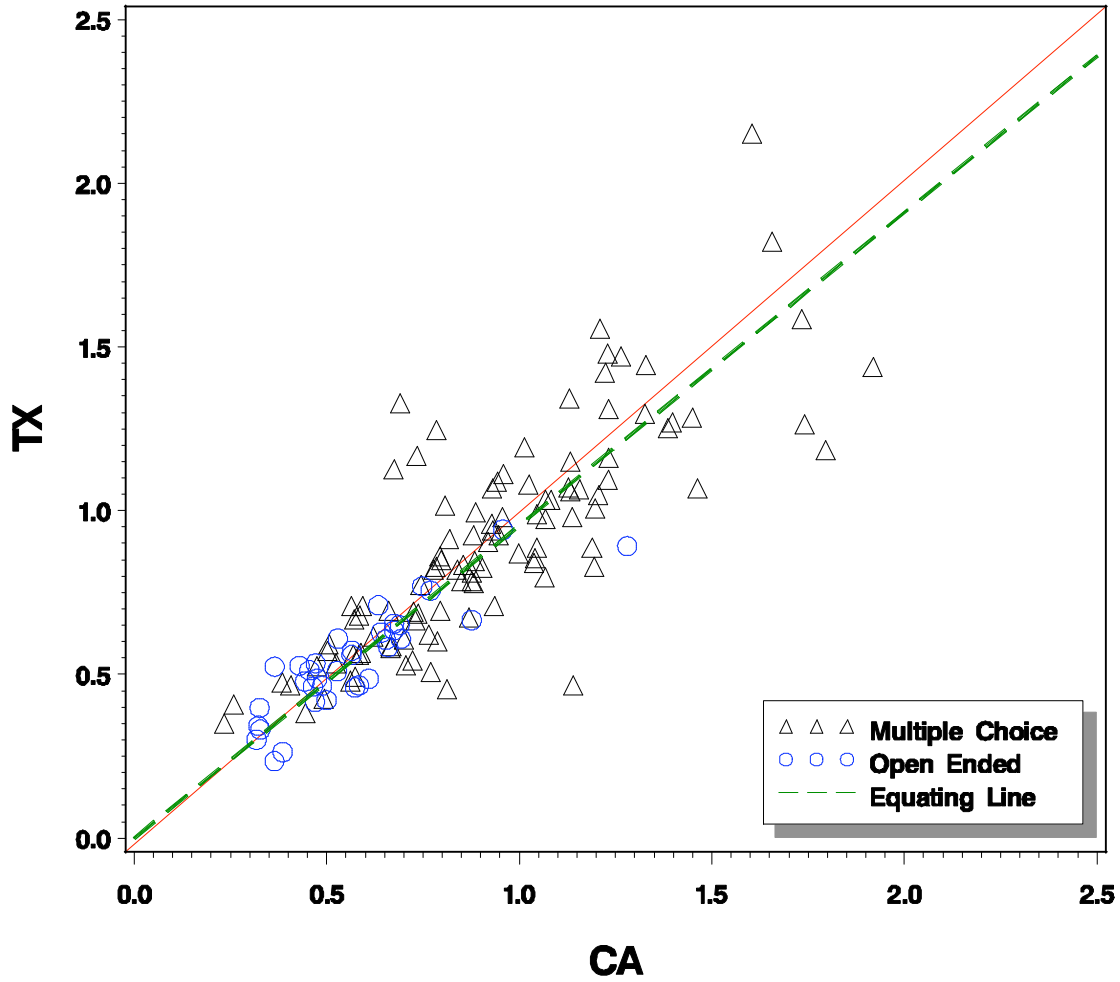
2005 NAEP Rdng Gr 8 b—plot: TX vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

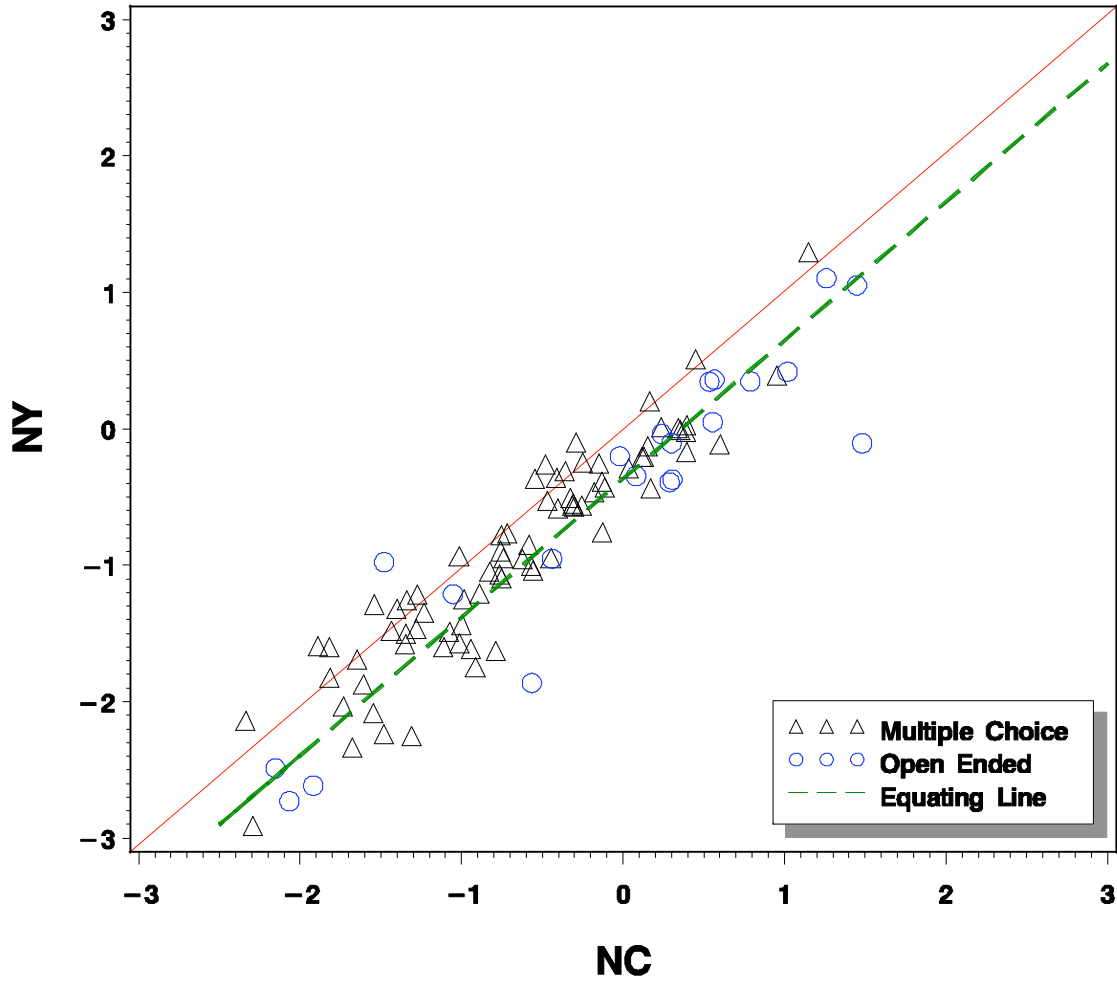
2005 NAEP Rdng Gr 8 a–plot: TX vs CA



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

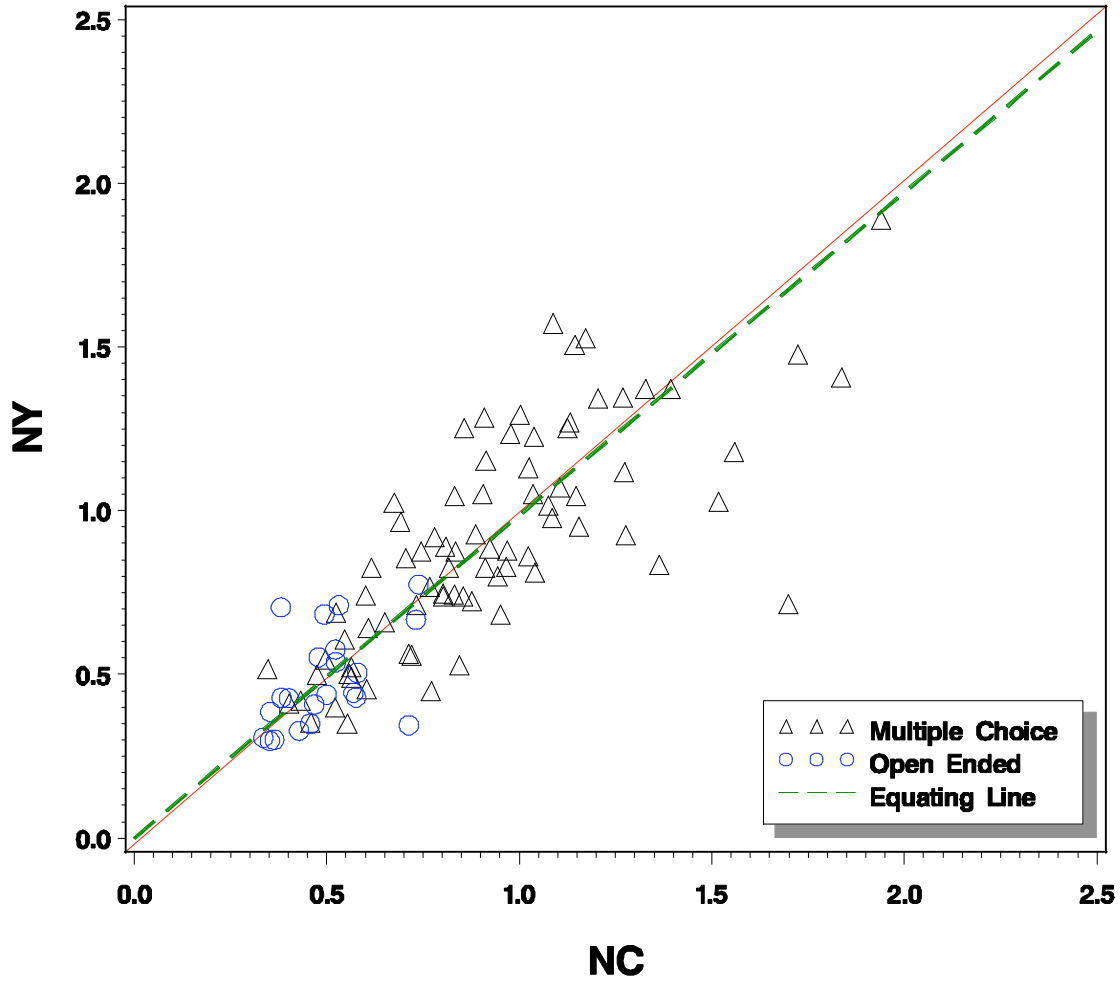
2005 NAEP Rdng Gr 8 b–plot: NY vs NC



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

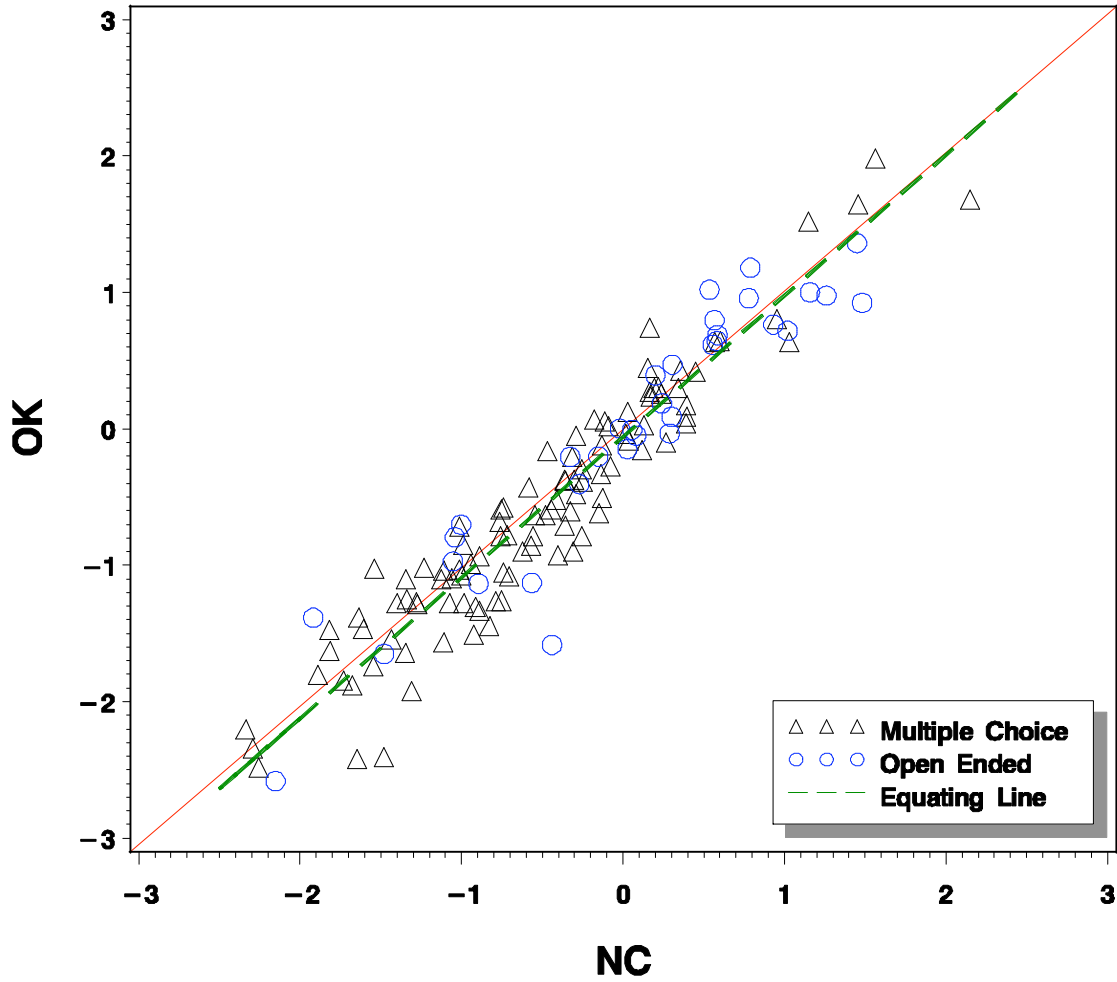
2005 NAEP Rdng Gr 8 a–plot: NY vs NC



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

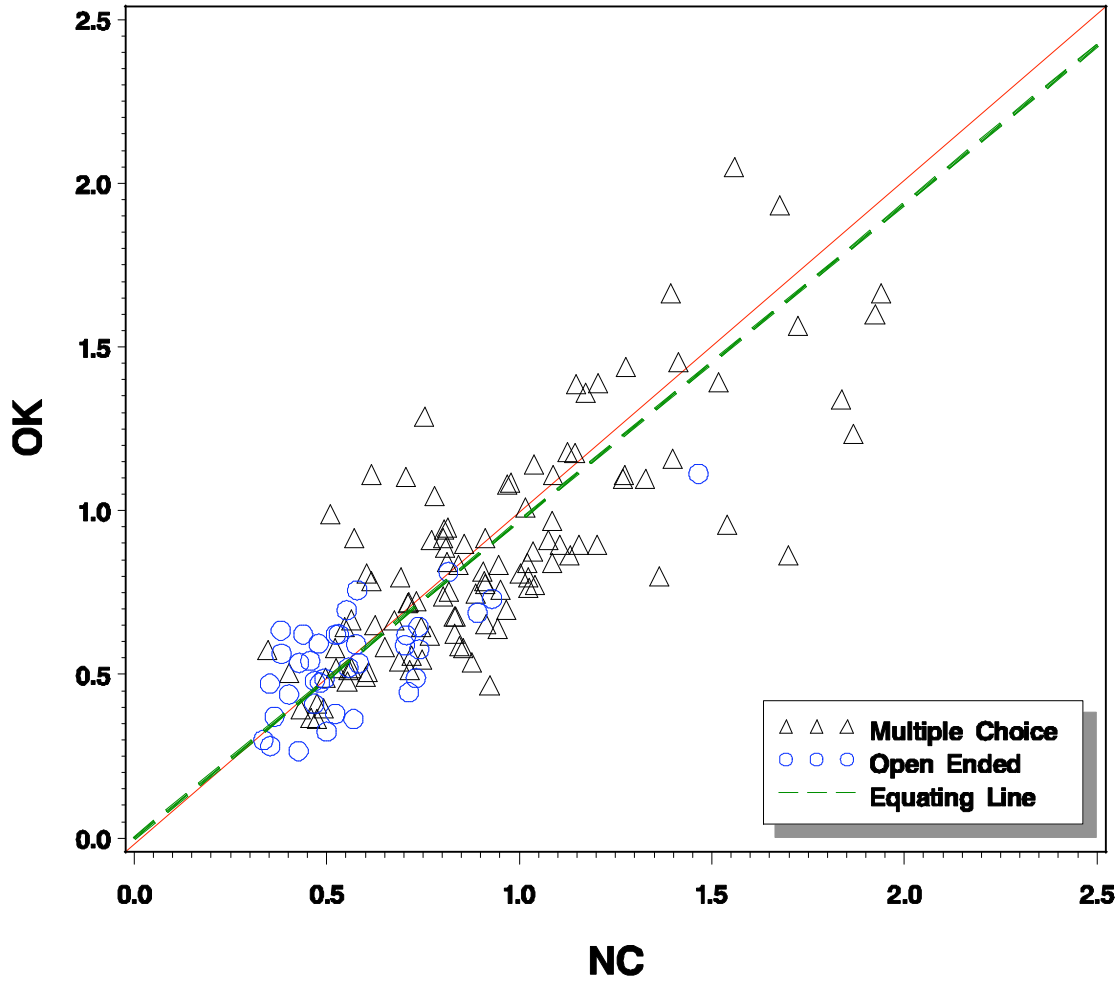
2005 NAEP Rdng Gr 8 b—plot: OK vs NC



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

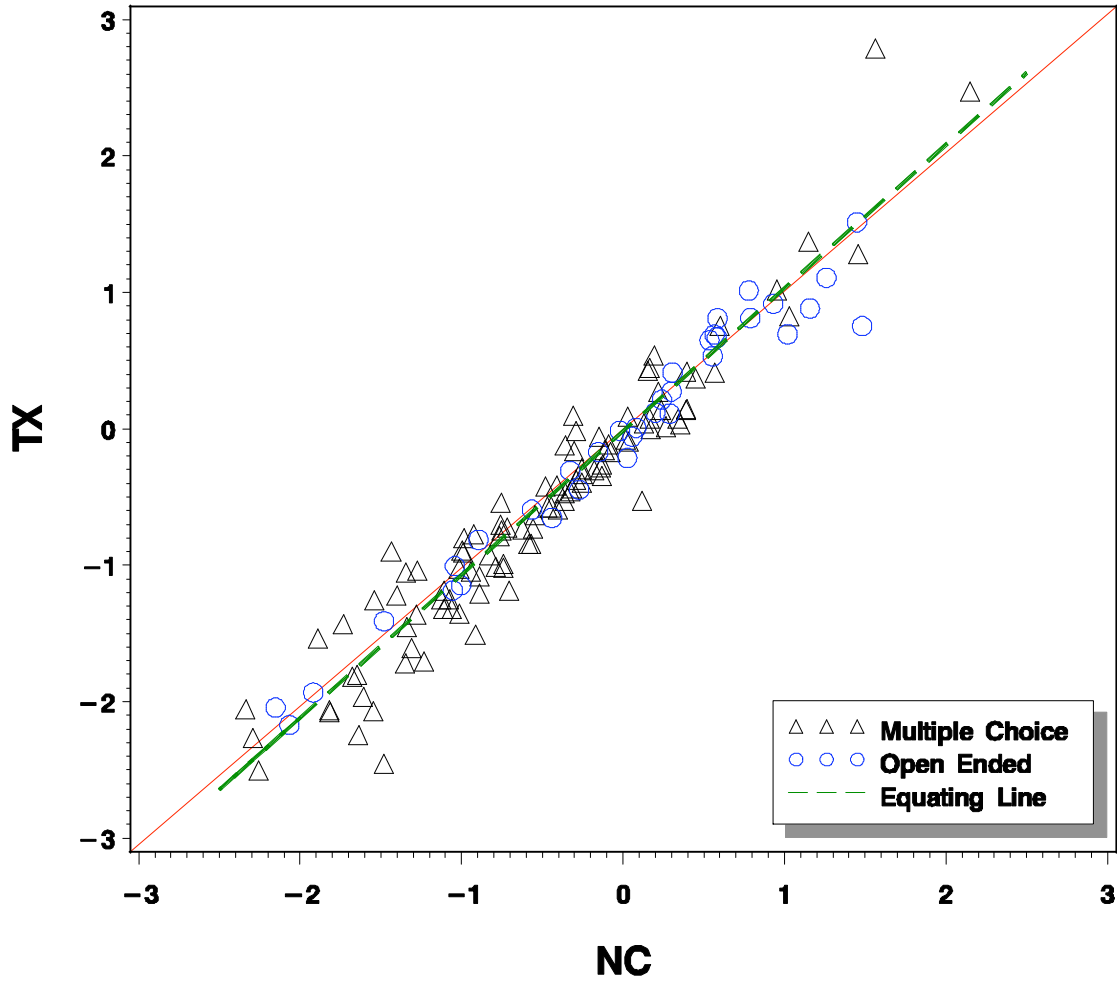
2005 NAEP Rdng Gr 8 a – plot: OK vs NC



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

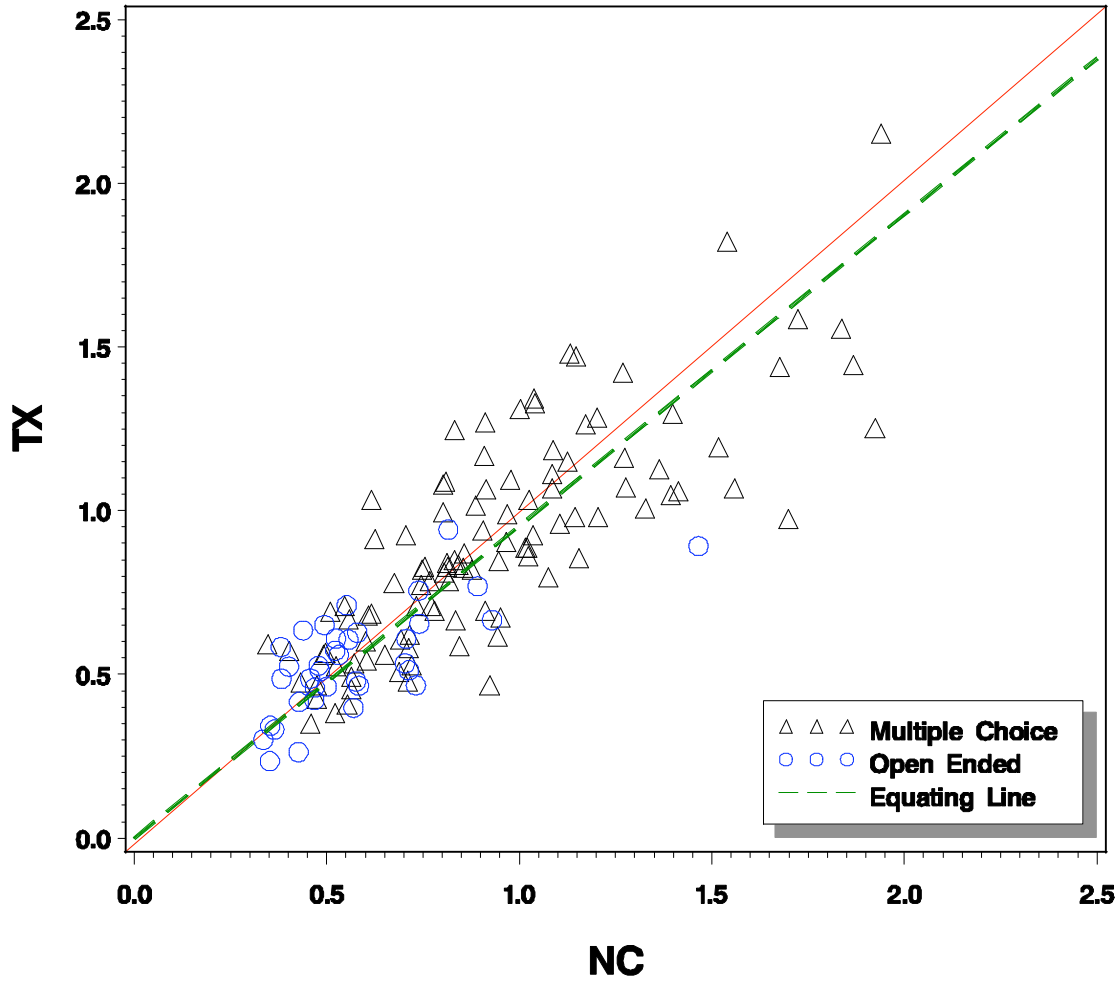
2005 NAEP Rdng Gr 8 b–plot: TX vs NC



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

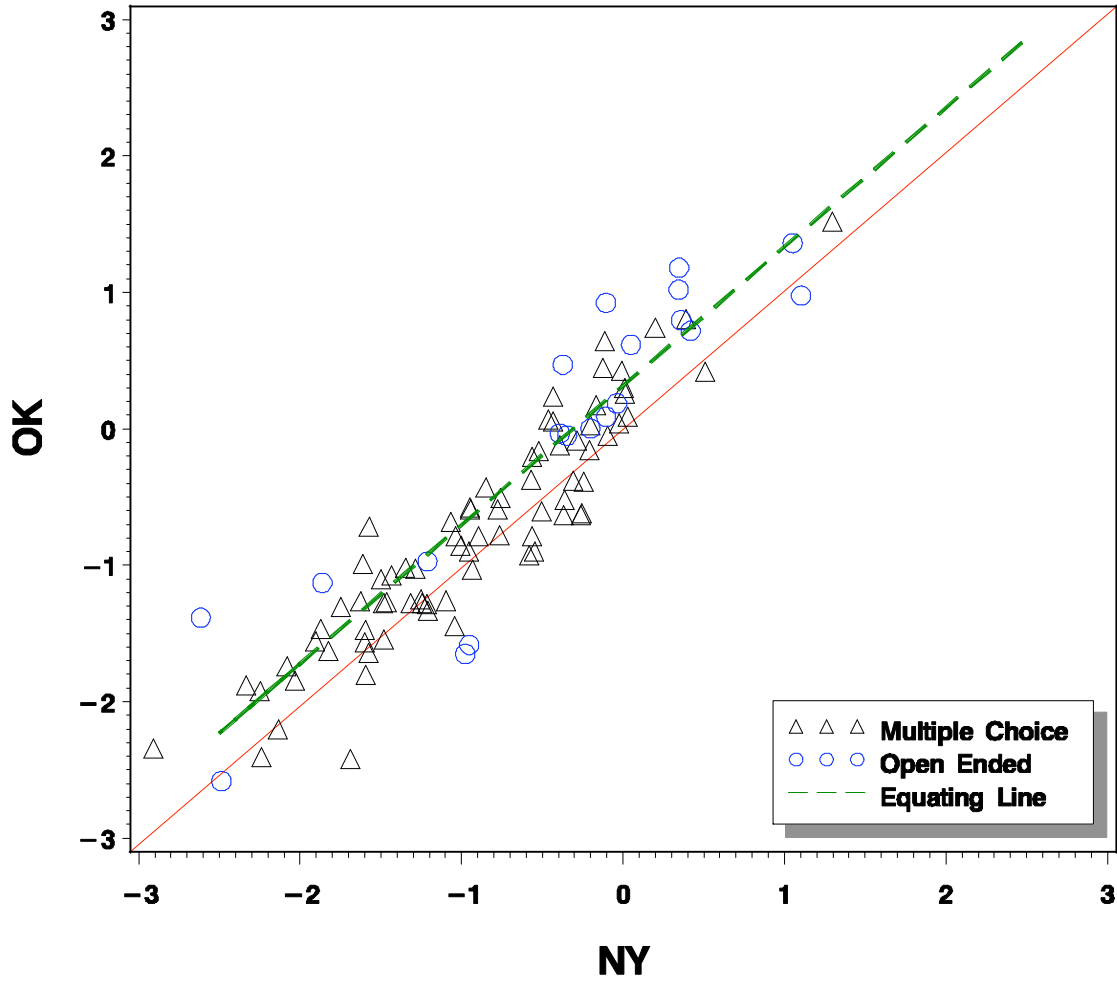
2005 NAEP Rdng Gr 8 a–plot: TX vs NC



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

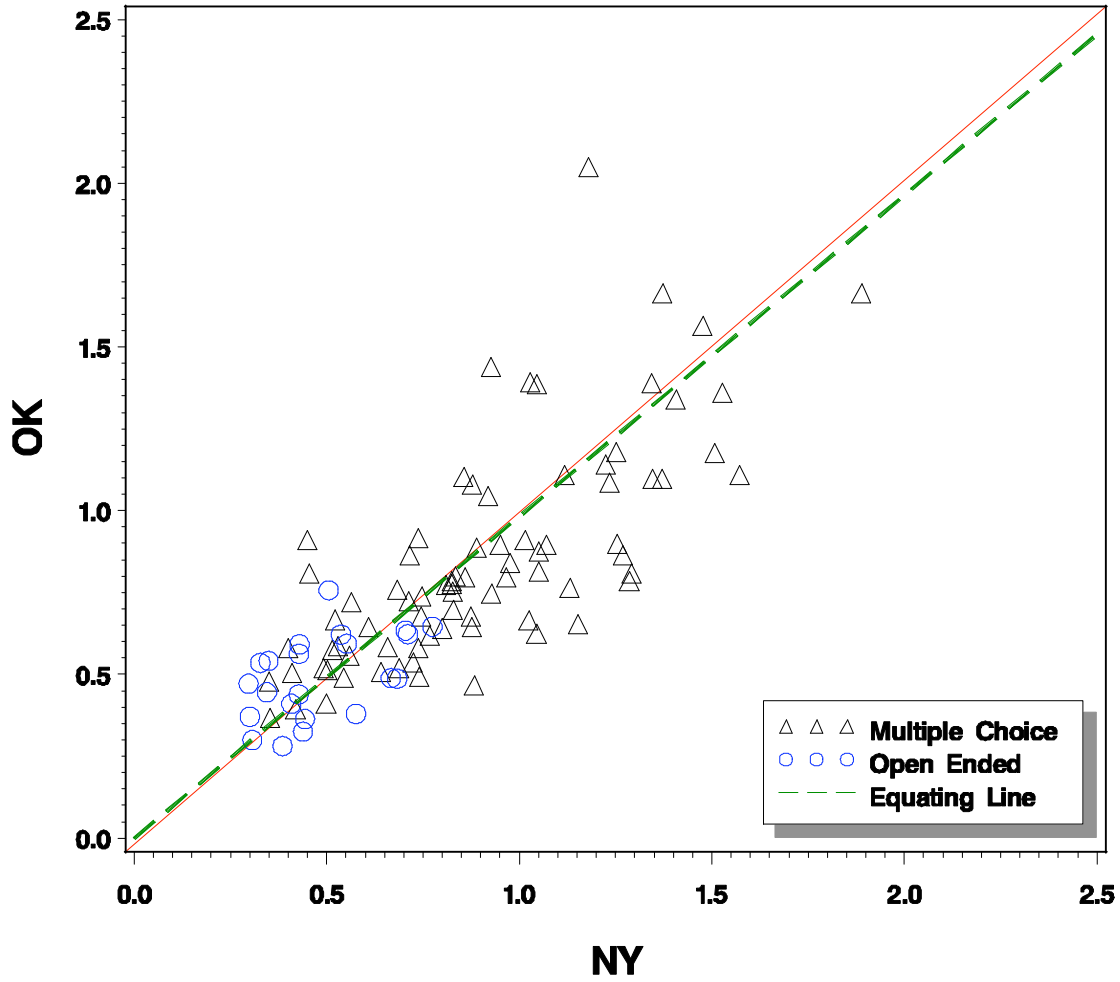
2005 NAEP Rdng Gr 8 b–plot: OK vs NY



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

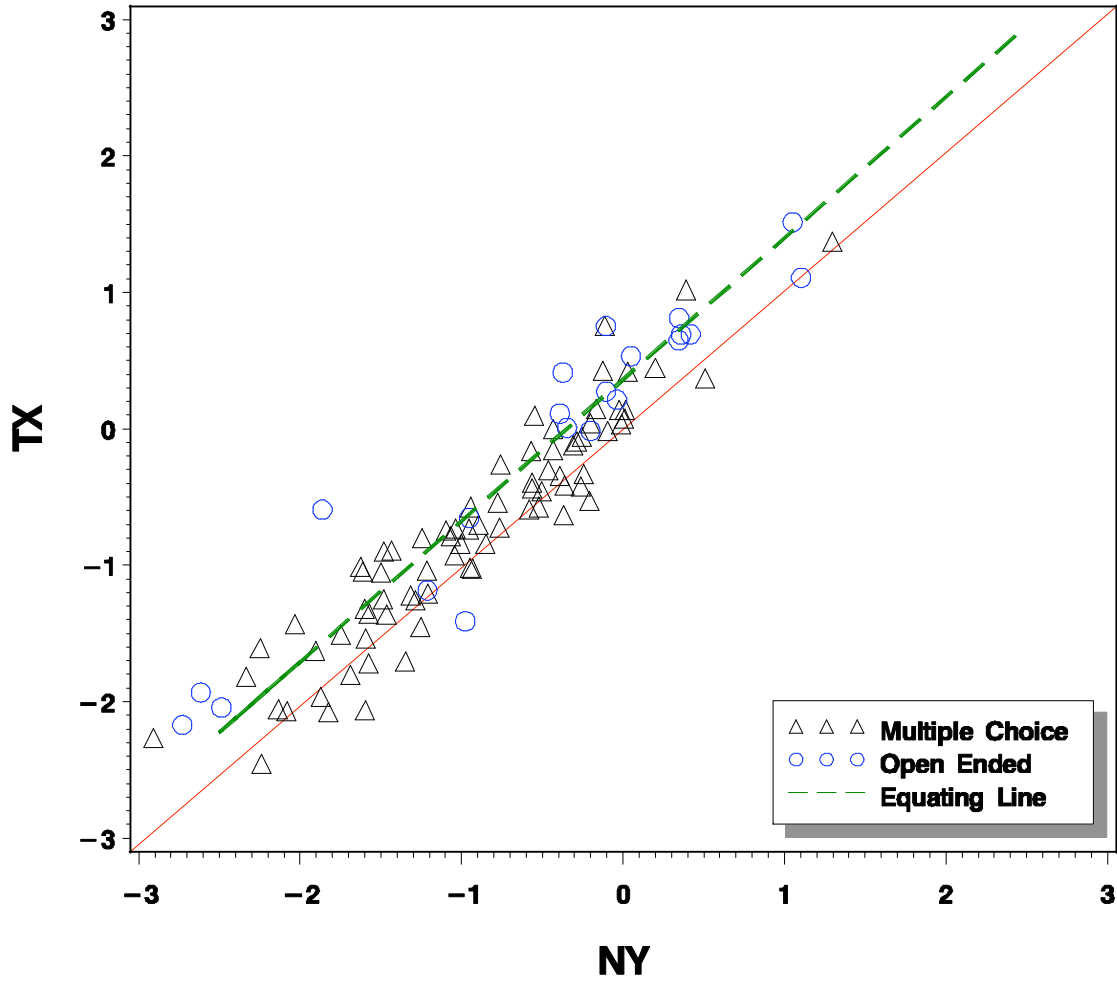
2005 NAEP Rdng Gr 8 a–plot: OK vs NY



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

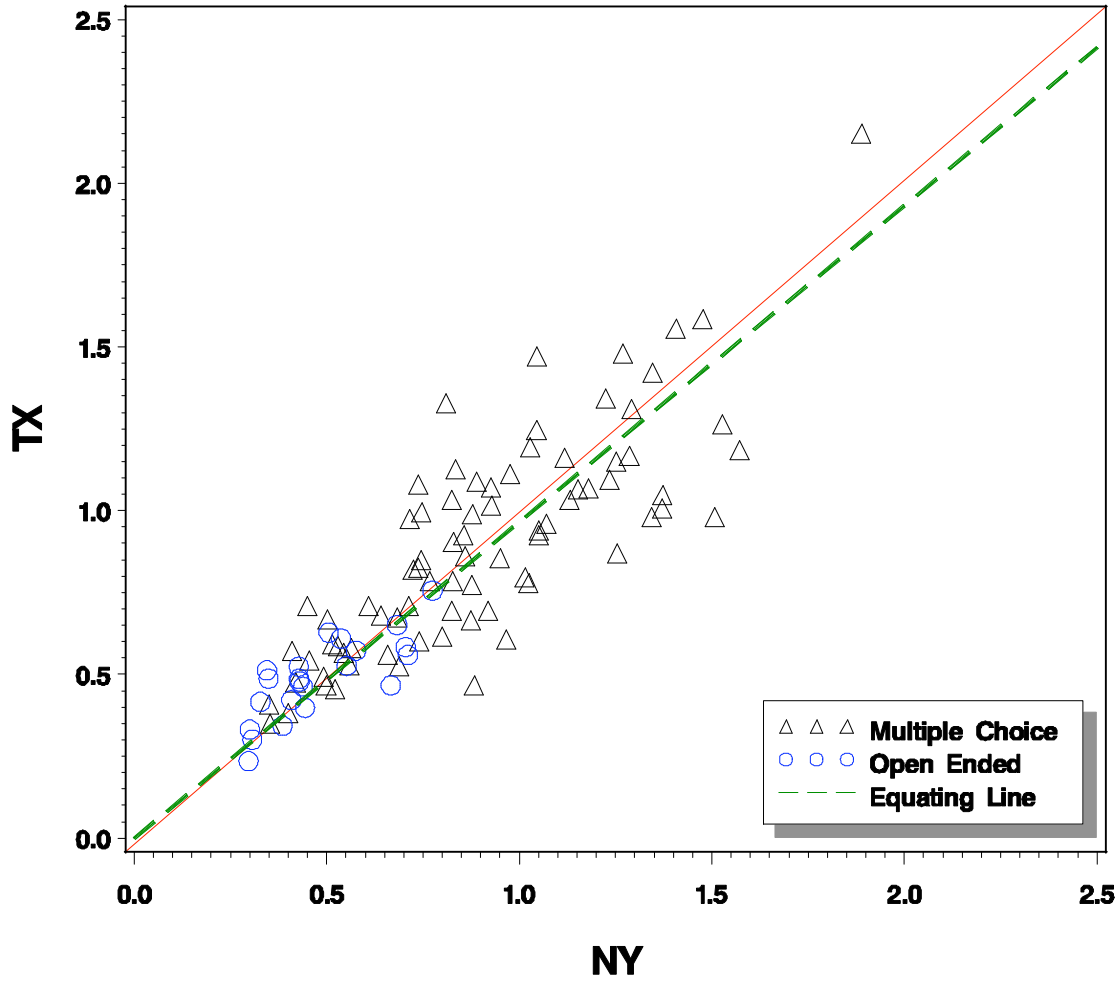
2005 NAEP Rdng Gr 8 b—plot: TX vs NY



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

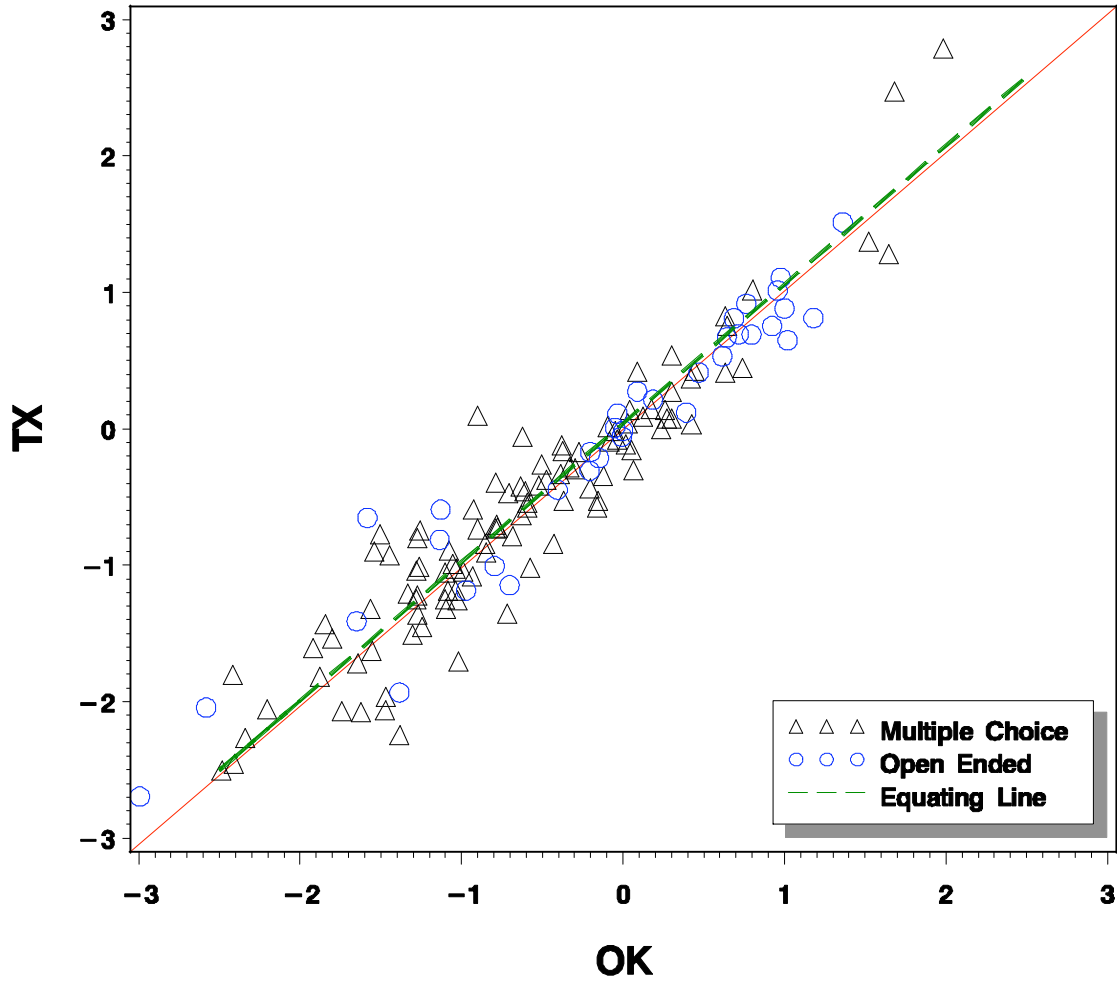
2005 NAEP Rdng Gr 8 a–plot: TX vs NY



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

2005 NAEP Rdng Gr 8 b–plot: TX vs OK



Continues next page

Figure B-2. 2005 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

2005 NAEP Rdng Gr 8 a–plot: TX vs OK

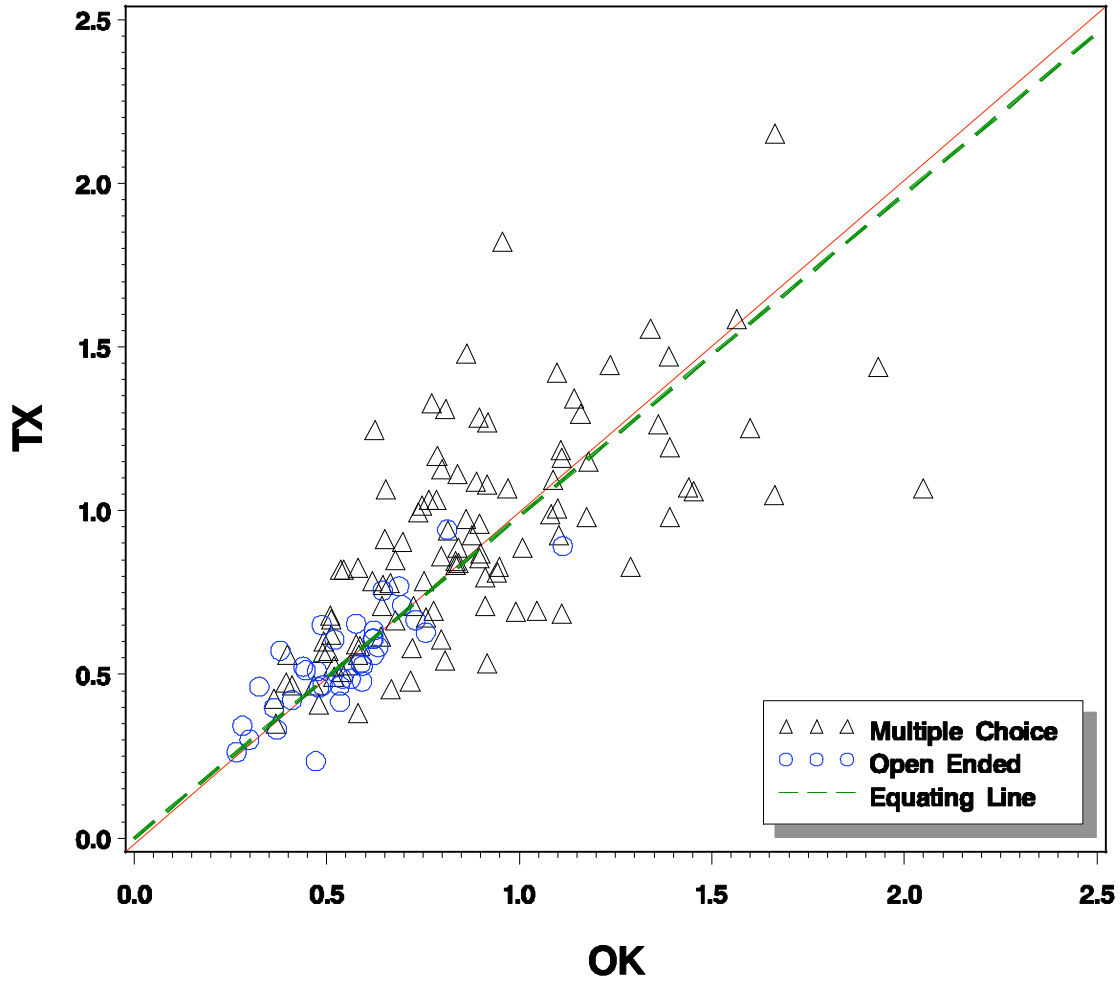
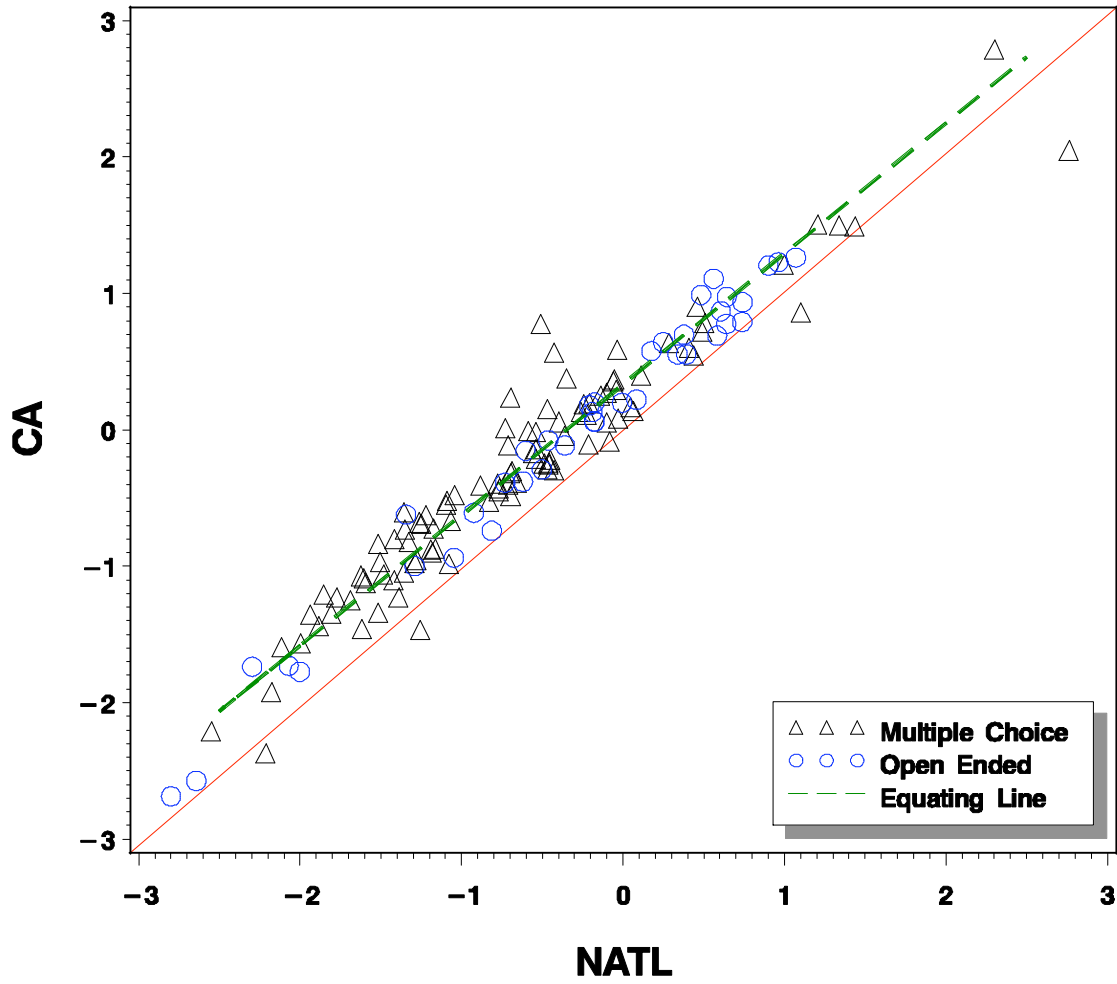


Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National

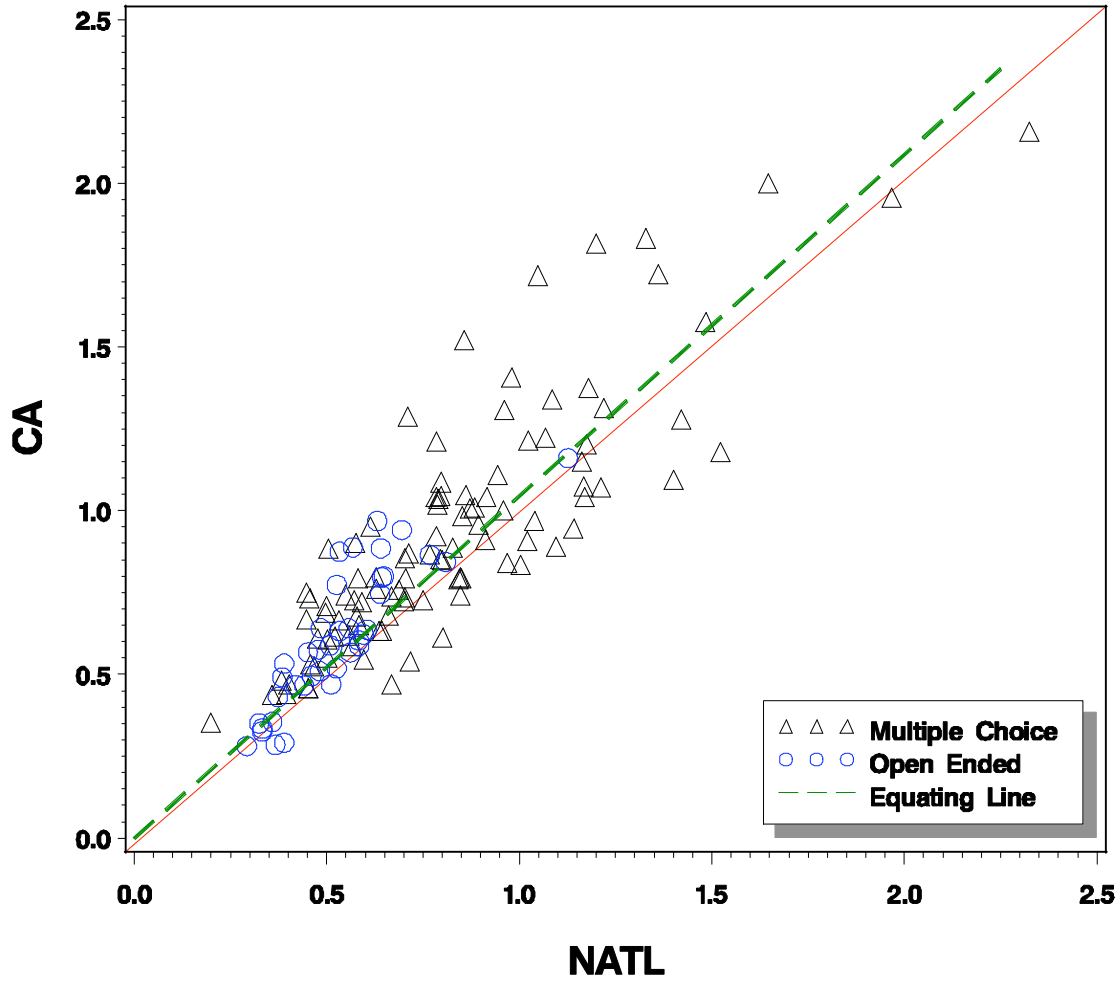
2003 NAEP Rdng Gr 8 b—plot: CA vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

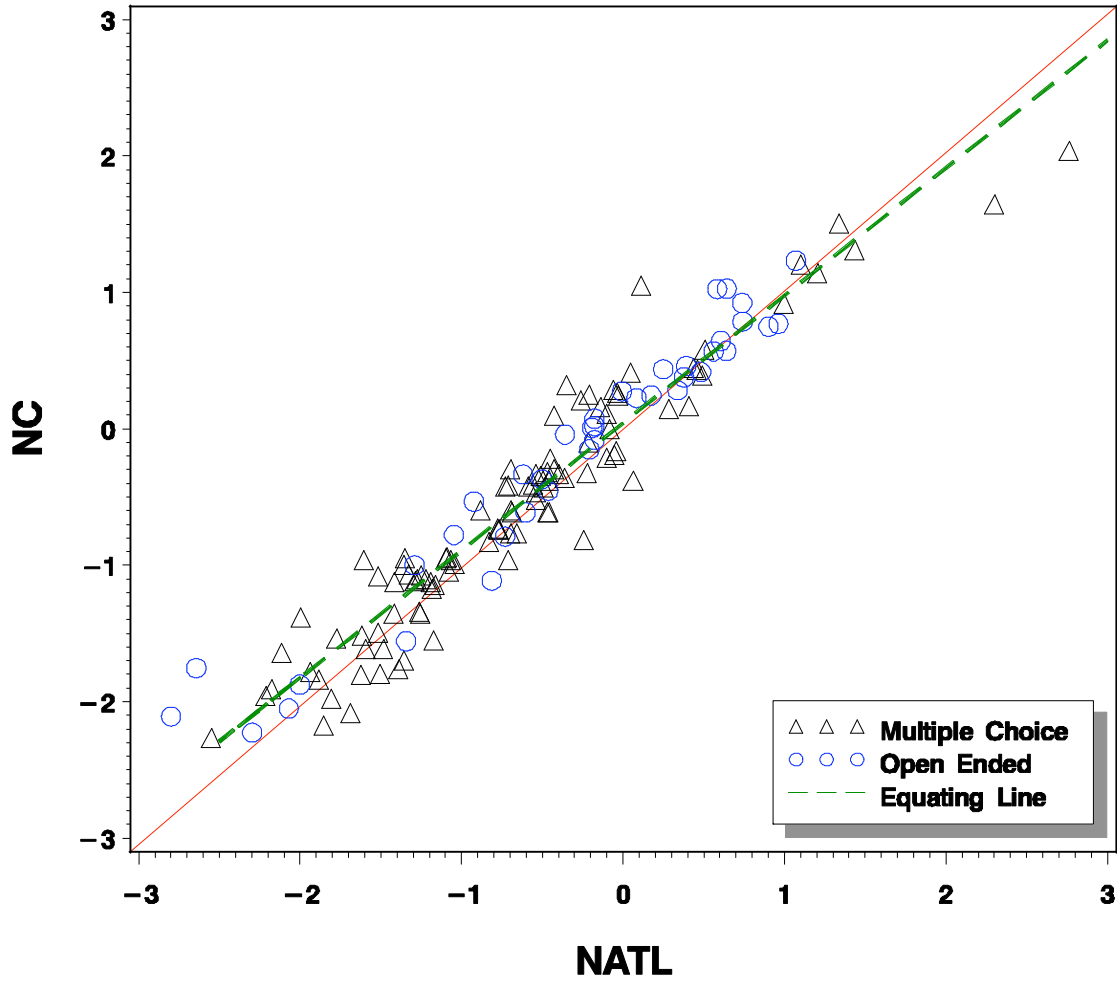
2003 NAEP Rdng Gr 8 a – plot: CA vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

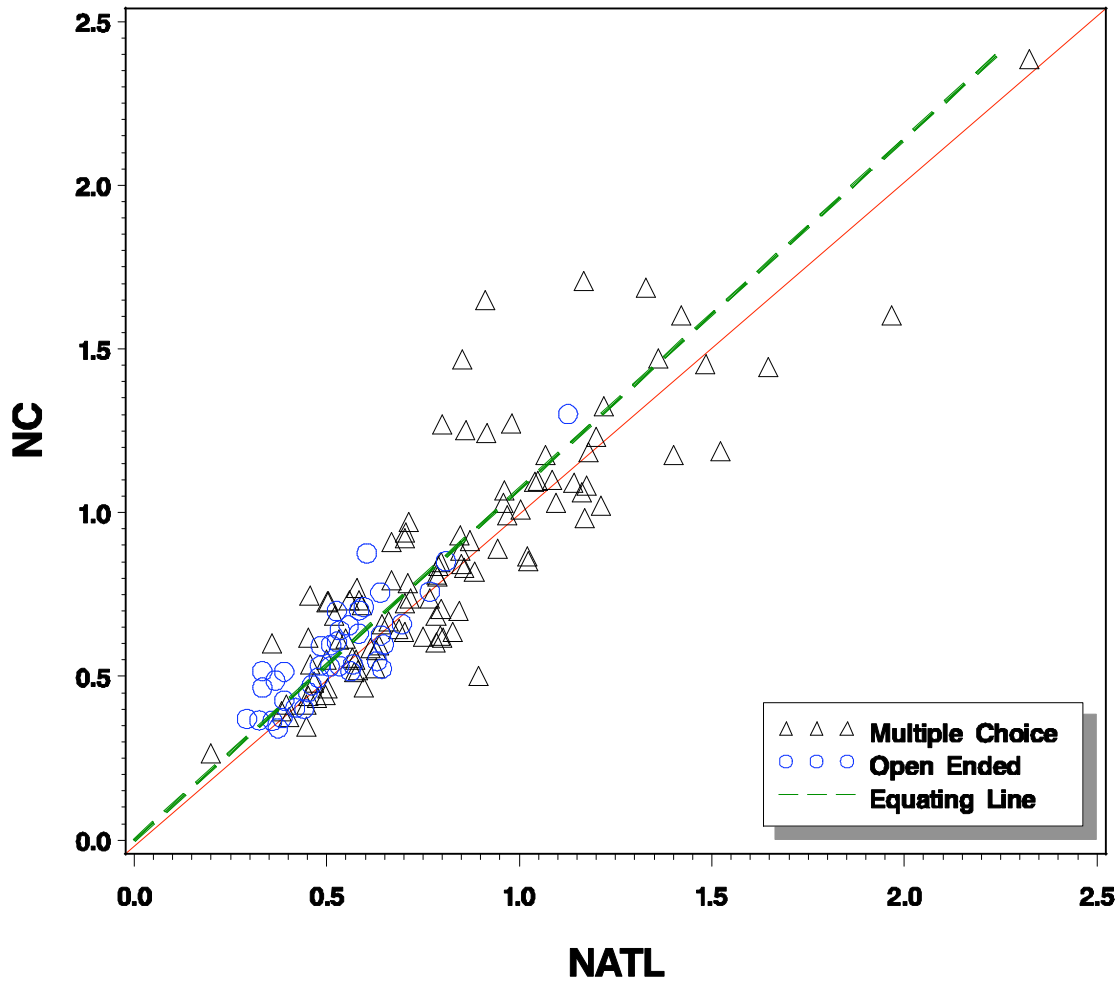
2003 NAEP Rdng Gr 8 b–plot: NC vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

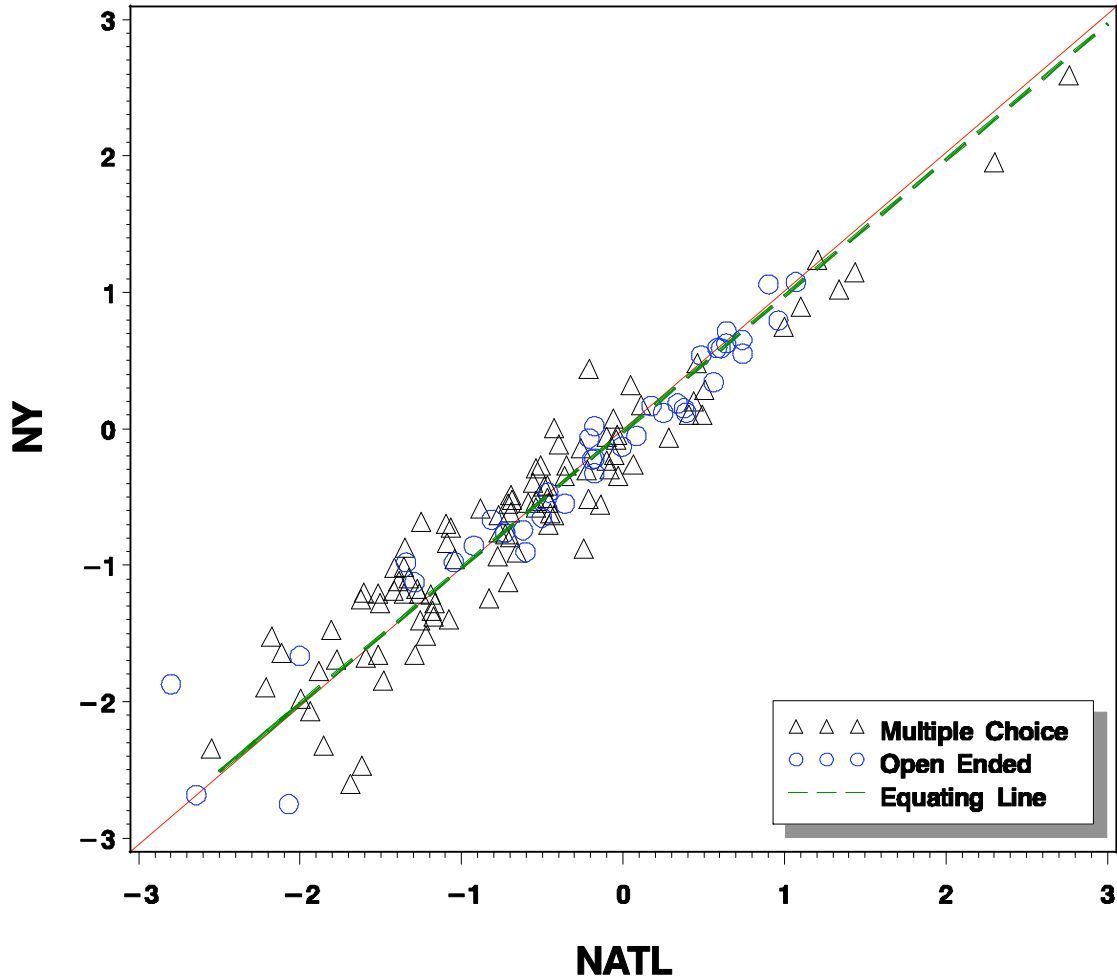
2003 NAEP Rdng Gr 8 a–plot: NC vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

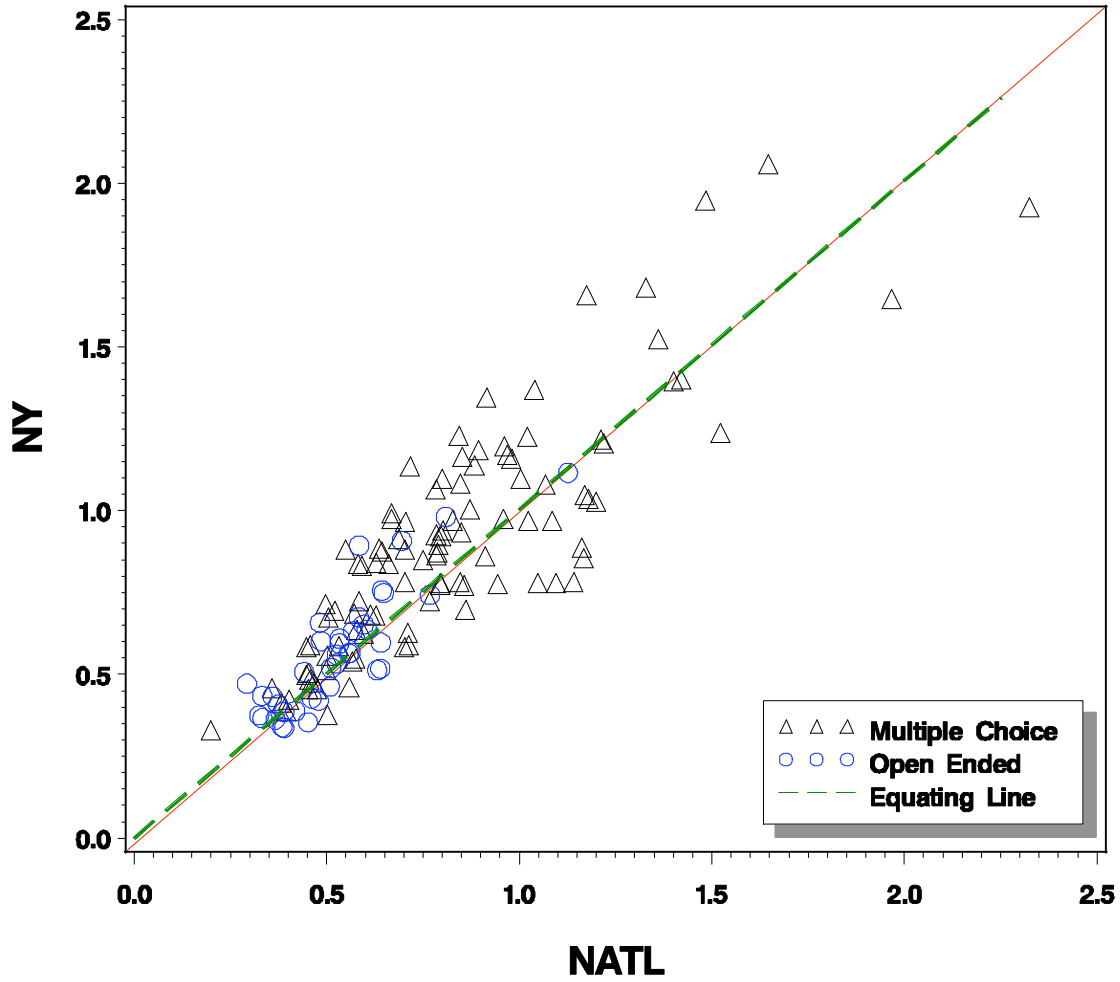
2003 NAEP Rdng Gr 8 b—plot: NY vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

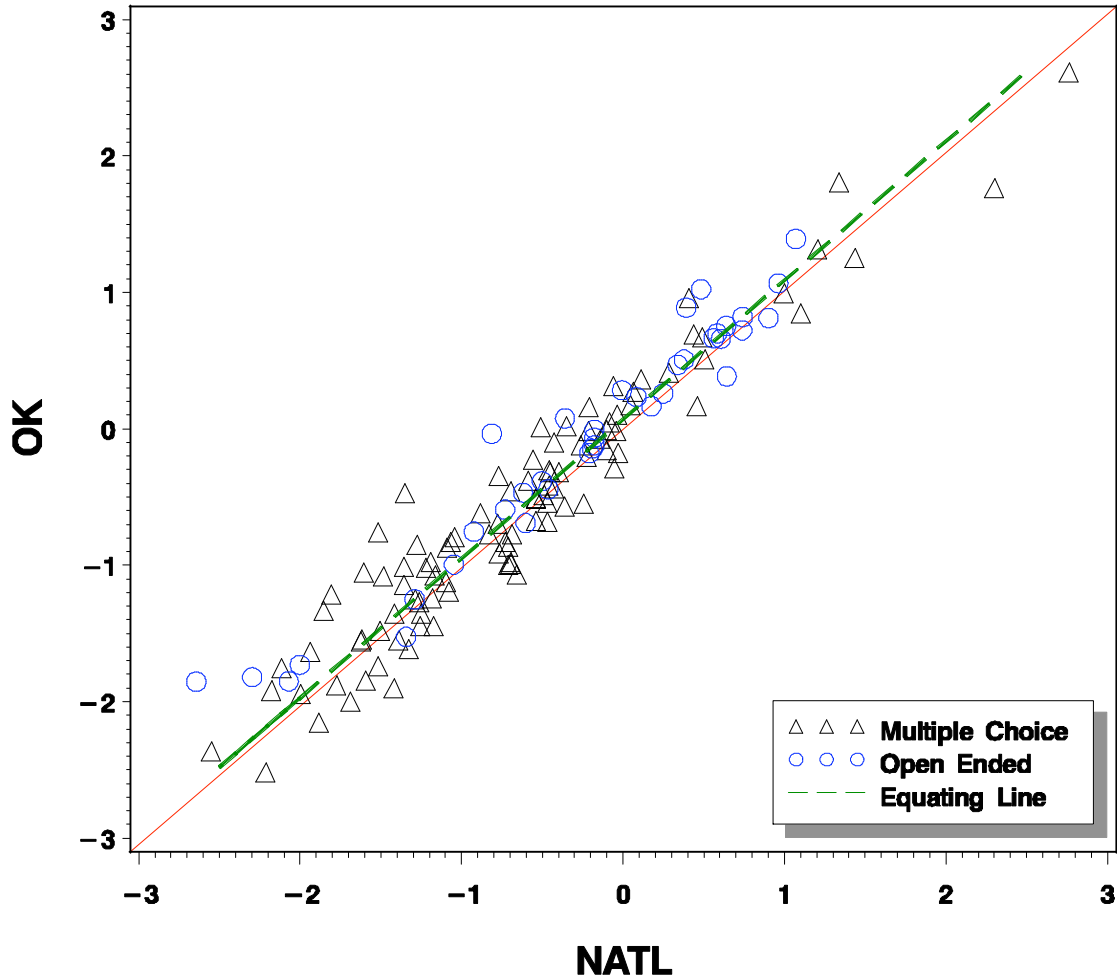
2003 NAEP Rdng Gr 8 a–plot: NY vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

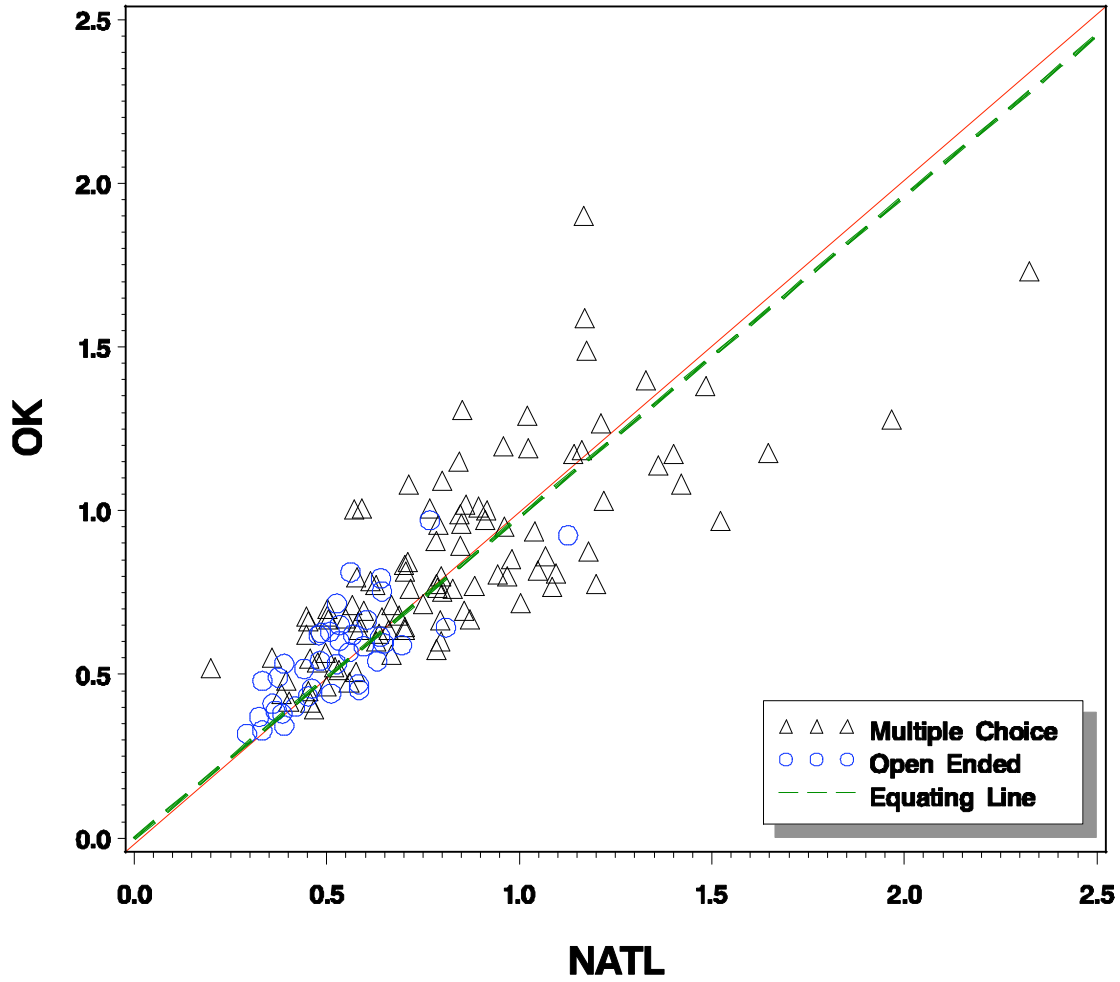
2003 NAEP Rdng Gr 8 b–plot: OK vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

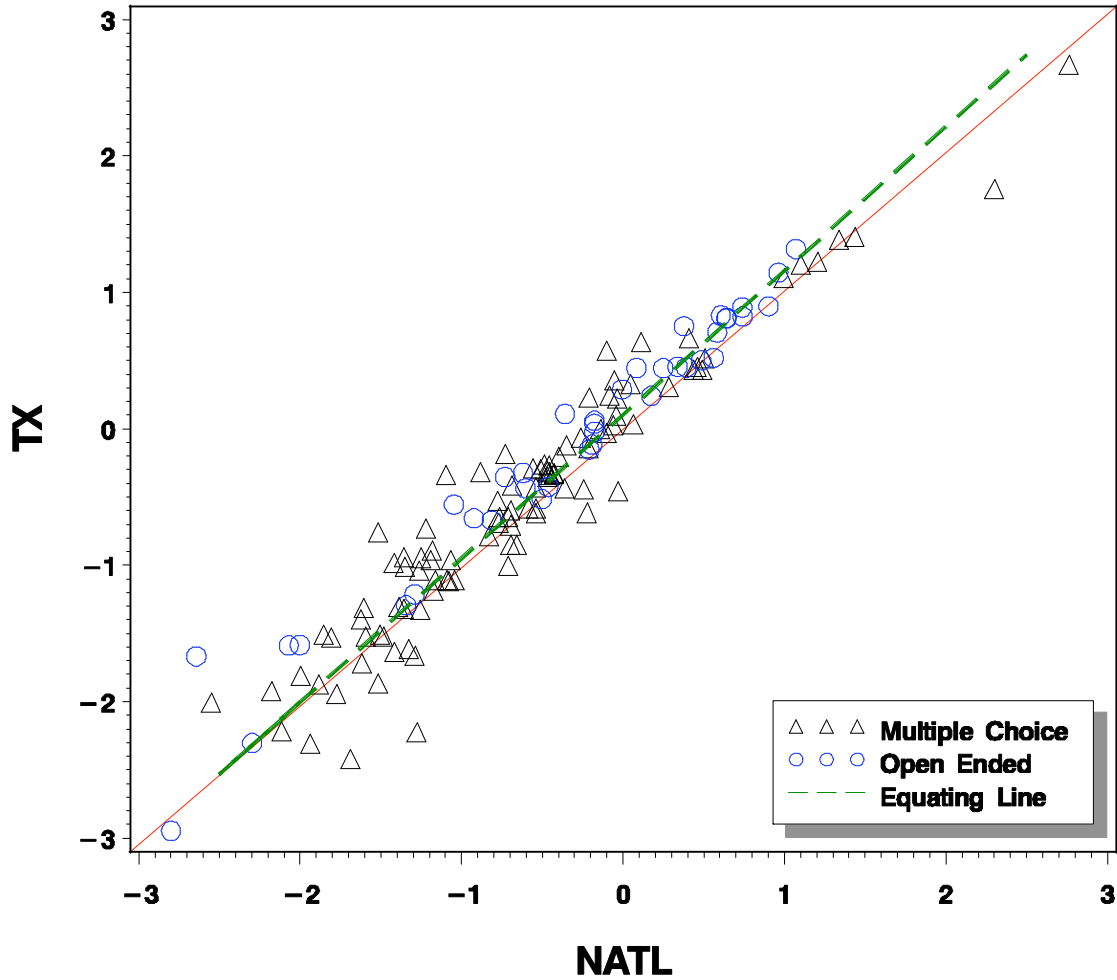
2003 NAEP Rdng Gr 8 a–plot: OK vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

2003 NAEP Rdng Gr 8 b—plot: TX vs NATL



Continues next page

Figure B-3. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs National (Continued)

2003 NAEP Rdng Gr 8 a – plot: TX vs NATL

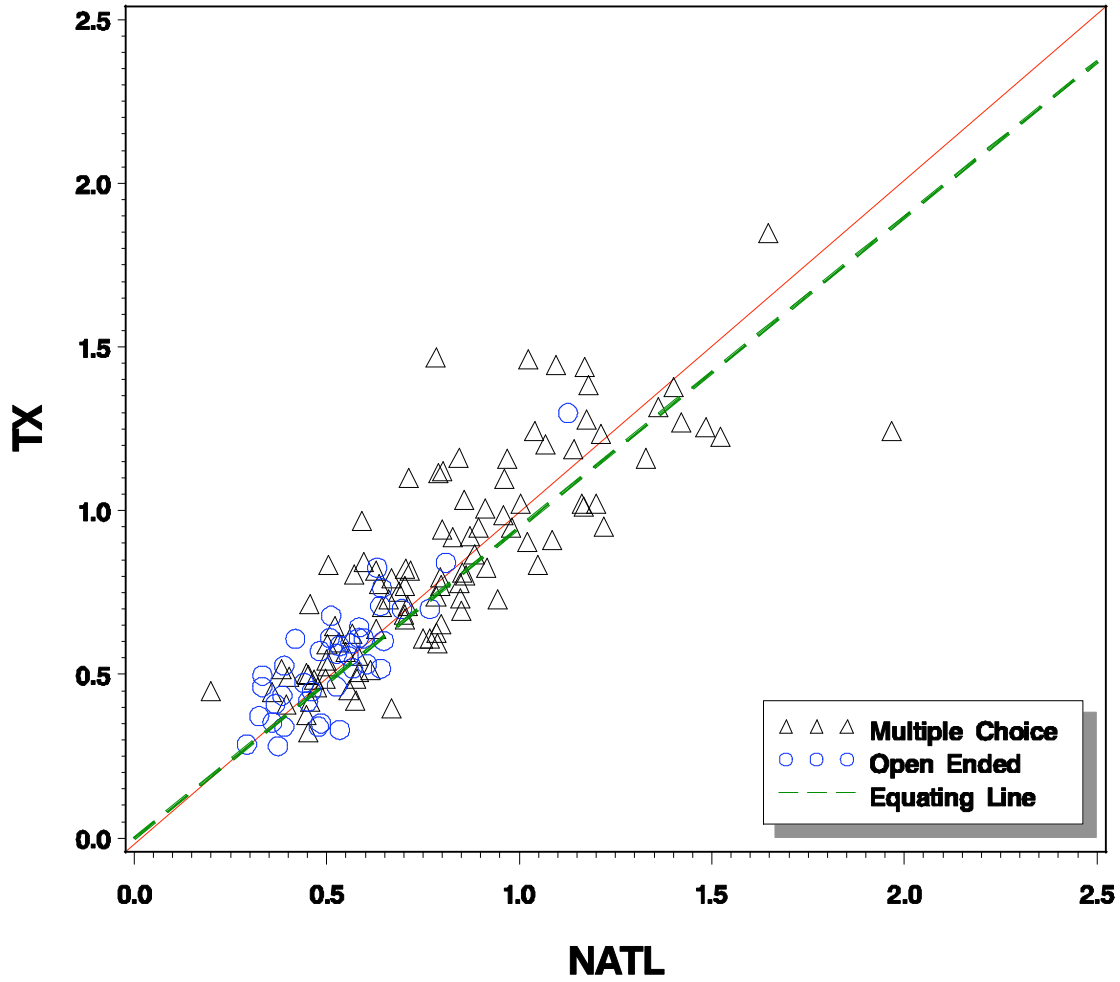
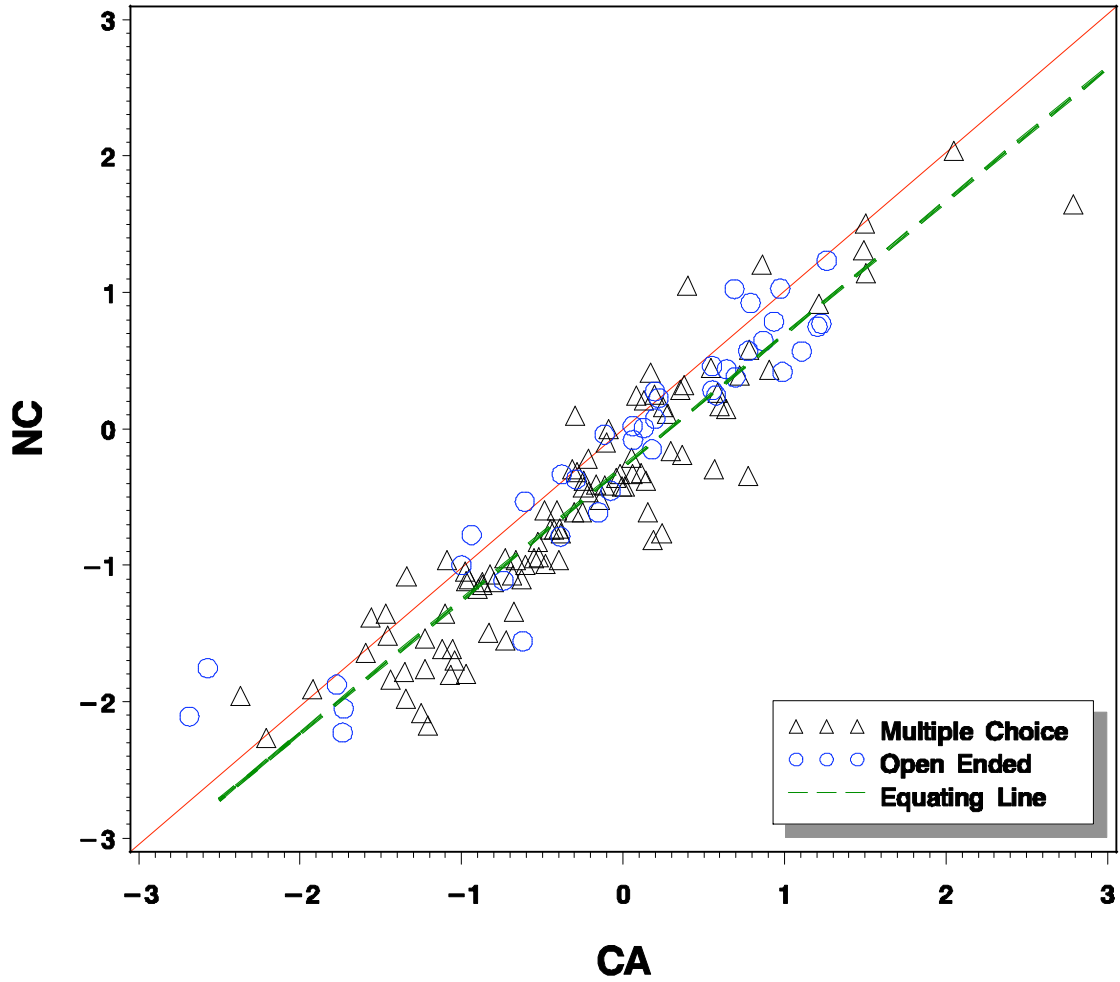


Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States

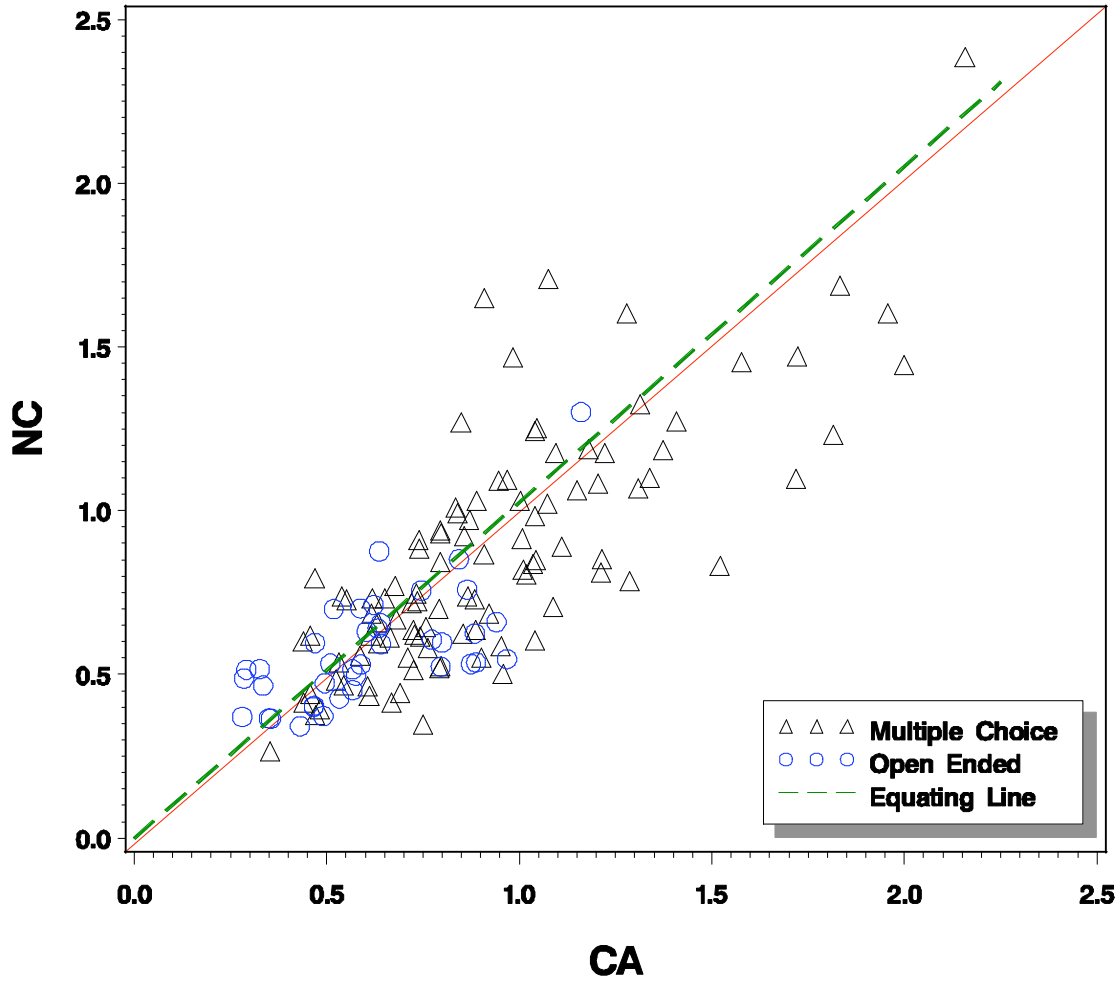
2003 NAEP Rdng Gr 8 b–plot: NC vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

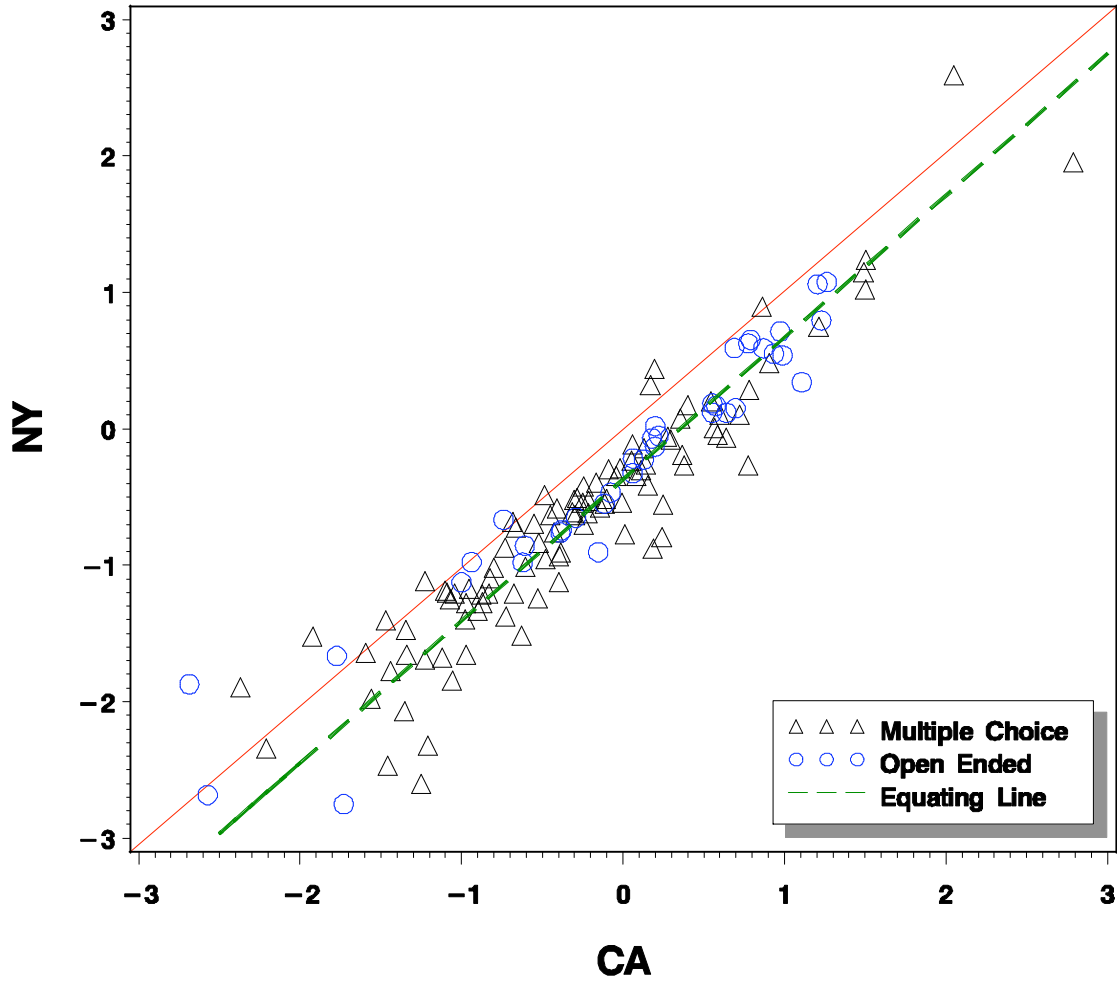
2003 NAEP Rdng Gr 8 a–plot: NC vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

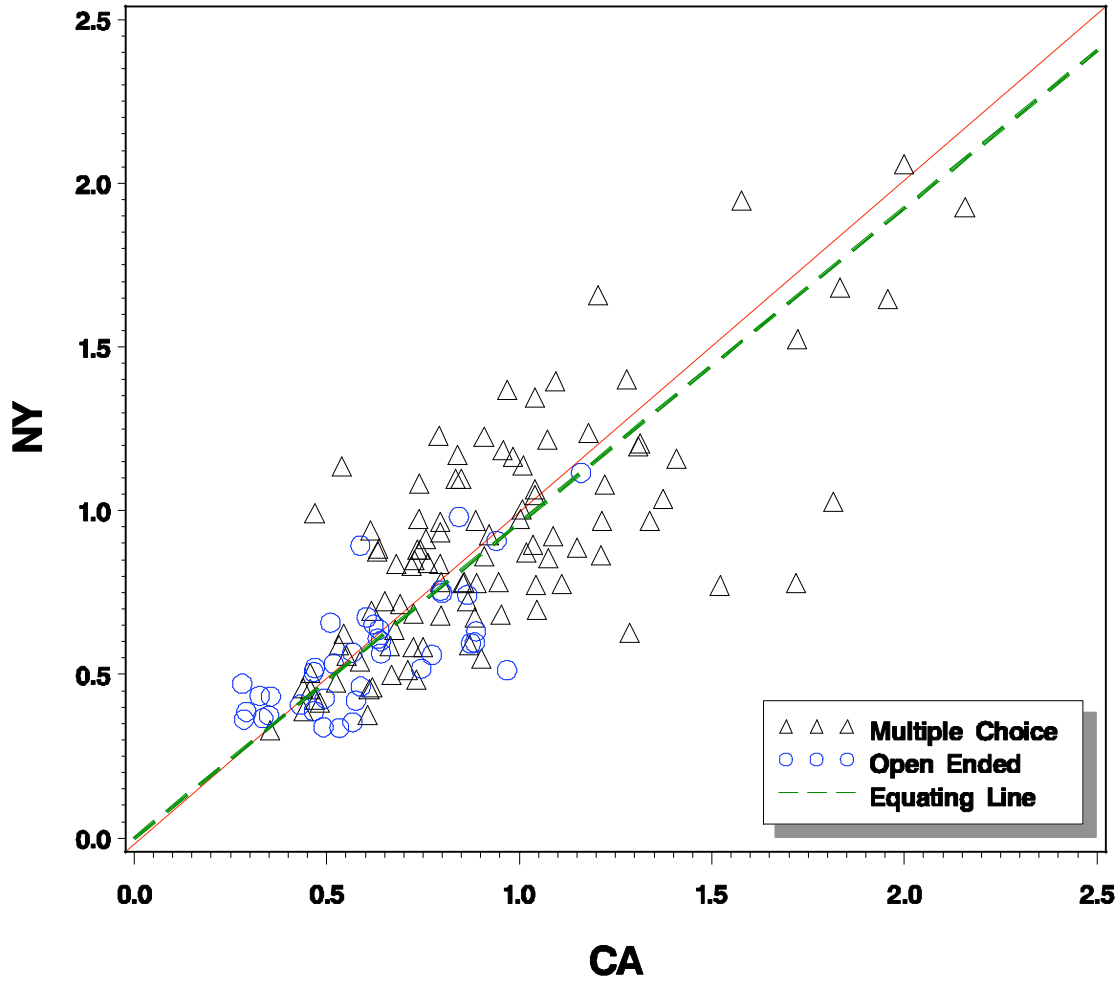
2003 NAEP Rdng Gr 8 b—plot: NY vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

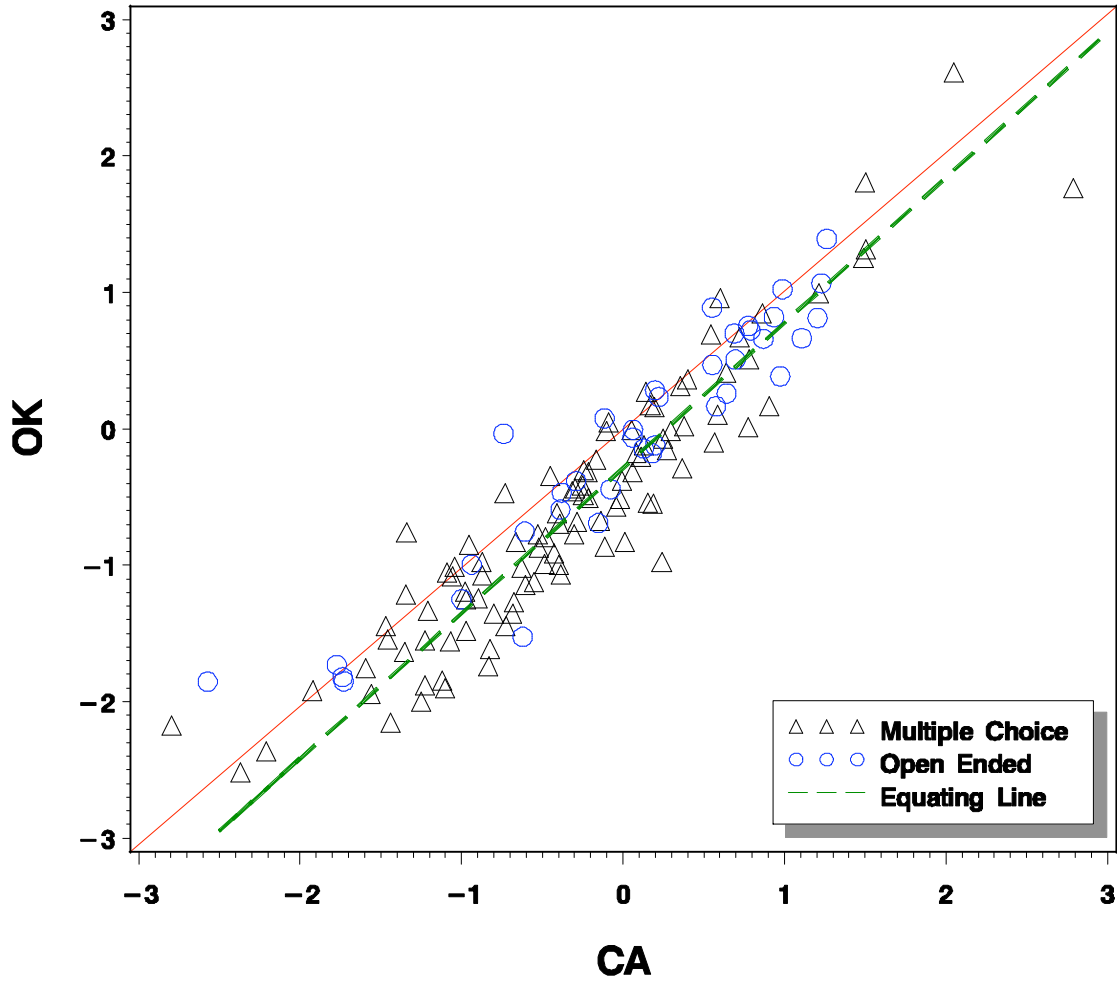
2003 NAEP Rdng Gr 8 a–plot: NY vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

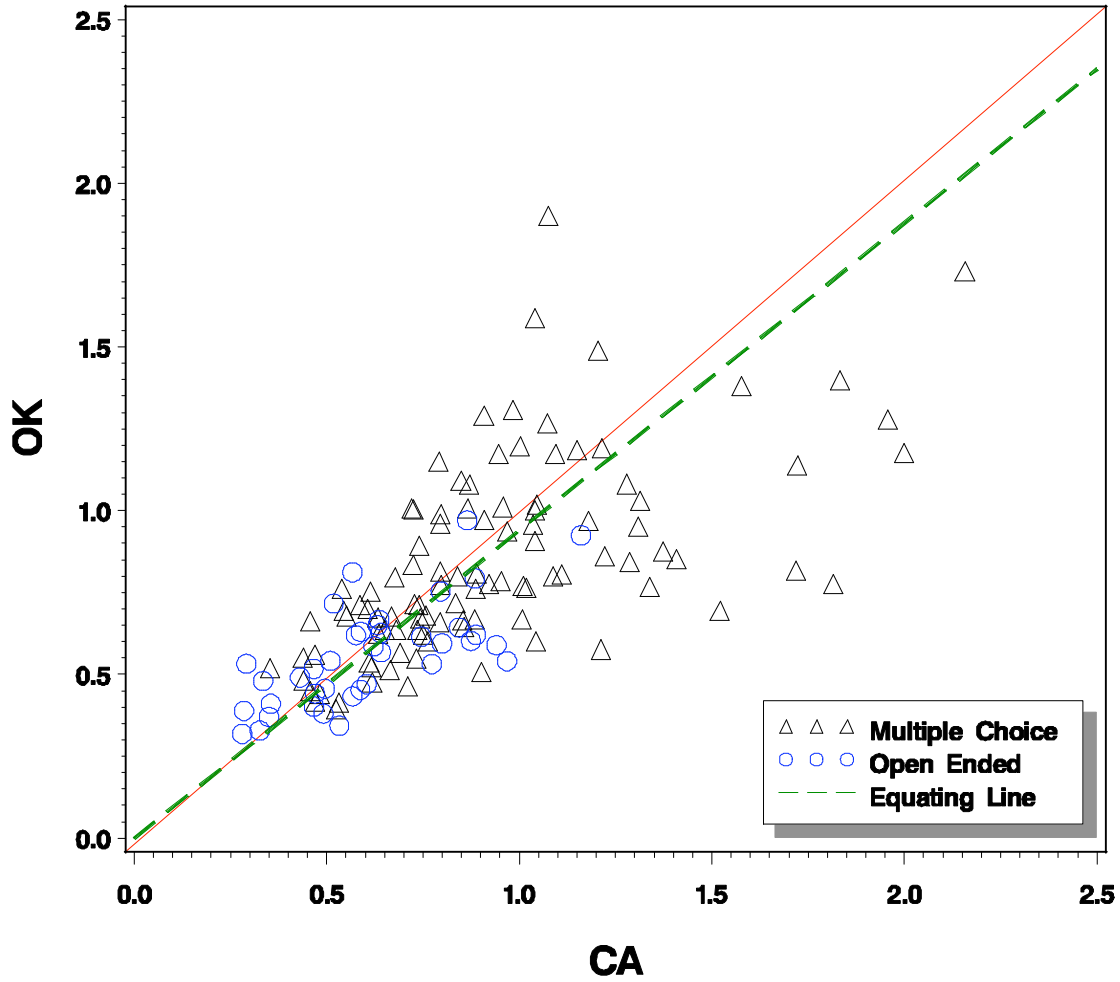
2003 NAEP Rdng Gr 8 b—plot: OK vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

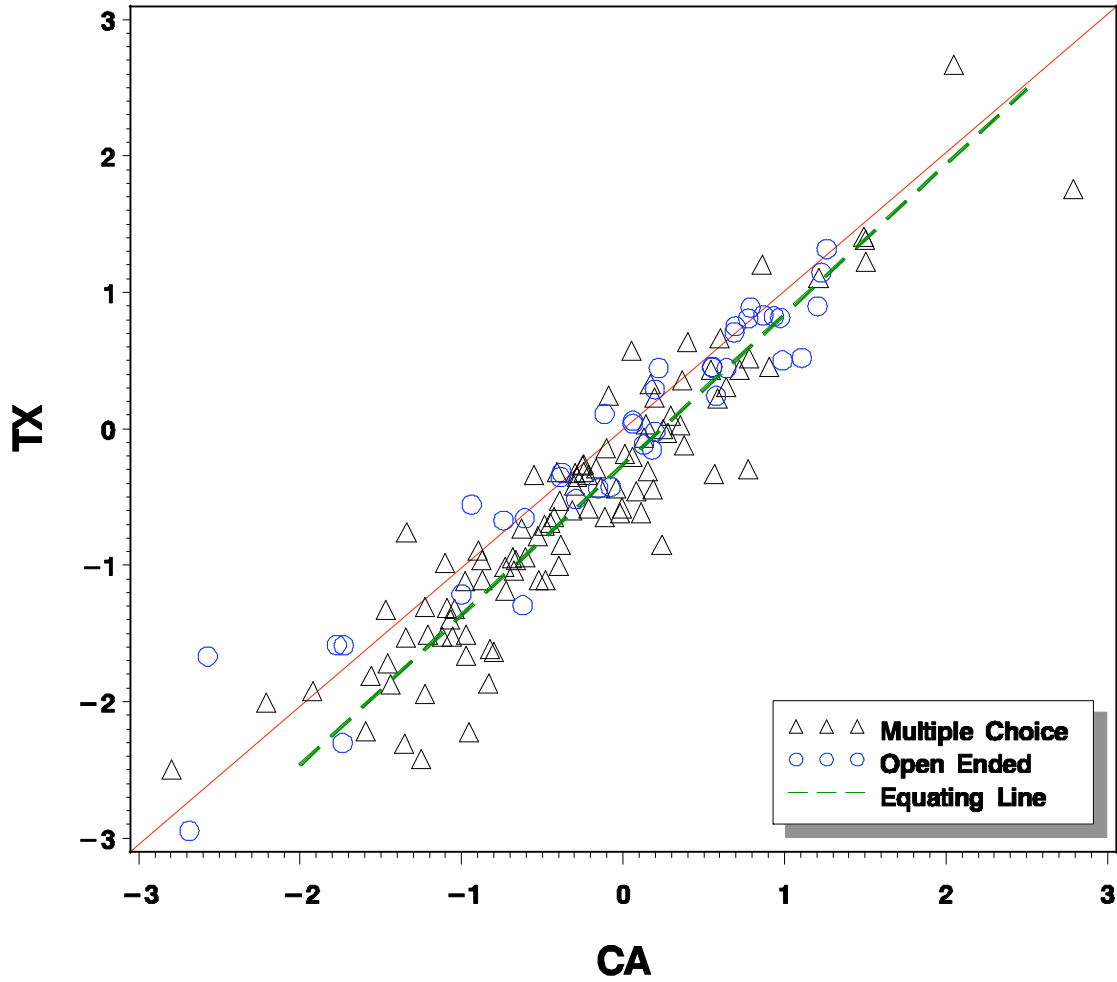
2003 NAEP Rdng Gr 8 a – plot: OK vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

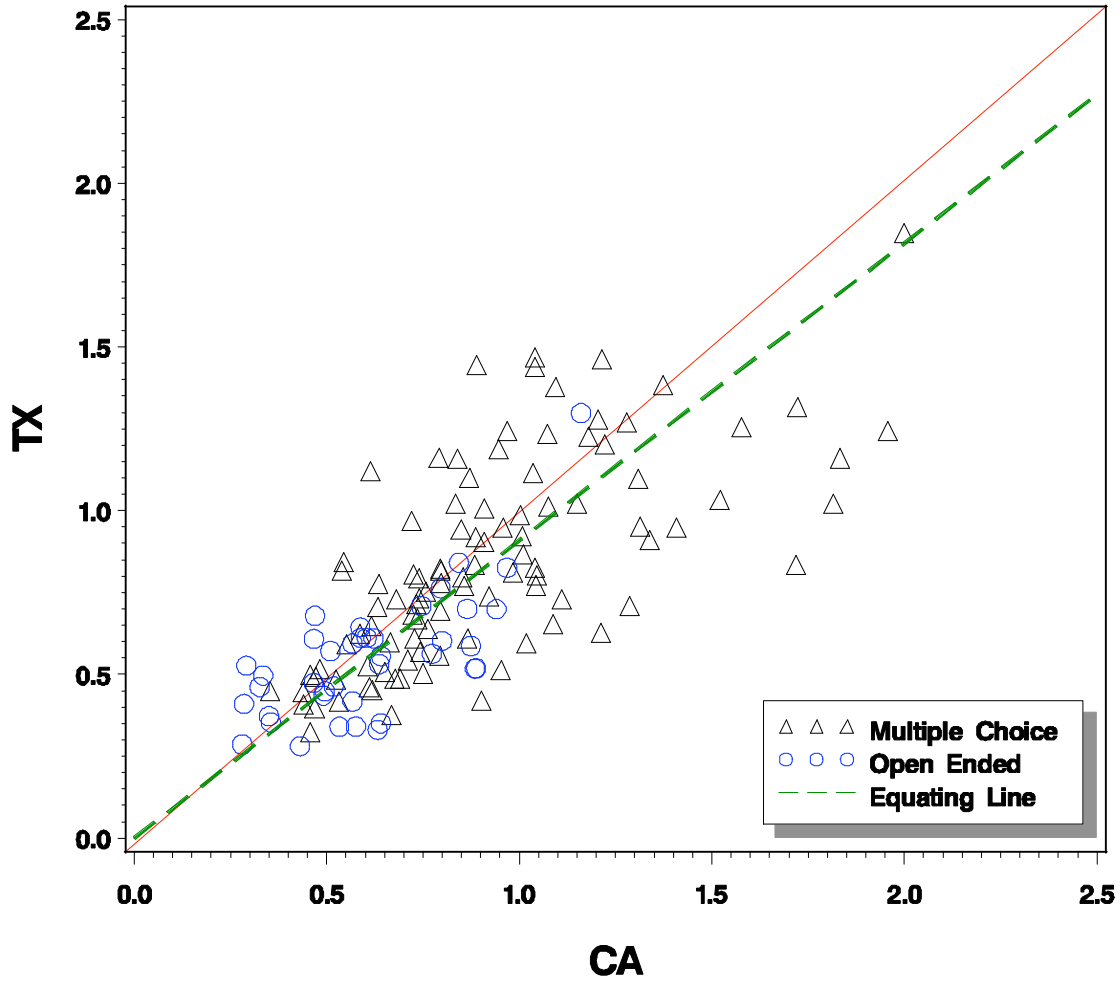
2003 NAEP Rdng Gr 8 b—plot: TX vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

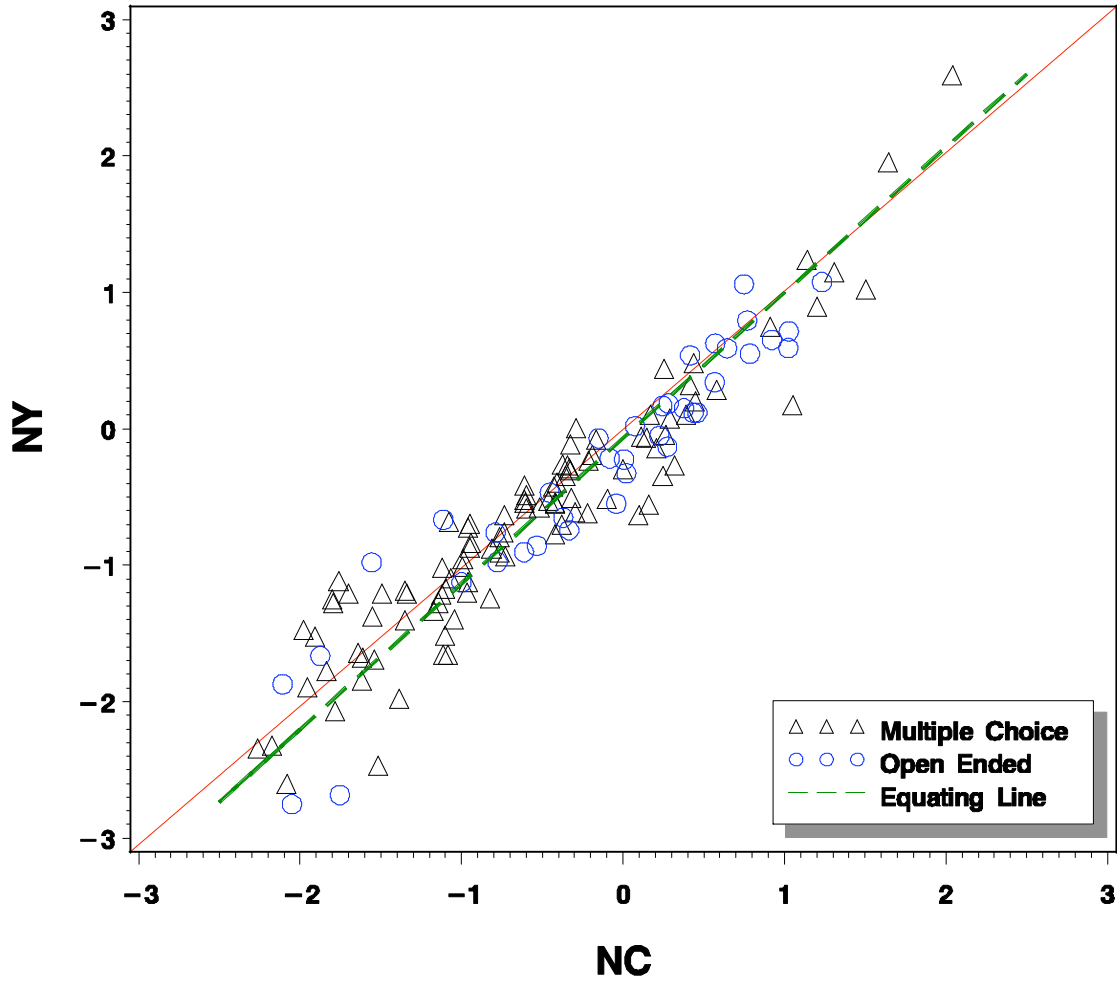
2003 NAEP Rdng Gr 8 a–plot: TX vs CA



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

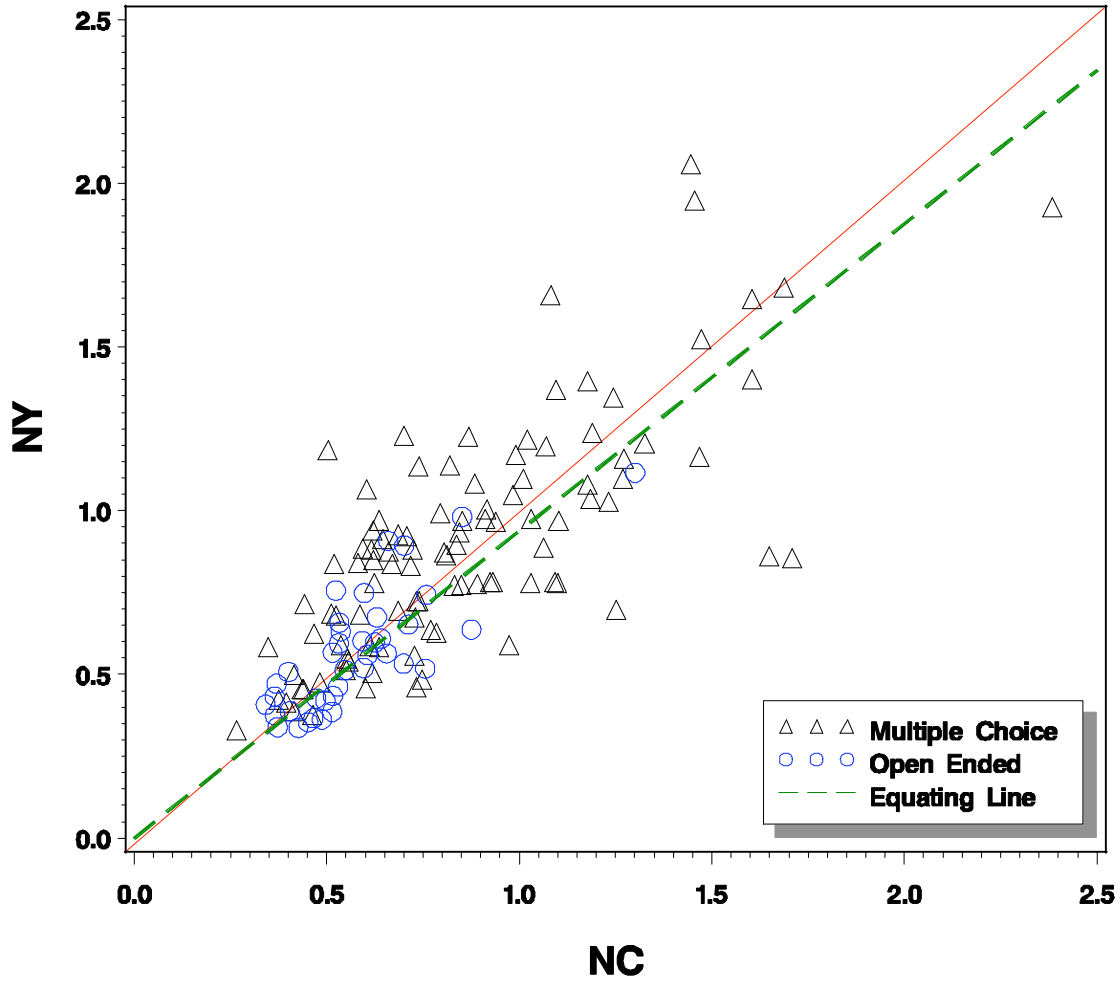
2003 NAEP Rdng Gr 8 b–plot: NY vs NC



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

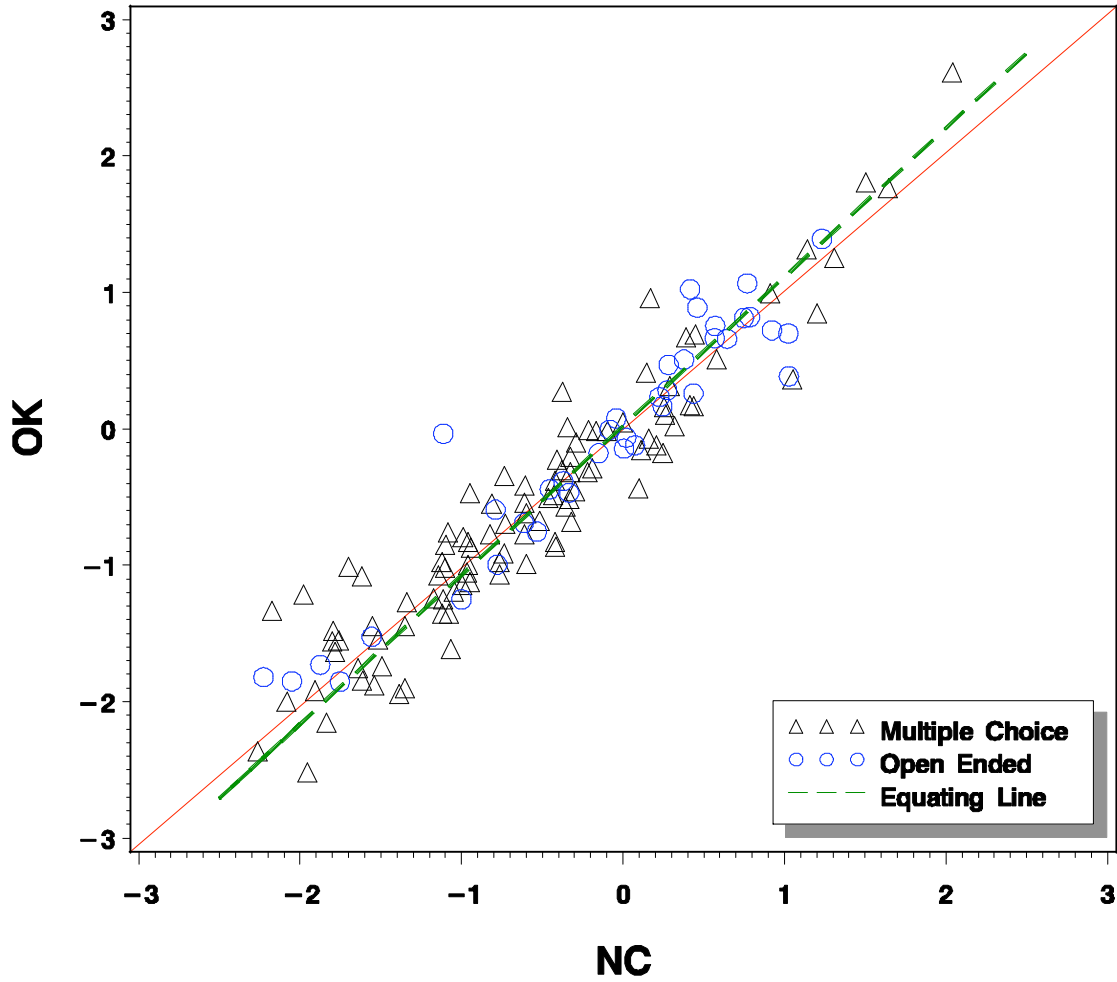
2003 NAEP Rdng Gr 8 a–plot: NY vs NC



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

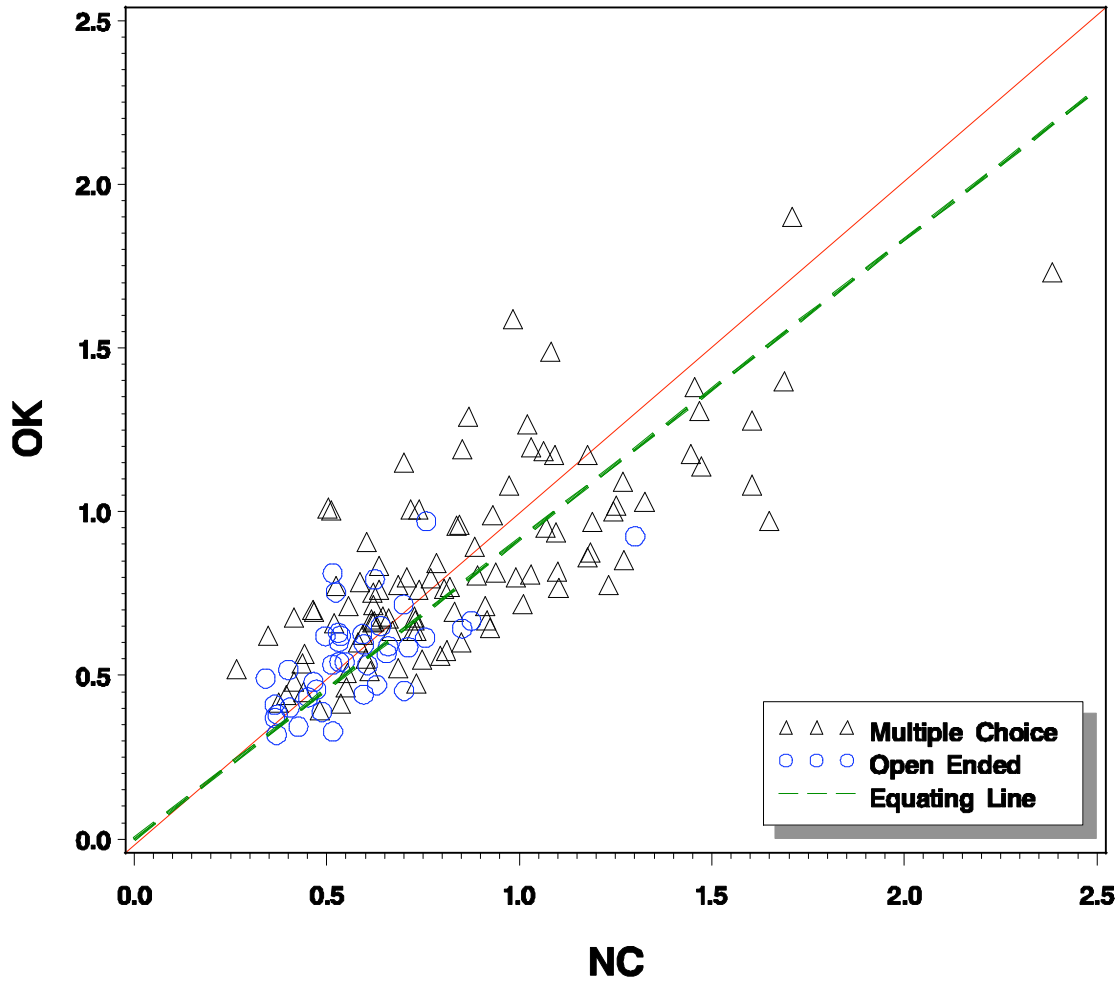
2003 NAEP Rdng Gr 8 b–plot: OK vs NC



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

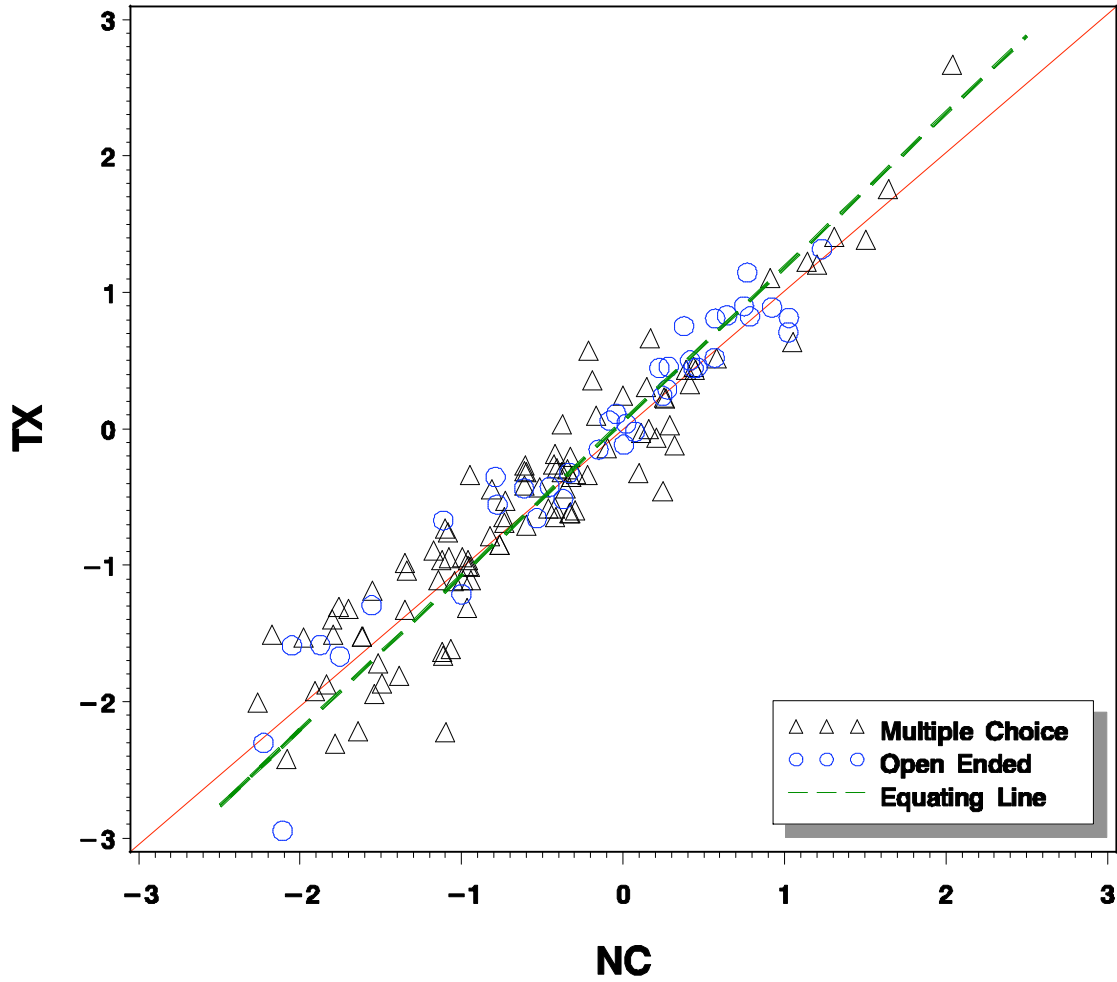
2003 NAEP Rdng Gr 8 a – plot: OK vs NC



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

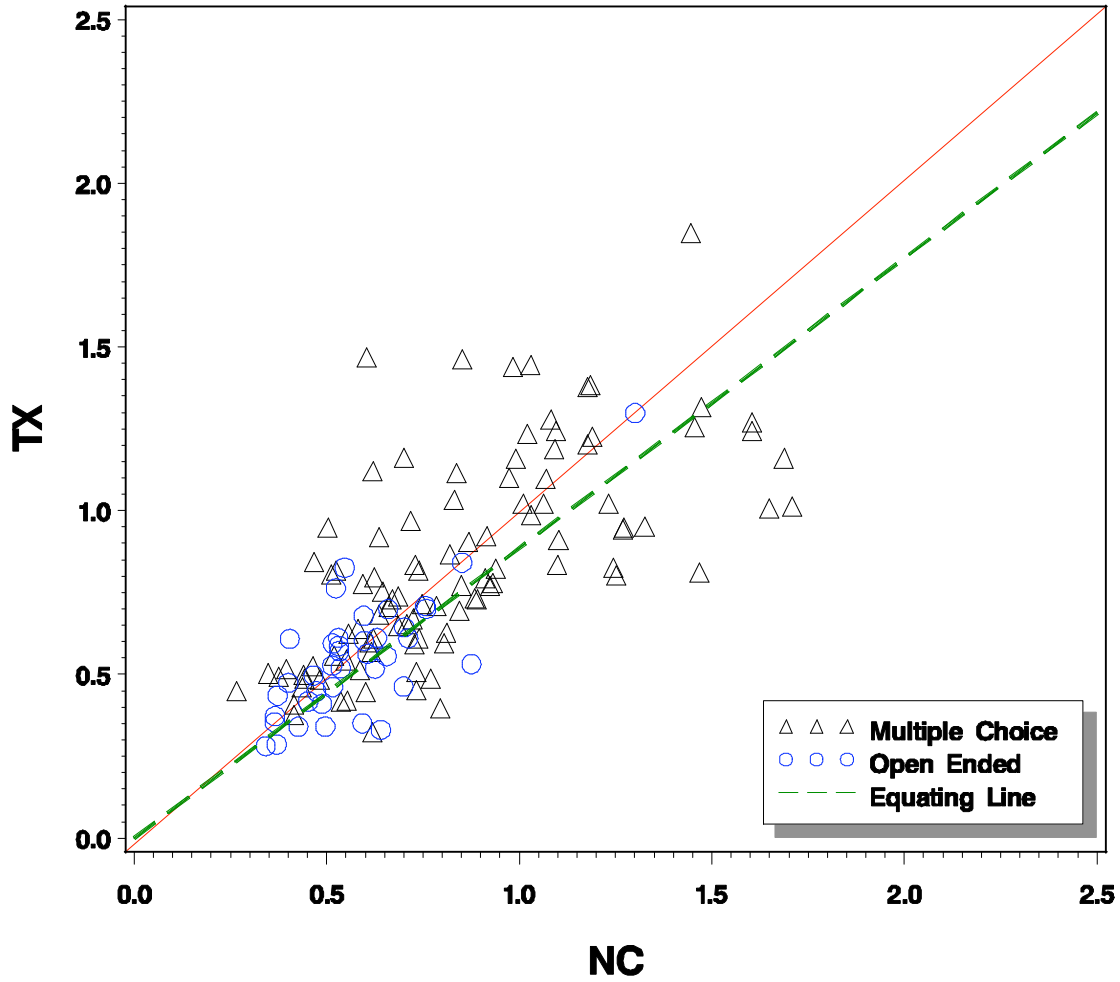
2003 NAEP Rdng Gr 8 b–plot: TX vs NC



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

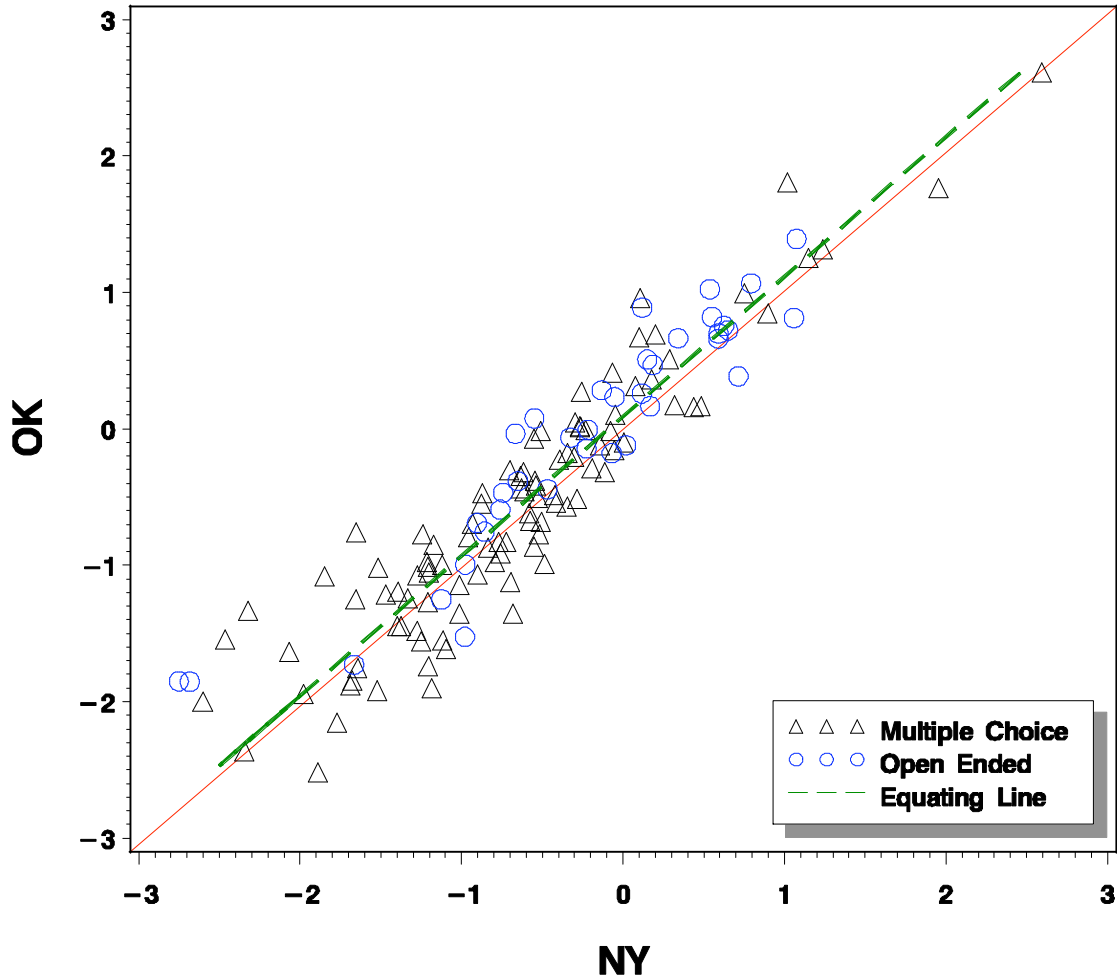
2003 NAEP Rdng Gr 8 a–plot: TX vs NC



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

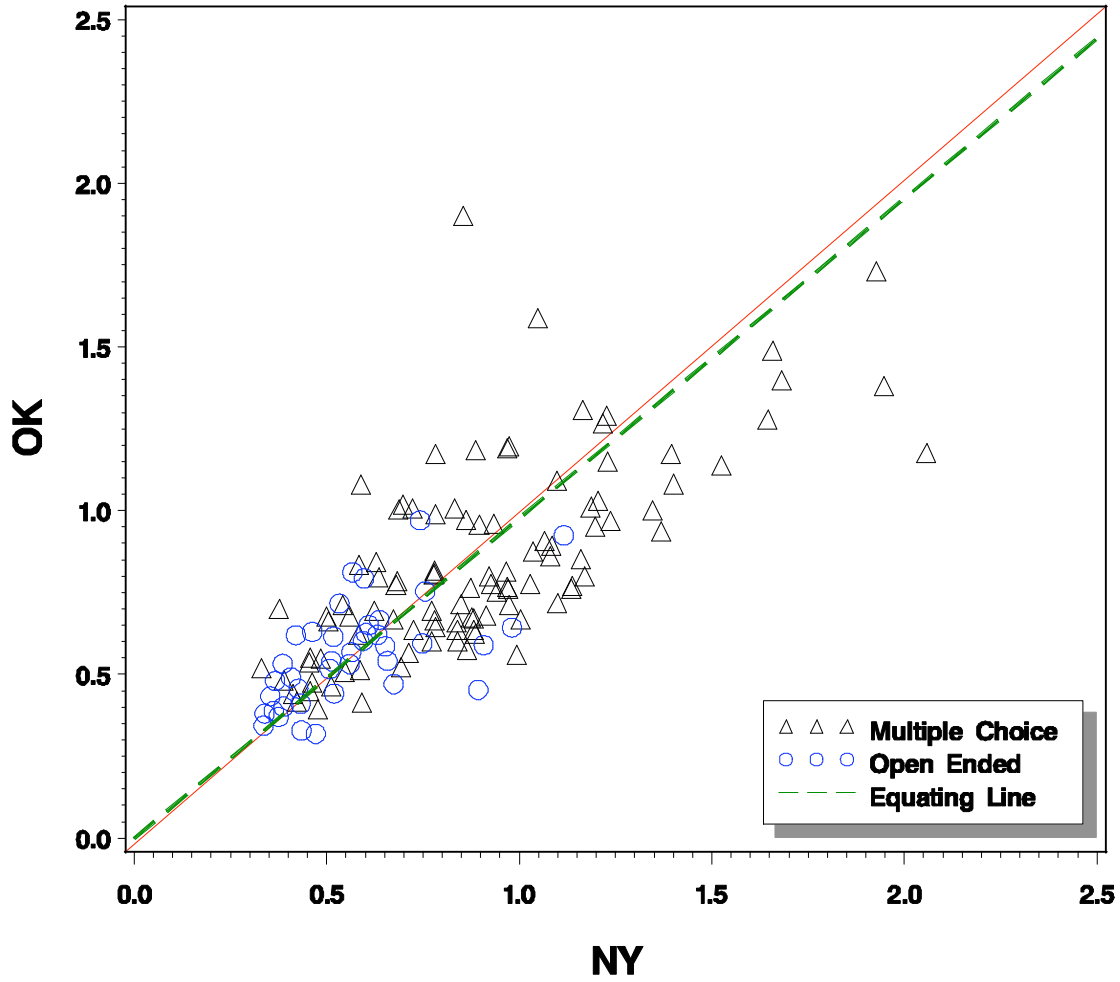
2003 NAEP Rdng Gr 8 b–plot: OK vs NY



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

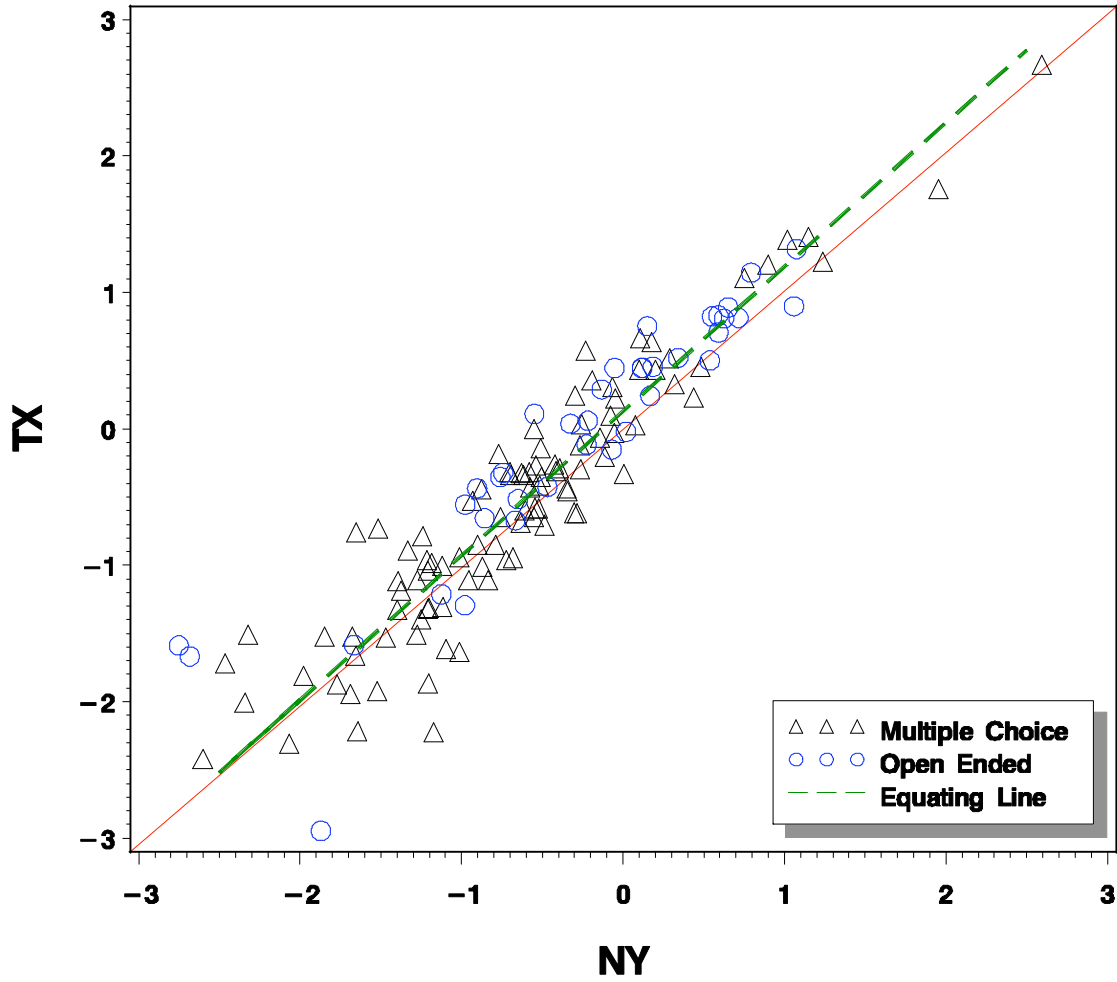
2003 NAEP Rdng Gr 8 a–plot: OK vs NY



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

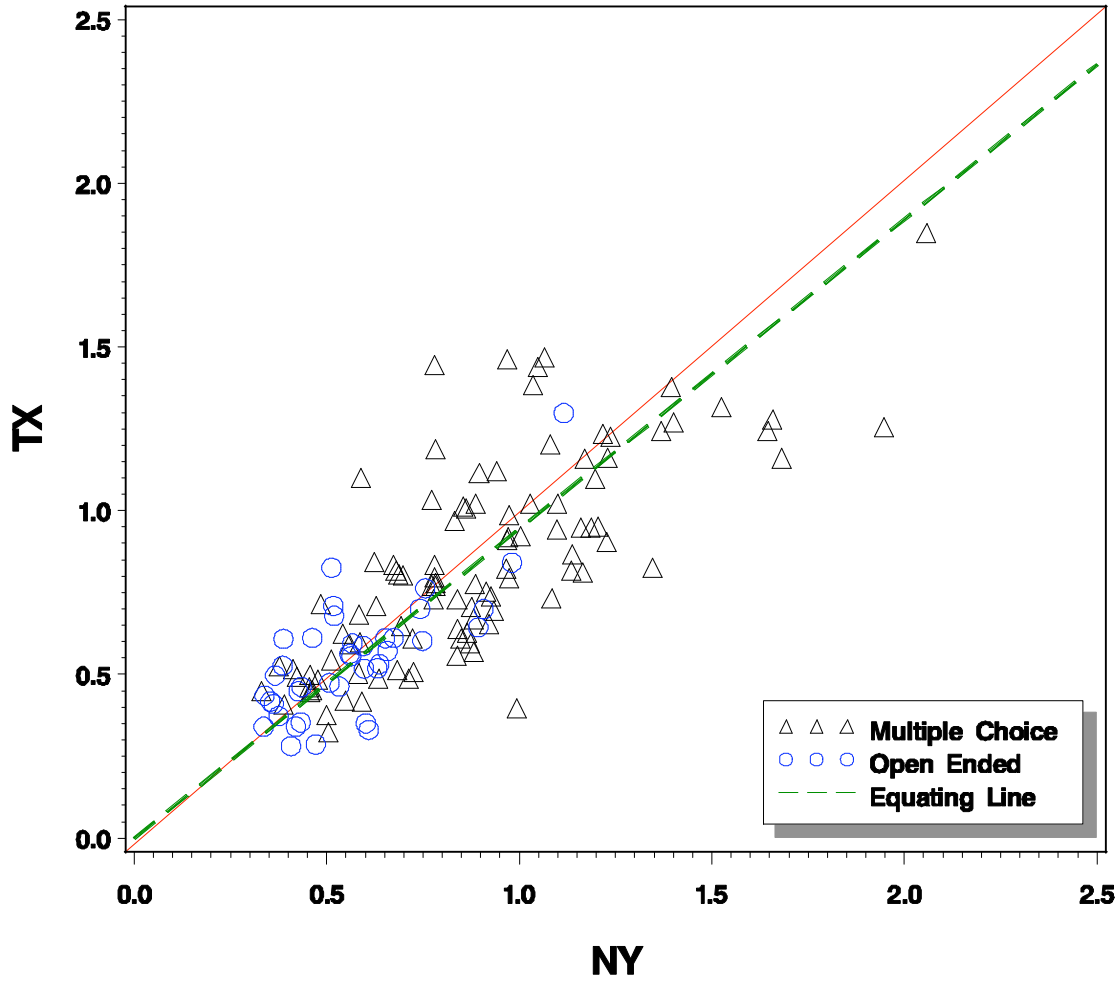
2003 NAEP Rdng Gr 8 b—plot: TX vs NY



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

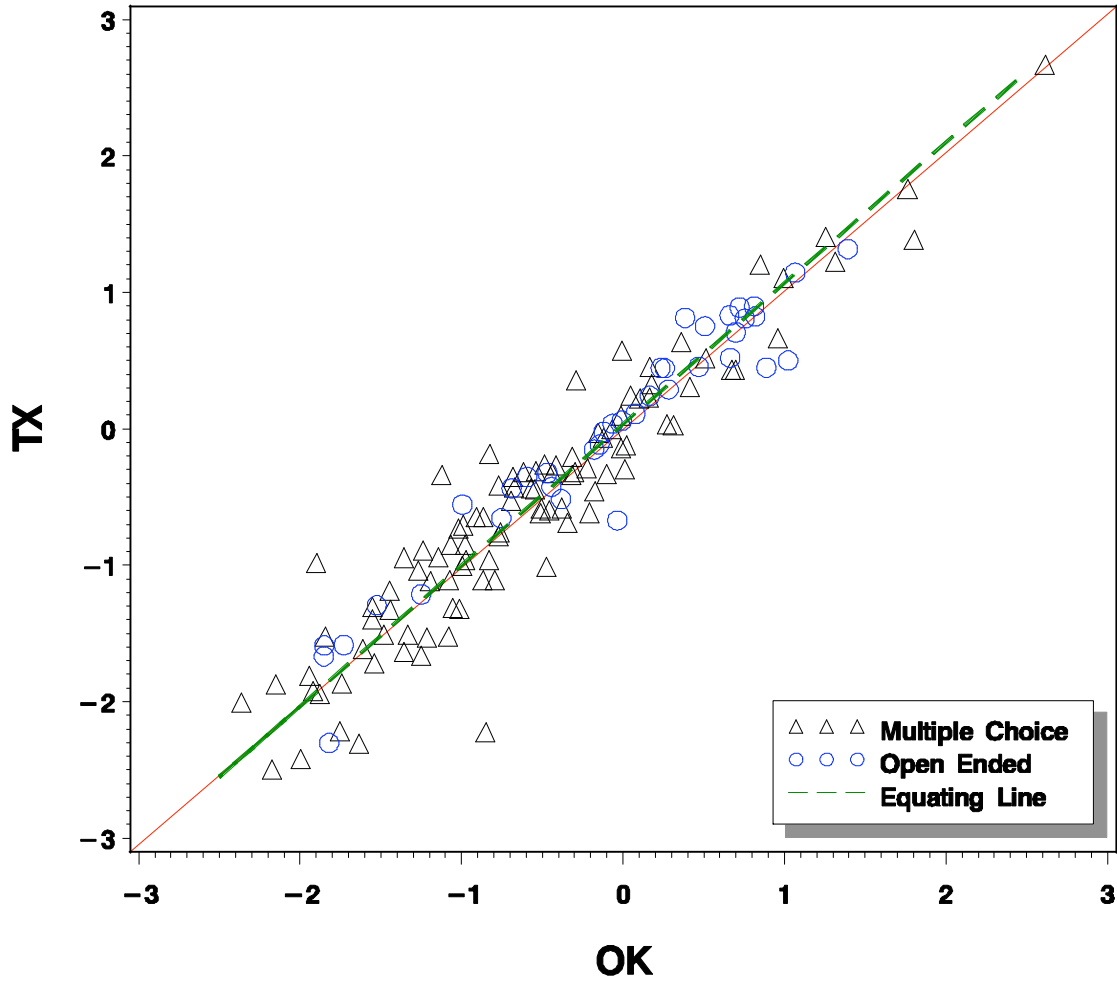
2003 NAEP Rdng Gr 8 a–plot: TX vs NY



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

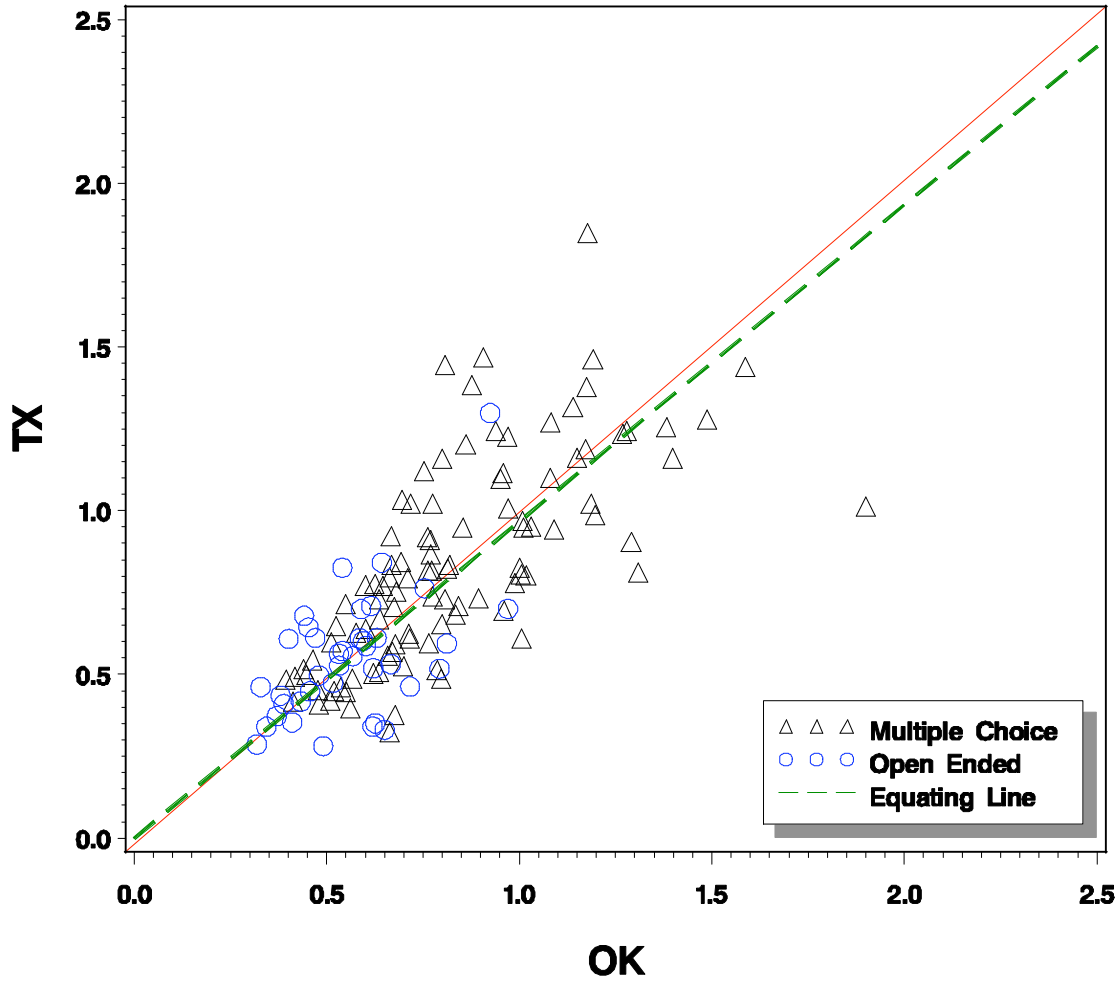
2003 NAEP Rdng Gr 8 b–plot: TX vs OK



Continues next page

Figure B-4. 2003 NAEP Reading Gr 8 a- and b- plots: Selected States vs States (Continued)

2003 NAEP Rdng Gr 8 a–plot: TX vs OK



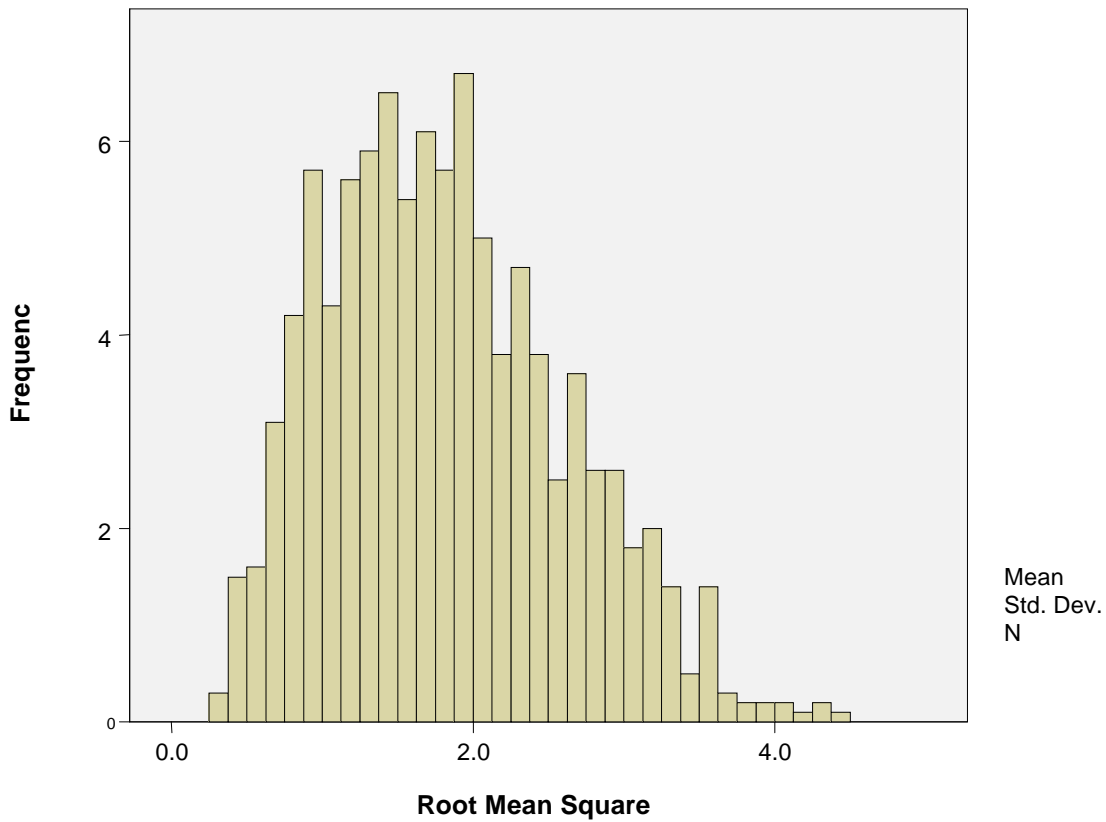
This page left intentionally blank

**Appendix C: Empirical Distribution for *RMSD*, Based on the Math Assessment, for Three
Sample Sizes: $N=2,800$, $N=4,000$, and $N=11,000$**

Figure C-1. Empirical Distribution for RMSD, Based on the Math Assessment for Three Sample Sizes

$N=2,800$

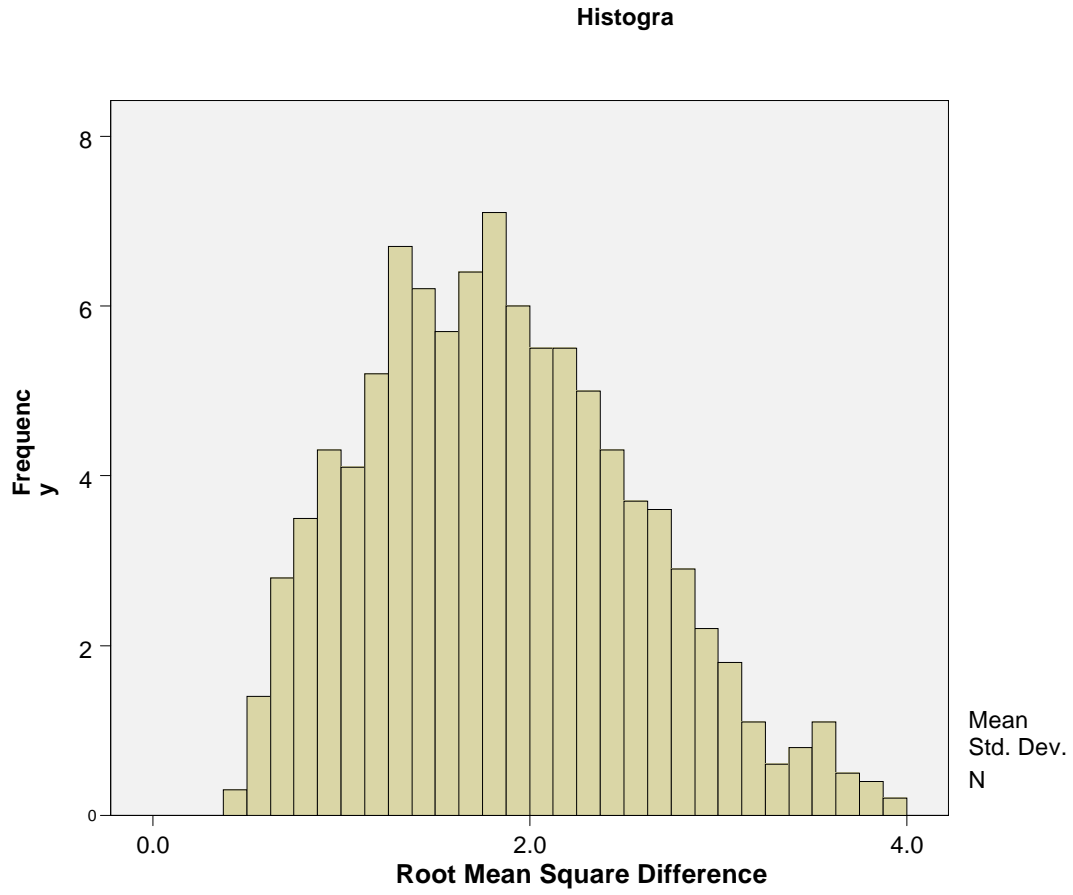
Histogram



Continues next page

Figure C-1. Empirical Distribution for RMSD, Based on the Math Assessment for Three Sample Sizes (Continued)

$N=4,000$

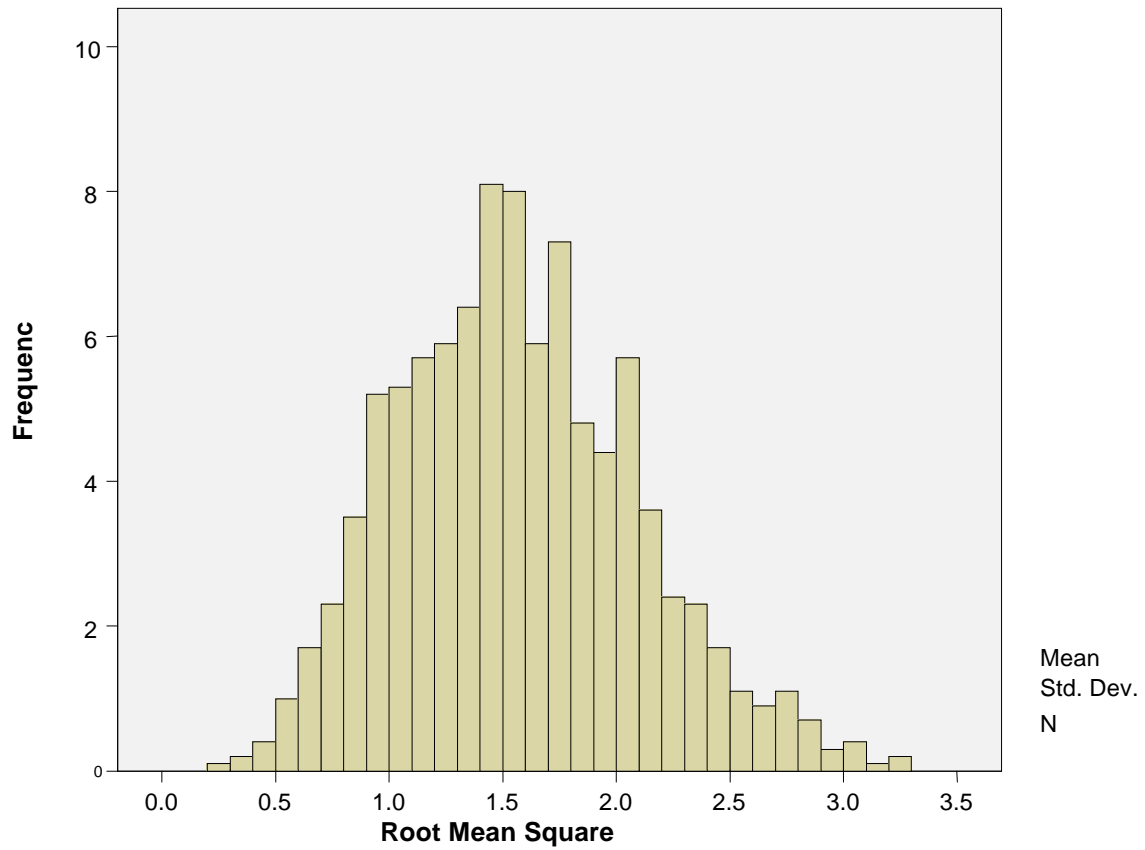


Continues next page

Figure C-1. Empirical Distribution for RMSD, Based on the Math Assessment for Three Sample Sizes (Continued)

$N=11,000$

Histogram

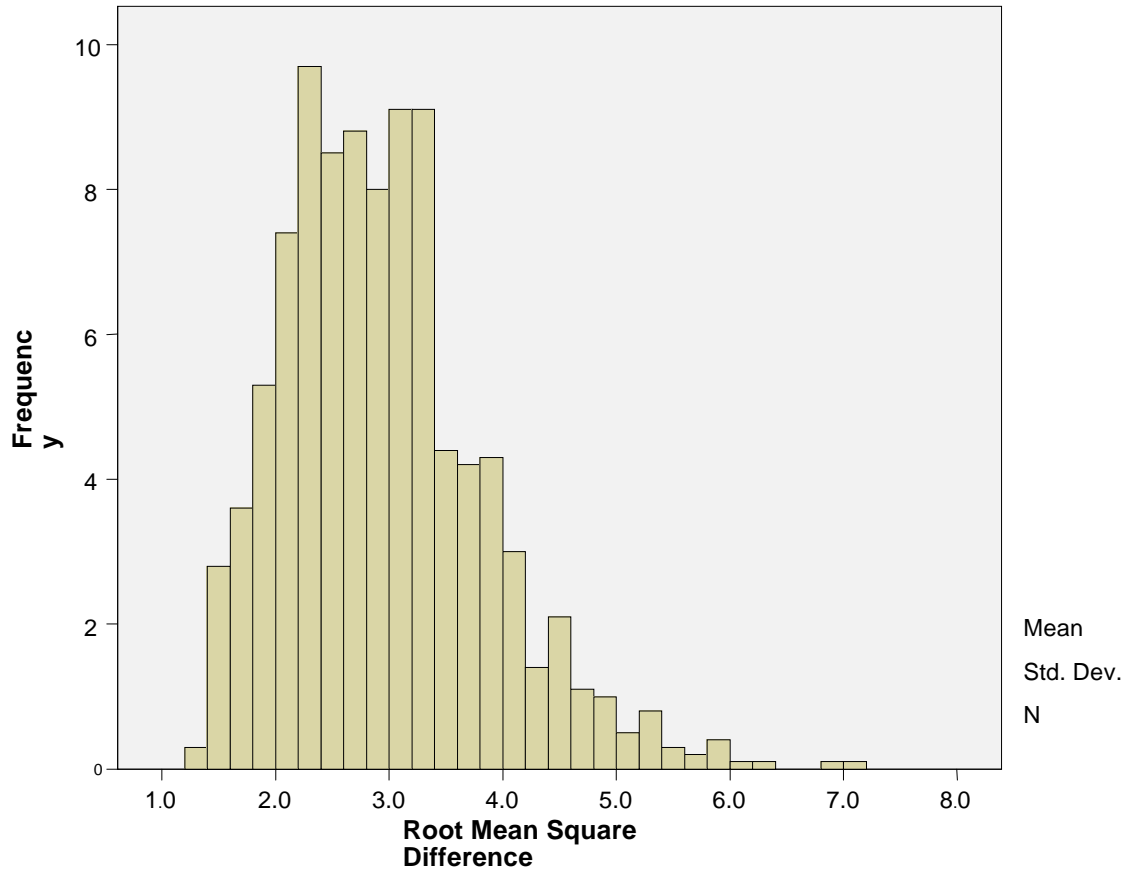


**Appendix D: Empirical Distribution for *RMSD*, Based on the Reading Assessment, for
Four Sample Sizes: $N=2,560$, $N=4,000$, $N=7,520$, and $N=10,000$**

Figure D-1. Empirical Distribution for RMSD, Based on the Reading Assessment, for Four Sample Sizes

$N=2,560$

Histogram –

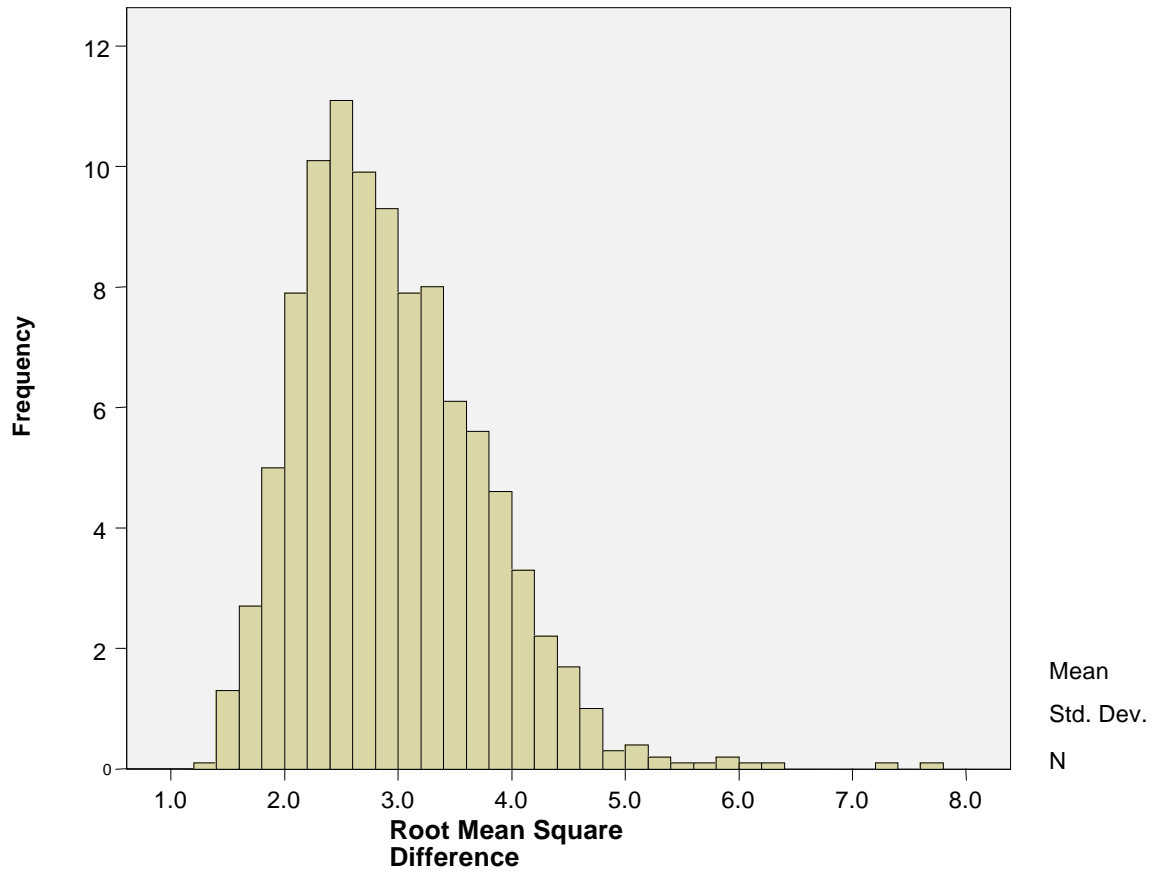


Continues next page

Figure D-1. Empirical Distribution for RMSD, Based on the Reading Assessment, for Four Sample Sizes (Continued)

$N=4,000$

Histogram –

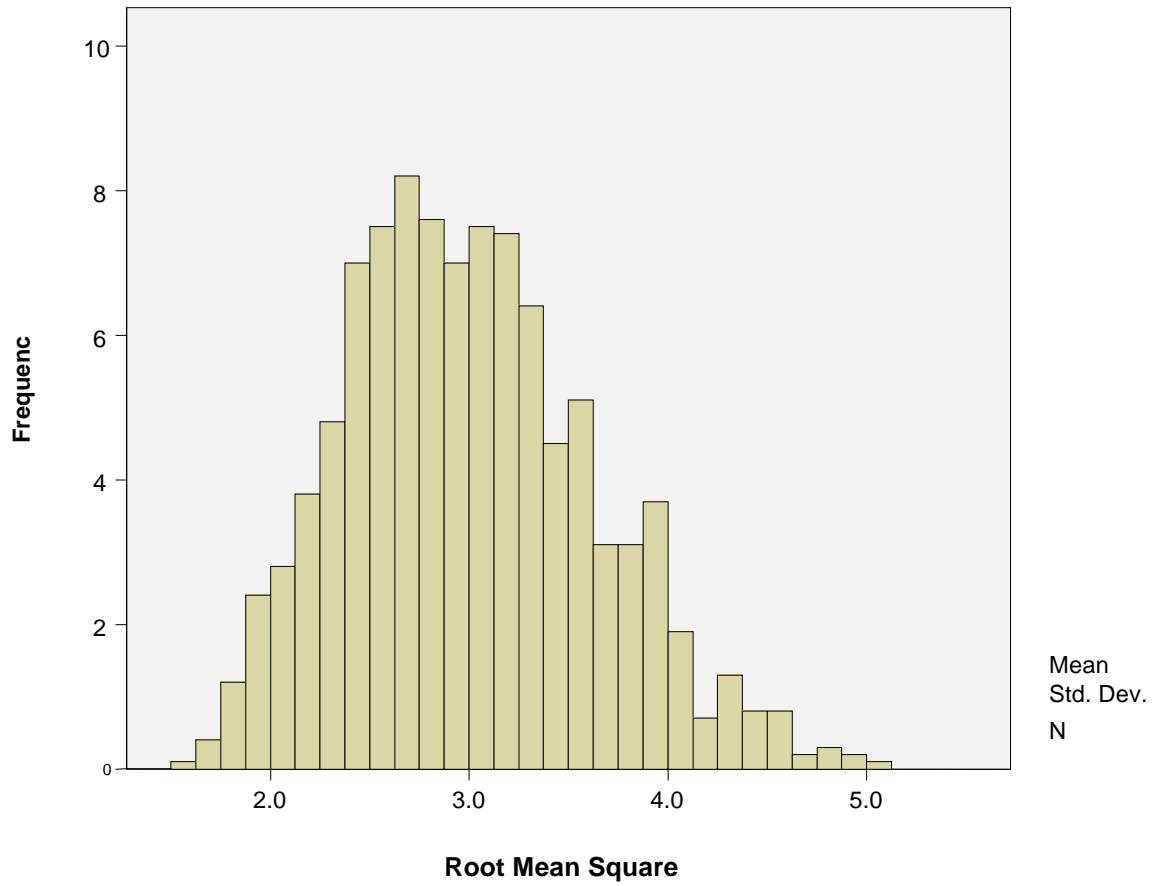


Continues next page

Figure D-1. Empirical Distribution for RMSD, Based on the Reading Assessment, for Four Sample Sizes (Continued)

$N=7,520$

Histogram –

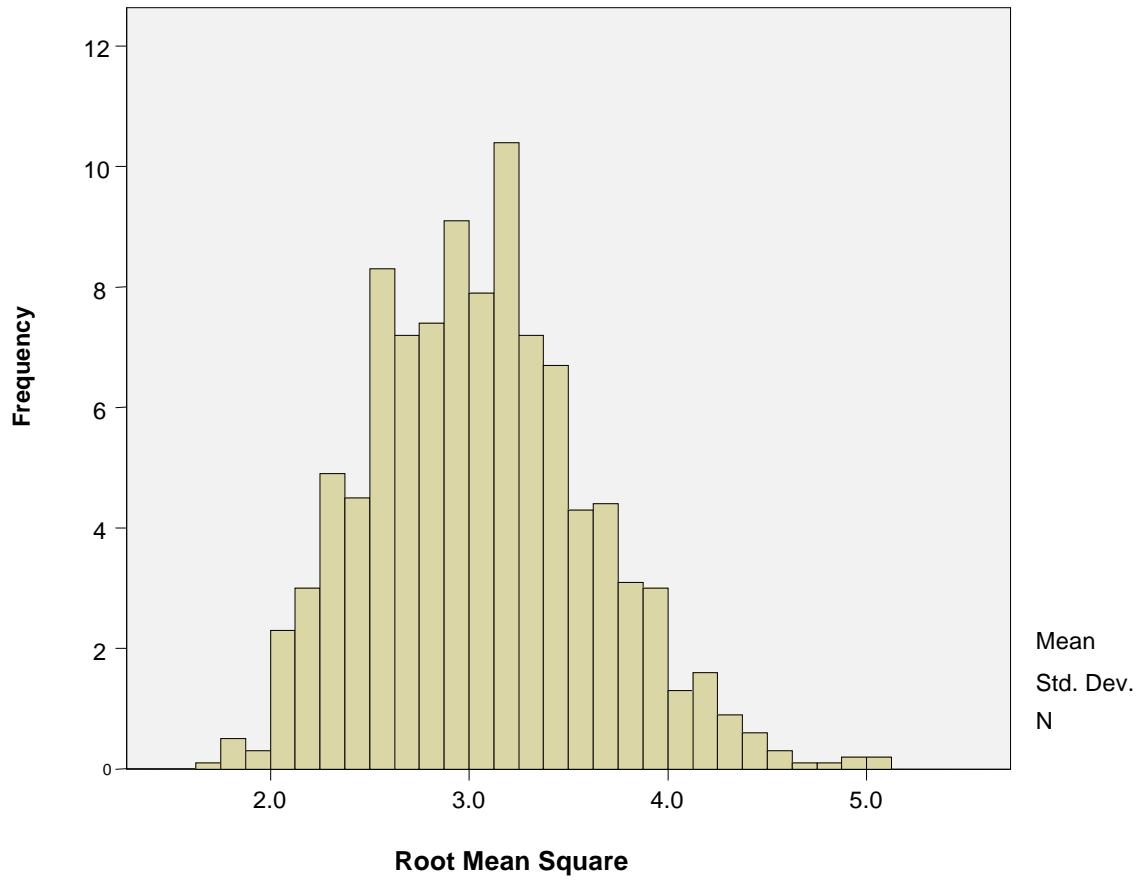


Continues next page

Figure D-1. Empirical Distribution for RMSD, Based on the Reading Assessment, for Four Sample Sizes (Continued)

$N=10,000$

Histogram – 10,000



This page left intentionally blank

Appendix E: Test Characteristic Curves for 2005 and 2003 Grade 8, Math Assessment

Figure E-1. TCCs for the 2005 and 2003, Grade 8, Math Assessment for the National Sample

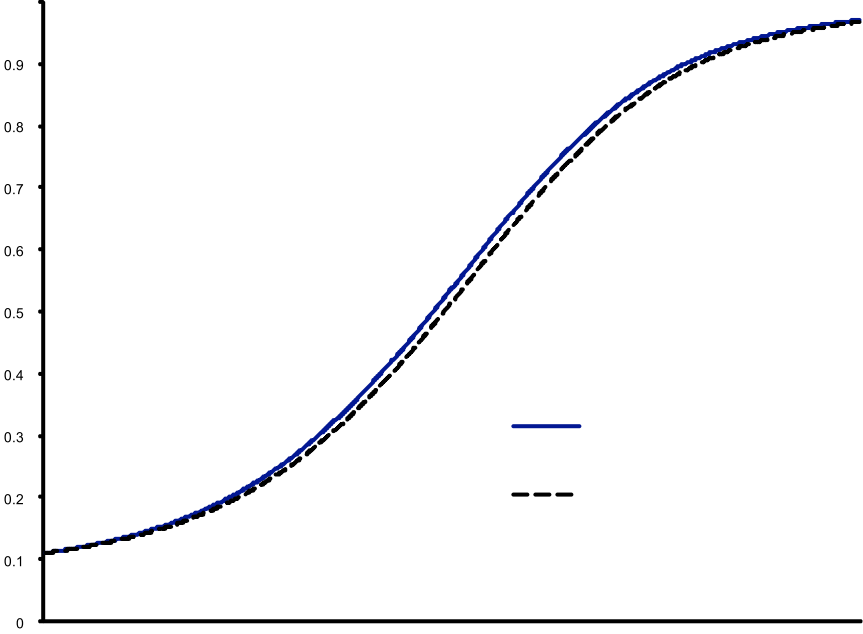


Figure E-2. TCCs for the 2005 and 2003, Grade 8, Math Assessment for Florida

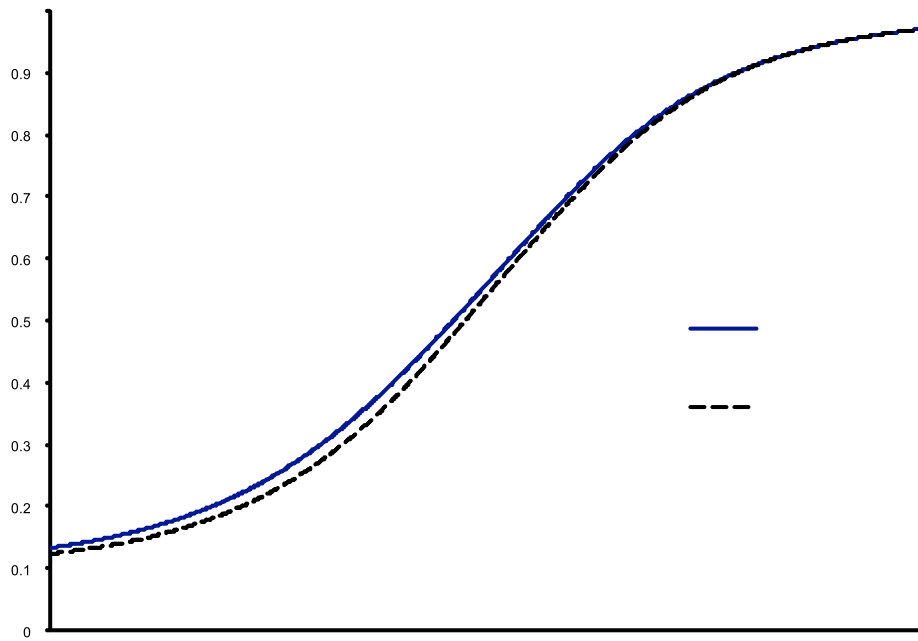


Figure E-3. TCCs for the 2005 and 2003, Grade 8, Math Assessment for Massachusetts

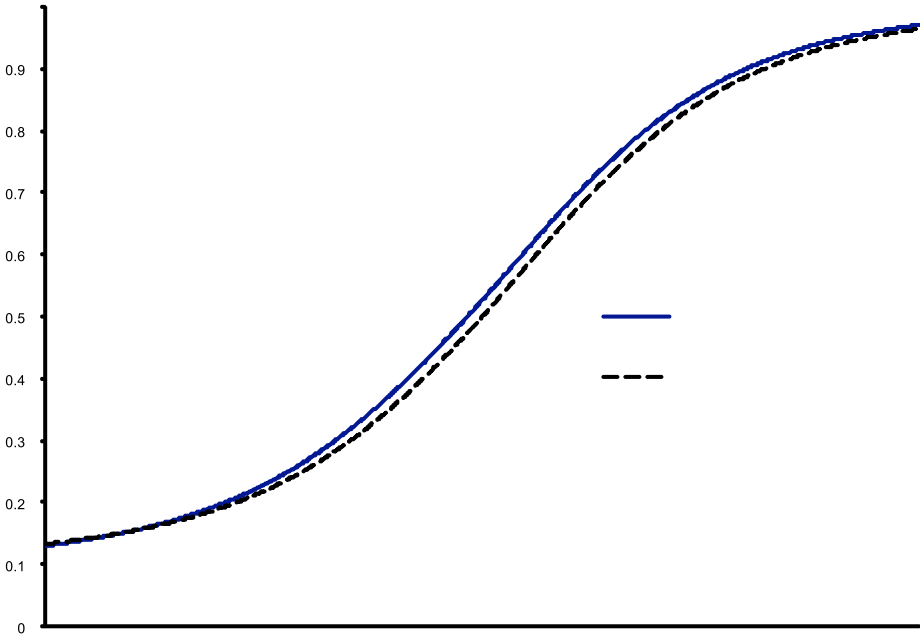


Figure E-4. TCCs for the 2005 and 2003, Grade 8, Math Assessment for California

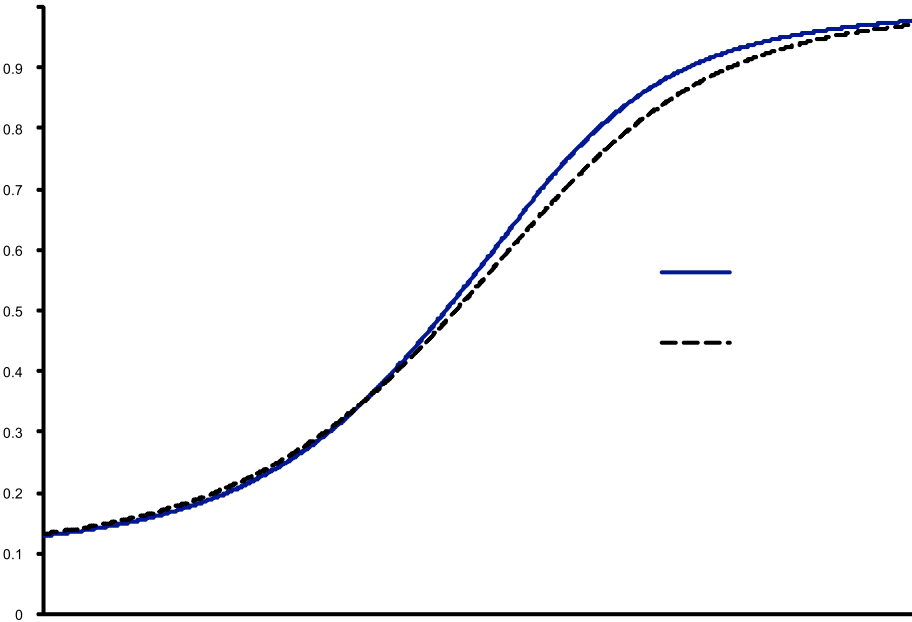


Figure E-5. TCCs for the 2005 and 2003, Grade 8, Math Assessment for North Carolina

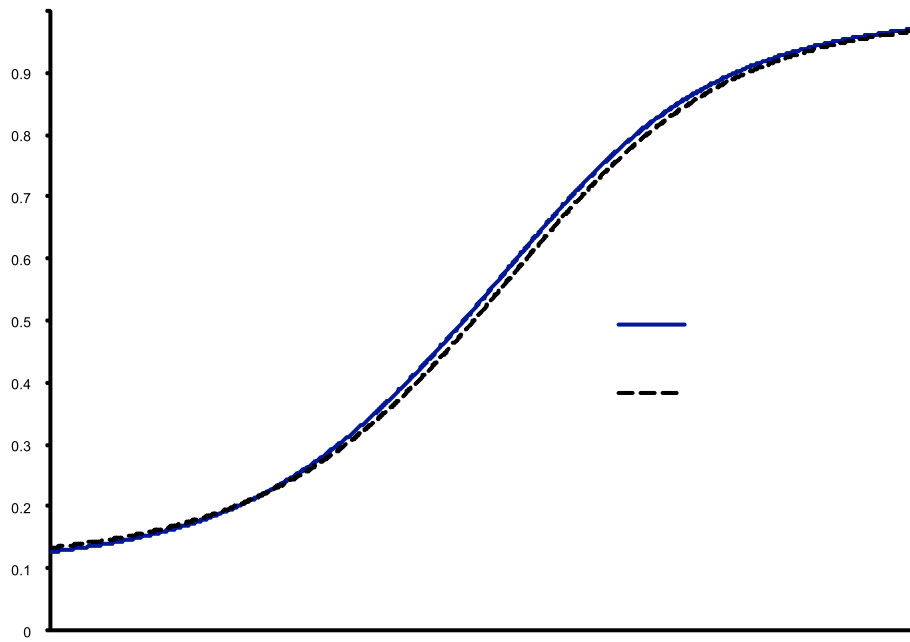
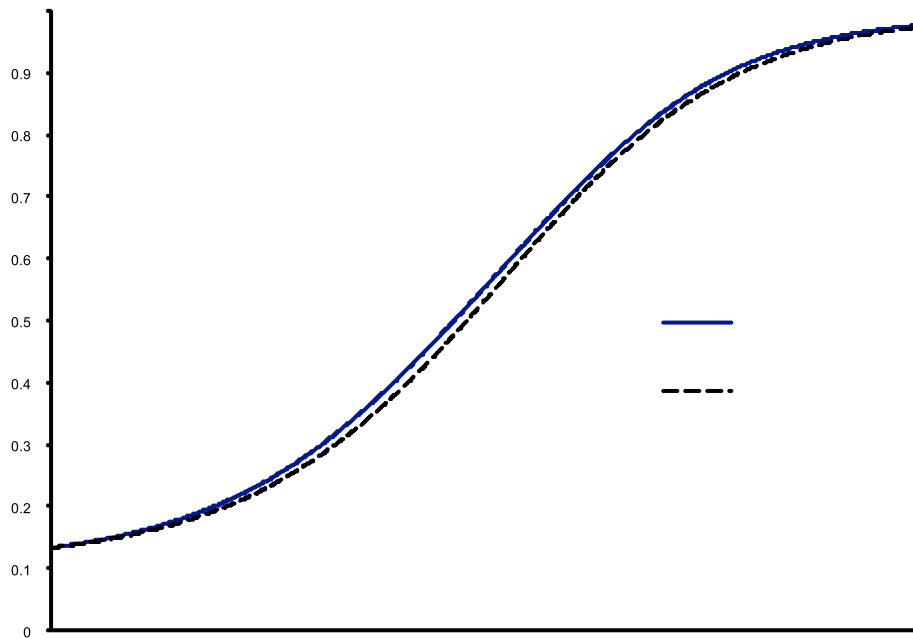


Figure E-6. TCCs for the 2005 and 2003, Grade 8, Math Assessment for Oklahoma



This page intentionally left blank

Appendix F: Test Characteristic Curves for 2005 and 2003, Grade 8, Reading Assessment

Figure F-1. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for the National Sample

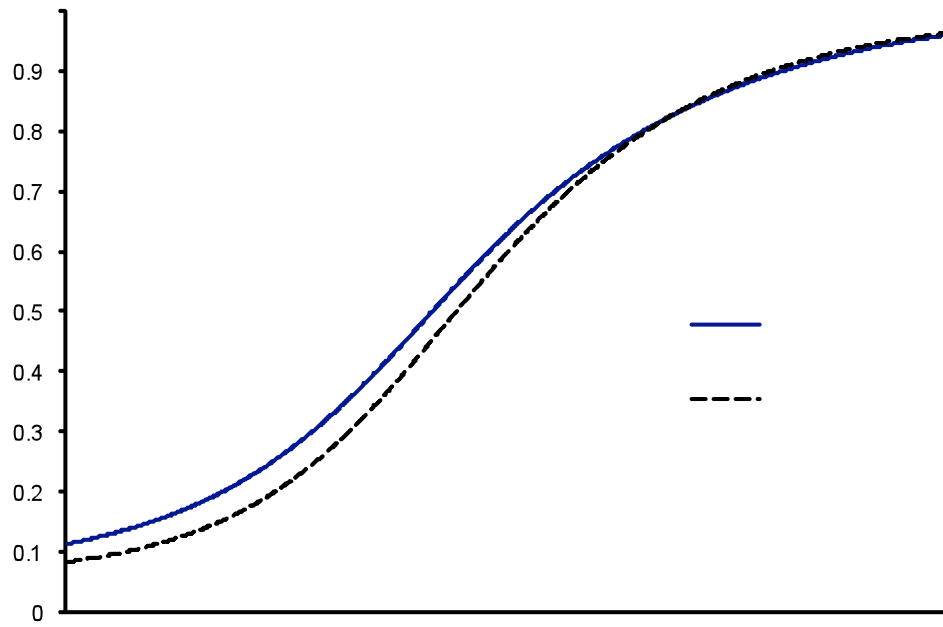


Figure F-2. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for California

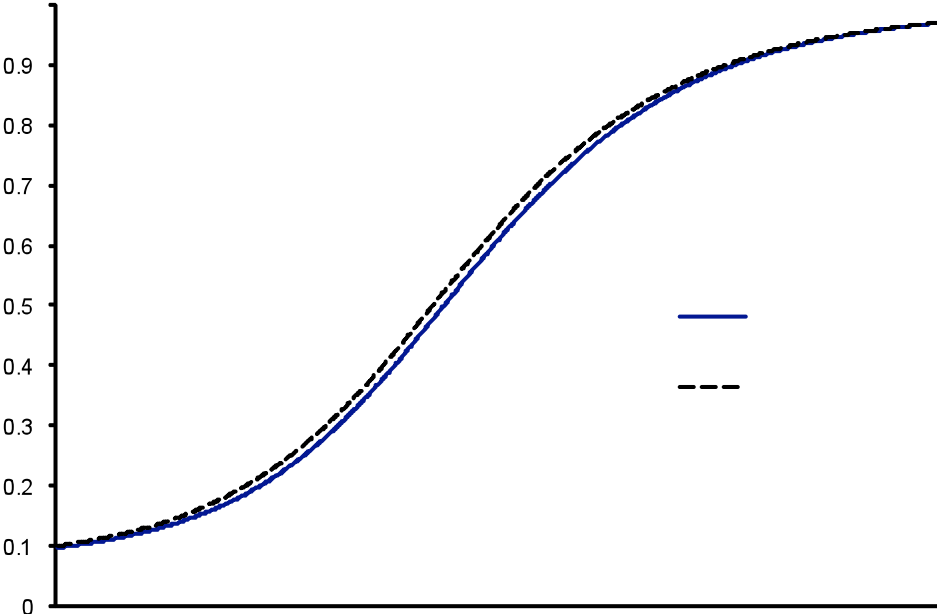


Figure F-3. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for New York

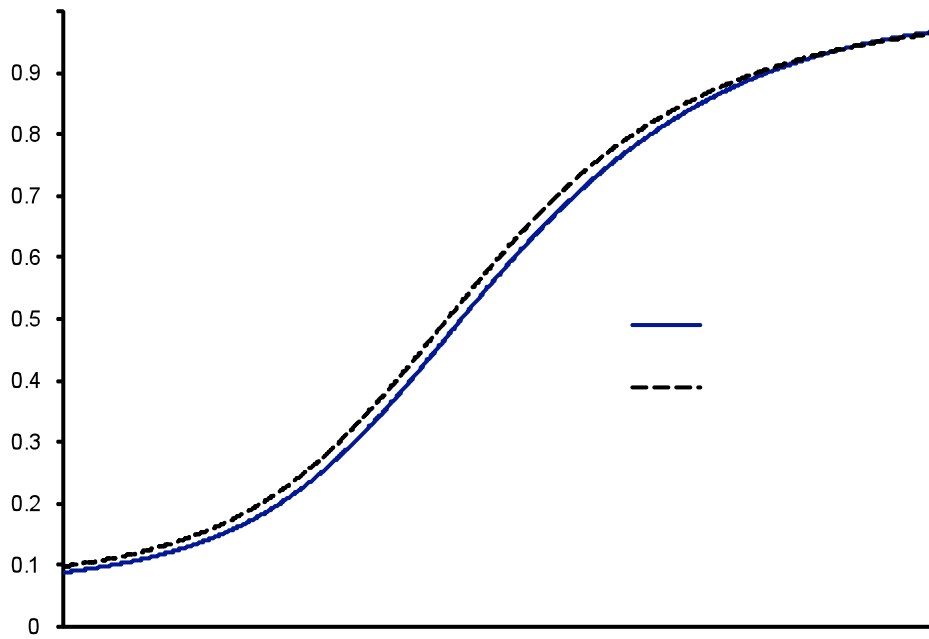


Figure F-4. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for North Carolina

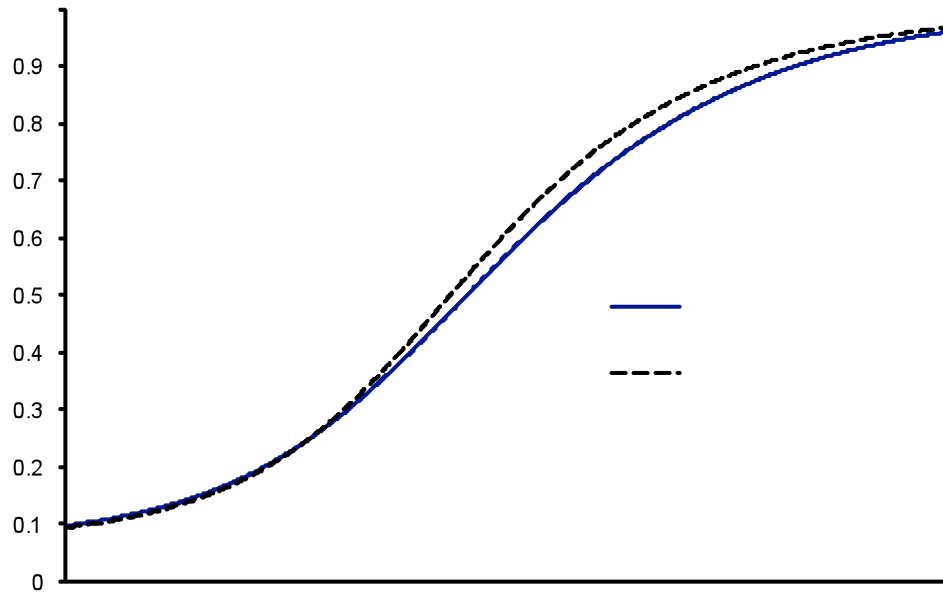


Figure F-5. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for Oklahoma

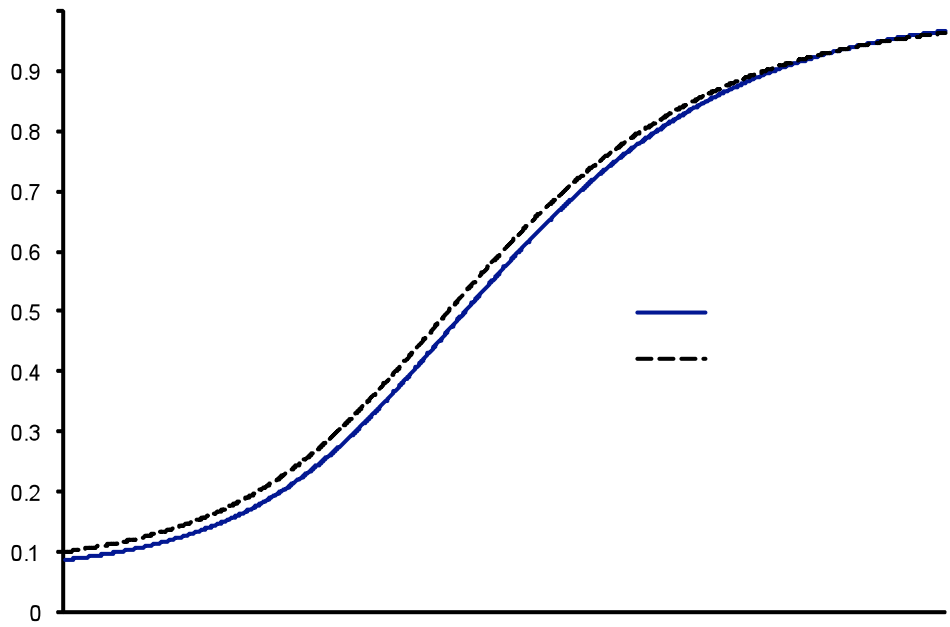
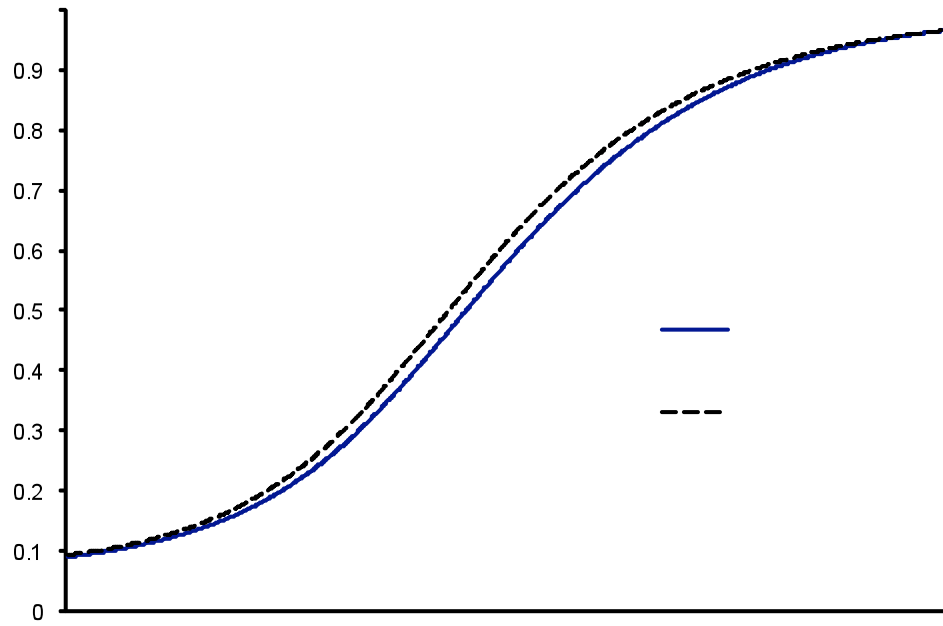


Figure F-6. TCCs for the 2005 and 2003, Grade 8, Reading Assessment for Texas



This page left intentionally blank

Chapter 6:
Methods for Evaluating the Alignment Between State Curriculum Frameworks and State Assessments: A Literature Review

Drey Martone, Stephen G. Sireci, and Jill Delton

Center for Educational Assessment

University of Massachusetts Amherst

This page intentionally left blank

Contents

List of Figures and Tables.....	6-v
Abstract.....	6-vii
Introduction.....	6-1
Overview of Alignment.....	6-5
The Relationship of Alignment to Content Validity.....	6-7
Approaches to Alignment Research.....	6-9
Webb Methodology.....	6-11
Achieve Methodology.....	6-17
Surveys of Enacted Curriculum (SEC) Methodology.....	6-21
Summary of Alignment Methodologies.....	6-25
Importance of Subject Matter Experts (SMEs).....	6-27
Alignment Methods Used by States.....	6-29
Conclusion.....	6-31
References.....	6-33
Appendix A: Sample of Webb Reports.....	6-37

This page intentionally left blank

Figures and Tables

Figures

Figure 1. Example of SEC Content Matrixes.....	6-44
Figure 2. Example of an SEC “Topographical” (Content) Map.....	6-45

Tables

Table 1. Categorical Concurrence Table.....	6-37
Table 2. Depth-of-Knowledge Consistency Table.....	6-38
Table 3. Range-of-Knowledge Correspondence and Balance of Representation....	6-39
Table 4. Summary of Attainment of Alignment Criteria.....	6-41
Table 5. A Comparison of the Three Most Popular Alignment Approaches.....	6-42

This page left intentionally blank

Abstract

In this paper, we (a) discuss the importance of alignment for facilitating proper assessment and instruction, (b) describe the three most common methods for evaluating the alignment between state curriculum frameworks and assessments, (c) discuss the relative strengths and limitations of these methods, (d) discuss examples of applications of each method, and (e) discuss which methods are being applied across the nation. We conclude that choice of alignment method depends on the specific goals of a state or district and that alignment research is critical for ensuring the curriculum-instruction-assessment cycle facilitates student learning. Additional benefits of alignment research include valuable professional development for teachers and better understanding of the results from standardized assessments. The implications of alignment research for understanding the results of the National Assessment of Educational Progress are also discussed.

This page intentionally left blank

Introduction

A great deal of discourse and debate exist, both professional and political, regarding state-mandated testing and testing under the *No Child Left Behind (NCLB)* legislation. The main criticisms of mandated testing in our nation's schools are reduced teaching time, a narrowed curriculum, and decreased morale of teachers and students (Sireci, Lewis, and Martone, 2006; Smith and Rottenberg, 1991). There is evidence, however, to support the view that mandated testing provides a necessary lens to view the educational opportunities presented to students. Without a means to understand what goes on in the classroom and a way to compare how students are performing, it is difficult to truly understand if all students are provided with adequate educational opportunities. Well-designed tests provide important data to learn about student performance and aid in decisions regarding funding (Cizek, 2001).

Although politicians, educators, and parents debate the merits of standardized testing, the psychometric characteristics of the tests are rarely the basis of concern. Rather, the main criticisms have focused on "opportunity to learn" issues such as failure to test students on what they are taught and a narrowing of the curriculum due to mandated testing. Ideally, to address such claims, researchers must demonstrate that what is covered on mandated tests supports what occurs in the classroom, both in terms of the curriculum and the instruction. *Alignment* research is one means to demonstrate the connection between testing, content standards (i.e., curriculum), and instruction. If these components work together to deliver a consistent message about what should be taught and assessed, students will have the opportunity to learn and to truly demonstrate what they have achieved.

The results of an alignment study can help policymakers, assessment developers, and educators make refinements so curriculum, assessment, and instruction support each other in what is expected of students. Alignment research has allowed the public to understand how testing does or does not support what is purported to occur in classrooms and what changes may be needed in components of educational systems.

The issue of alignment between tests and curricula has also been raised with respect to "Our Nation's Report Card," the National Assessment of Educational Progress (NAEP). Since 1969, NAEP has tested nationally representative samples of U.S. students to evaluate and track their performance in several subject areas. Since 1992, one way in which NAEP test results are reported is by the percentages of students in a state who fall within specific achievement level classifications. The reported NAEP achievement levels are Below Basic, Basic, Proficient, and Advanced. Note that Below Basic is an unofficial level. Under *NCLB*, states are required to set at least three performance standards on their tests, one of which should be used as a designation of "proficient." Although states are not required or encouraged to align their tests with NAEP assessments, an emerging use of NAEP results is to evaluate the rigor of standards across states, and to corroborate educational gains observed on state tests. In fact, Cross, Rebarber, and Torres (2004) claimed that state tests are "audited by the National Assessment of Educational Progress" (p. ii).

Given the current situation of both NAEP and state testing, whenever state test results portray a different picture of students' performance than that portrayed by NAEP results, misalignment between the state curriculum and the NAEP test frameworks (or between the NAEP and state tests themselves), is offered as a possible explanation (Fuller, Gesicki, Kang, and Wright, 2006). Many researchers have pointed out that NAEP and state accountability systems have different goals and produce different results (e.g., Barth, 2006; Cohen, Seburn, Gushta, Chan, and Jiang, 2005; Linn, 2005). Nevertheless, the popular press has been very critical of state testing programs when state test results are better than the corresponding NAEP results (e.g., de Vise, 2005; Dillon, 2005), as have policy researchers (e.g., Cross et al., 2004; Fuller et al., 2006; School Matters, 2005). Therefore, the degree to which the content and achievement standards of state assessments are similar to NAEP assessments is likely to be of interest to state and national policymakers as they evaluate test results at the national and state levels.

Importance of Alignment Research

Alignment research has resulted in multiple positive outcomes. First, like traditional studies of content validity, alignment studies provide important evidence that can support the validity of test score interpretations (Le Marca, 2001). Second, alignment studies have helped to better understand the number and frequency of content standards currently being assessed and help determine changes that need to be made in future assessments (Ananda, 2003a; Le Marca, 2001; Webb, 1997). In so doing they address the complaint that large-scale assessments result in a narrowed curriculum. Third, alignment studies have also been used as a legal defense to demonstrate that students are assessed on what they are given an “opportunity to learn” (Phillips, 2000; Webb, 1997) and to compare the assessment approaches among states or districts (Ananda, 2003a). Fourth, alignment research benefits teachers when they see the connection between classroom instruction and assessments (Webb, 1997), and this research has served as professional development for teachers (Porter and Smithson, 2001). Fifth, alignment studies inform future item writing activities (Ananda, 2003a), which helps test developers and provides another form of professional development for teachers, whenever they are involved in the item writing or item review processes. Sixth, states have used the results of alignment research to inform local planning and decision-making with respect to establishing a baseline to measure future progress (Porter and Smithson, 2001).

Clearly, alignment studies have produced positive outcomes across multiple levels of educational systems and have allowed all components of such systems to work toward similar goals to improve student achievement. As Norman Webb, a pioneer of alignment research, stated, “Better aligned goals and measures of attainment of these goals will increase the likelihood that multiple components of any district or state education system are working towards the same ends” (1997, p. 2). Beyond just the alignment of standards and assessments, the instructional content delivered to the students also needs to be in agreement. If this is not the case, if teachers are teaching what they want irrespective of what the curriculum calls for, students could potentially do well in the classroom and then fail on the assessments without understanding where they need additional help (McGehee and Griffith, 2001). Through alignment research, policymakers and educators involved in the educational process can see where they are headed, and will know where they stand relative to an agreed upon goal.

Purpose of this Paper

Our review was funded as part of a congressionally mandated evaluation of NAEP, but it should be understood from the outset that the test frameworks for NAEP are not intended to set a national curriculum and that state assessments are not designed to be aligned with NAEP. Thus, this review is not a review of NAEP-state test alignment, but rather a review of alignment methods developed to evaluate the alignment of state assessments to state curriculum frameworks. This review will enable consumers of NAEP results to better understand how states are building curricular validity into their assessments via alignment methodology. Knowledge of the different alignment methods used by states may be of interest to the U.S. Department of Education if it decides to formally look at NAEP-state test alignment.

Our review focuses on the use of alignment methodology to facilitate strong links among curriculum standards, instruction, and assessment. The purpose of our review is to describe why an understanding of alignment is an important characteristic of a statewide testing process and how undertaking alignment research can be beneficial both to the participants in the process and to the consumers of the results. Our review is structured around three areas of discussion. First, we present an overview of how alignment is defined in the educational measurement literature. This overview includes formal definitions of alignment and describes how alignment builds on earlier notions of content validity. In the second section, we describe the three most widely used alignment evaluation methods. While these methods share some common components, a closer

look at each approach highlights the relative strengths and limitations of each method. We also provide examples of specific applications of each of these methodologies. The final section discusses the alignment methods used by specific states, including studies that investigated the alignment of NAEP and state tests.

This page left intentionally blank

Overview of Alignment

Alignment means many things in the educational world. A Webster's dictionary definition states that to align is "to bring into a straight-line; to bring parts or components into proper coordination; to bring into agreement, close cooperation" (Le Marca, Redfield, Winter, and Despriet, 2000, p. 1). In a classroom setting, instructional alignment refers to agreement between a teacher's objectives, activities, and assessments so they are mutually supportive (Tyler, 1949). On a schoolwide level, curricular alignment refers to the degree to which the curriculum across the grades builds and supports what is learned in earlier grades (Tyler, 1949). Alignment, as described by the authors in this review, takes curricular alignment a step further to look at "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (Webb, 1997, p. 4). LeMarca et al. (2000) presented a more comprehensive definition of alignment:

Alignment is defined here as the degree to which assessments yield results that provide accurate information about student performance regarding academic content standards at the desired level of detail, to meet the purposes of the assessment system. To satisfy this definition, the assessment must adequately cover the content standards with the appropriate depth, reflect the emphasis of the content standards, provide scores that cover the range of performance standards, allow all students an opportunity to demonstrate their proficiency, and be reported in a manner that clearly conveys student proficiency as it relates to the content standards (p. 24).

In a perfect world, what a student is tested on should be derived from what is expected of the student as detailed in the school or district curriculum, as well as what is taught to the student by his or her teachers. While not everything that is listed in the curriculum or taught to the student can or should be assessed, alignment research has illuminated how much and to what degree the curriculum coverage or instructional content has been assessed. An understanding of alignment dimensions is sometimes used at the outset to create curriculum frameworks and assessments that are aligned from their inception (Rothman, 2003). The results of alignment research have been used in conjunction with the priorities determined by educational stakeholders to meaningfully inform future educational decisions.

The theory underlying alignment research is that a consistent message from all aspects of the educational structure will result in systematic, standards-based reform in which:

An instructional system is to be driven by content standards, which are translated into assessments, curriculum materials, and professional development, which are all, in turn, tightly aligned to the content standards. The hypothesis is that a coherent message of desired content will influence teachers' decisions about what to teach, and teachers' decisions, in turn, will translate into their instructional practice and ultimately into student learning of the desired content (Porter, 2002, p. 5).

Assessments, standards, and instruction are all integral to student achievement, but they have each been determined and enacted at multiple levels of the educational structure. Curriculum frameworks represent policy documents, but sources outside the policymakers created the assessments, and the curriculum and assessments are implemented locally in the educational setting. Alignment studies allow researchers to systematically study the different components of the educational structure as a means to compare their content and make judgments about the adequacy of the match.

Webb noted that the *Goals 2000: Educate America Act* supported the development of a consistent message about student learning among the policy, assessment, and instruction

perspectives. As he put it, this act “indicated alignment of curriculum, instruction, professional development, and assessments as a key performance indicator for states, districts, and schools striving to meet challenging standards” (Webb, 1997, p. 1). Additionally, *NCLB* requires that a state’s academic achievement standards be aligned with the state’s academic content standards. If the alignment between academic achievement and content standards is low, a state is likely to have trouble meeting the requirements of *NCLB*. Alignment research culminates in a report about the relationships of the components that can be used for future decision-making rather than just a simple yes or no response (Rothman, Slattery, Vranek, and Resnick, 2002). The results of alignment research provide a measure of how well assessments cover the underlying curriculum. Some alignment approaches also provide information regarding the degree to which assessments and curriculum match classroom instruction. Once the degree of alignment is understood, subsequent changes in any of the educational components can be made to improve the curriculum-instruction-assessment cycle.

In summary, alignment studies provide data that can be combined with the priorities of educational stakeholders to guide changes in assessments, curriculum, and instruction. By focusing on the match between test content and what is intended to be taught, alignment research shares some common goals and methodology with traditional methods for studying content validity. In the next section, we discuss some similarities between contemporary evaluations of alignment and traditional studies of content validity.

The Relationship of Alignment to Content Validity

Generally defined, *content validity* refers to the degree to which a test appropriately represents the content domain it is intended to measure. When a test is judged to have high content validity, its content is considered to be congruent with the testing purpose and with prevailing notions of the subject matter tested. Thus, content validity does not specify particular aspects of the educational process such as curriculum frameworks or instruction. Rather, it is more general and refers to tests both within and outside educational systems (e.g., licensure and certification tests).

There are at least four aspects to content validity—domain definition, domain representation, domain relevance, and appropriateness of the test construction procedures (Sireci, 1998a, 1998b). *Domain definition* refers to the process used to operationally define the content domain tested. In the case of K–12 achievement testing, the domain is typically derived from state-established curriculum frameworks. *Domain representation* refers to the degree to which a test represents and adequately measures all facets of the intended content domain. To evaluate domain representation, inspection of all the items and tasks on a test must be undertaken. Studies of domain representation typically use subject matter experts (e.g., teachers) to scrutinize test items and judge the degree to which they are congruent with the test specifications (Crocker, Miller, and Franks, 1989; Sireci, 1998a). *Domain relevance* addresses the extent to which each item on a test is relevant to the domain tested. An item may be considered to measure an important aspect of a content domain and so it would receive high ratings with respect to domain representation. However, if it were only tangentially related to the domain, it would receive low ratings with respect to relevance. *Appropriateness of test development procedures* refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material. The content validity of a test can be supported if there are strong quality control procedures in place during test development, and if there is a strong rationale for the specific item formats used on the test.

Traditional studies of content validity typically use subject matter experts (SMEs) to rate test items with respect to their congruence to the test specifications or their relevance to the intended domain. Hence, traditional content validity studies and contemporary alignment studies are similar in that they both gather data from SMEs, and structure the data collection procedure in a way that independently evaluates specific aspects of content domain representation.

Sireci, Robin, Meara, Rogers, and Swaminathan (2000) provided an example of a traditional content validity approach to alignment using the Grade 8 1996 NAEP Science Assessment. A primary goal of their study was to evaluate the congruence between the NAEP Science Framework and the NAEP Science Assessment. Ten carefully selected SMEs reviewed a sample of NAEP Science items and were asked to assign each item to (a) one of the three content areas (“fields of science”), (b) one of the three cognitive levels (“ways of knowing and doing science”), and (c) one of the four “themes of science” listed in the NAEP test specifications (framework). Each item was given an item congruence index rating based on the number of raters who agreed with the original classification. For example, if an item was intended to measure Earth Science and 8 out of 10 SMEs rated it as Earth Science, it had an item-content area congruence rating of 0.8. An index of 0.7 and greater was used to judge an item as adequately congruent with its content area, cognitive level, or theme. (Sireci, 1998a, provides examples of traditional and innovative content validity studies in several other contexts.)

While the traditional content validity approach involves rating or matching items to more global levels within test specifications (such as “domains,” “strands,” or “content areas”), contemporary alignment research uses the same expert rating approach but delves deeper to examine the match between items and the objectives or benchmarks *within* a strand. For example, a state’s curriculum framework may have the strand Grade 4 Number Sense (4N), which is the level at which test specification tables are typically written. However, within strand

4N there are multiple objectives. For example 4N-1.1 might be “Read, write, order and compare numbers up to 1,000,000.” In this example, the objective provides the detail regarding the specific skill being measured by an item. Alignment research often matches items to these detailed objectives and then reports findings summarized by objective. In fact, some alignment approaches do not provide summaries of the alignment at the strand level. However, in some cases alignment research considers what was actually taught to the students. In this way, alignment research can offer a deeper view of the educational process, and can be thought of as an extension of a more traditional content validity evaluation. However, as we discuss later, traditional content validity studies may have some advantages for evaluating the congruence of a particular test form to its test specifications.

Valid assessment requires significant overlap between the assessment and the desired curriculum to ensure decisions made based on test results are defensible. Alignment research is related to validity, but there is an important distinction that Webb (1997) highlighted: “Validity refers to the appropriateness of inferences made from information produced by an assessment (Cronbach, 1971). Alignment refers to how well all policy elements in a system work together to guide instruction and, ultimately, student learning” (p. 4). Alignment research has been most closely associated with content and consequential validity as a means to provide for a common understanding of what students should learn as a guide for instruction and to ensure equity for all students (Bhola, Impara, and Buckendahl, 2003; Webb, 1997). While alignment research examines how well several aspects of the educational system work together to impact student learning, validity research focuses on the appropriateness of the interpretations made from the results of the assessment. Thus, alignment research is an example of a validity study that supports test score interpretations.

Building on content validity studies, alignment research has helped various state departments of education to systematically compare what has been listed in the standards to what has been tested. In Webb’s (1997) work he found:

Most states’ frameworks and assessments were judged to be aligned if goals and learning objectives were considered in the design or selection of the assessment instruments. Most states lacked a formal and systematic process for determining the alignment among standards, frameworks, and assessments (p. 8).

Alignment research addresses states’ potential deficiencies by systematically comparing the different pieces of the educational process. If educational components are not well aligned, the system will not send a consistent message about what is valued in the educational process (Webb, 1999). Thus, alignment research can be used to evaluate concerns that the curriculum has been dumbed down (Linn, 2000), that students have not received a fair chance to learn the material on which they were tested (Winfield, 1993), and that states have not addressed the need to improve instructional quality (Rothman et al., 2002). Contrariwise, traditional studies of content validity have focused on the match of test items to the domains specified in a test blueprint.

Approaches to Alignment Research

In the previous sections, we defined alignment, related it to content validity, and described the importance of conducting alignment research to the educational process. In this section, we describe three contemporary alignment methods.

The development and application of alignment methods came about from a desire to ensure that the scores students receive on an assessment reflect their performance with respect to specific curricular expectations (Le Marca, 2001). Some alignment studies have focused on the content of the standards compared to the assessments while others have included the content of instruction as an additional variable. The following section elaborates on the three most common methods for alignment research—the Webb, Achieve, and Surveys of Enacted Curriculum methods. An application of each of these methodologies is also presented to illustrate their processes and findings. Throughout this section points of comparison among the three approaches are highlighted.

This page left intentionally blank

Webb Methodology

Norman Webb developed a comprehensive and complex methodology to investigate the degree of alignment between assessments and standards. His method explores five different dimensions to understand the degree of alignment: content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability (Webb, 1997). In this method, “standards” are the broad content domains within a subject and the skills within this domain are referred to as “objectives.” Understanding these definitional terms is critical to seeing how the alignment process has been applied, because these terms and levels of analyses differ across the different alignment methods.

Alignment Dimensions

Content focus

Webb’s content focus dimension comprises six subcategories for analysis: categorical concurrence, depth of knowledge, range of knowledge, balance of representation, structure of knowledge, and dispositional consonance. Each of these subcategories explores the relationship between the assessment and the standards in a different way. Together they contribute to a thorough understanding of the degree of alignment between assessments and standards.

Categorical concurrence compares the similarity of the expectations for student learning, as expressed through the content categories in the standards, to the assessments. This subcategory is most similar to traditional content validity and is a minimum requirement in alignment research. Like the test blueprint comparison in a traditional content validity study, categorical concurrence looks at broad content areas, such as Number Sense and Geometry. To have alignment relative to this dimension, an assessment must have at least six items measuring a standard. Using this approach, if there are four standards, an assessment needs at least 24 items to establish categorical concurrence. However, unlike a traditional content validity study in which a test item is matched to its standard by SME consensus (e.g., 70 percent of SMEs match an item to its intended standard), Webb’s criterion is simply that, across the SMEs, an average of at least six items is matched to the standard. That is, a standard could theoretically be considered adequately represented, even if the six items matched to it were specified to measure a *different* standard in the test blueprint. Popham (1992) and Sireci (1998a) suggested the use of seven out of ten SMEs correctly matching an item to its intended standard as a criterion for a congruent item-test specification match.

Depth-of-knowledge consistency compares the level of cognitive demand expressed in the specific objectives within each standard to the cognitive demand in each item that is matched to that objective. Webb initially defined the cognitive areas as recall, skill or concept, strategic thinking, and extended thinking, but these areas may be modified for a particular study (Webb, 1999). The main criterion is that what is tested should be at the same cognitive level as what is expected to be taught. To have alignment relative to this criterion, at least 50 percent of the items matched to an objective must be at or above the cognitive level of that objective. Fifty percent is based on the assumption that most cutoff points require students to answer more than half the items to pass, but some flexibility is allowed with this criterion. The main concern in this aspect of alignment is that assessment items should not be targeting skills that are below those required by the objectives.

Range-of-knowledge consistency analyzes the breadth of the standards as compared to the breadth of an assessment. This dimension looks at the number of objectives within a standard measured by at least one assessment item. To have sufficient alignment relative to range of knowledge, at least 50 percent of the objectives within a standard need to be measured by at least one assessment item. This assumes that students should be tested on at least half of the domain of knowledge. This part of the alignment process also assumes all of the objectives have equal weighting and all of the objectives accurately cover the skills needed to complete that standard.

The level of complexity within a state’s standards influences this aspect of alignment as more complexly written objectives might only be partially assessed but would still be considered a match from the perspective of this dimension.

Balance of representation focuses on the degree to which items are evenly distributed across objectives within a standard to represent the breadth and depth of the standards. Given the limited time for assessment, this dimension highlights what aspects of the standards are prioritized. Balance of representation focuses on the objectives assessed by the items and then looks at the proportion of objectives measured compared to the number of items. The goal is to measure every objective assessed with at least two items. The calculation for the balance index is:

$$1 - \frac{\left(\sum_{k=1}^0 \left| \frac{1}{O} - \frac{I_k}{H} \right| \right)}{2}, \quad [1]$$

where O=Total number of objectives hit for the subject domain; $I_{(k)}$ = Number of items corresponding to objective (k); and H = Total number of items hit for the subject domain (Roach, Elliott, and Webb, 2005). If the proportion approaches 0, it signifies many items are assessed by only a small number of objectives. If it approaches one it signifies that the items are evenly distributed across all objectives. Ideally, over time, assessments should shift in the balance of representation to cover the entire standards. Thus, evaluating the specific standards covered over time is necessary to ensure important standards are not being neglected. Evaluating balance of representation across grades can also demonstrate shifts in priorities as the content develops.

These first four areas of Webb’s content focus dimension—categorical concurrence, depth of knowledge, range of knowledge, and balance of representation—are most often used by alignment researchers as the basis for their alignment methodologies. These four aspects serve as the most direct way to view the degree of match between an assessment and the standards. The last two aspects of the content focus dimension—structure of knowledge and dispositional consonance—have *not* been applied in a research study (that we found), but they illustrate the potential complexity of the alignment process.

Structure of knowledge analyzes to what degree the assessment items target the broader goals of instruction. For example, if the goal is for students to have an integrated understanding of a concept, this variable examines to what degree the assessment is only targeting isolated skills. Webb emphasized that this might best be analyzed in the context of the broader assessment system.

Dispositional consonance is another view of structure of knowledge in that it assesses the degree to which the assessments support the broader stakeholder beliefs about education. For example, in the standards it may state that it is important that students be able to critique their own work. This skill is easier to assess in non-standardized settings and highlights the need for alignment studies to include the broader assessment policies of an educational setting. This aspect of content focus would address concerns about “narrowing of the curriculum.”

Webb’s approach to alignment also takes a broader look at the context within which assessment and instruction occur. Because these dimensions have not been applied in a detailed study, only a brief overview is provided here. Webb addresses the issue of *articulation across grades and ages*. It is important to analyze the change in content across grades and ages to highlight the content and cognitive complexity in standards. Webb believed that assessments should be developed with an understanding of how students change through the years and how this change can be assessed at different stages of development. *Cognitive soundness* is one aspect of articulation across grades and looks at how the cognitive complexity increases as students move through levels of understanding connecting new ideas to existing ideas. *Cumulative growth* in content knowledge during schooling is another aspect of articulation across grades and relates to the idea that students start with basic ideas and build on those

through schooling. While theoretically these are important pieces of the alignment puzzle, these topics have not been included in applied alignment research to date although they are important issues that are included in approaches to vertical alignment.

Webb also highlights issues of *equity and fairness* in his general approach to alignment. Equity and fairness is a means to ensure high standards are set and every student is given the opportunity to demonstrate understanding.

The area of *pedagogical implication* is also included in Webb's general approach to alignment. Pedagogical implications focus on teachers' interpretations of the expectations in the standards and the assessments and how their instruction fits within these expectations. At times teachers may think they are addressing the curriculum, but in reality they might be only superficially meeting the broader expectations (Cohen, 1990).

Finally, the system applicability dimension of Webb's original methodology examines the degree to which classroom instruction relates to real world needs. That is, this aspect evaluates whether instruction focuses on important skills needed to succeed in a global economy rather than on just basic skills.

Application of Webb's Method

Webb (1999) applied his methodology in a study of mathematics and science assessments and standards in four states. Here, we focus on the mathematics alignment process and results. The purpose of Webb's study was to better understand how his alignment methodology functioned, to examine in greater detail the different alignment dimensions, and to understand ways to improve the alignment process. In this study, six reviewers compared the match between assessment items and standards or objectives in mathematics. The results of this matching were used to judge the degree of alignment based on four of Webb's criteria: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge consistency, and balance of representation.

The review process involved multiple decision points by the reviewers. Applying this process across multiple states, the reviewers noted differences among the standards in terms of content covered, level of detail for the standards, and the overall organization of the standards, which impacted the comparability of the states. The first step was a review of each state's standards to match each objective to a depth-of-knowledge level representative of the highest level of knowledge needed to achieve that objective. This process allowed for systematically linking items to objectives and cognitive levels. The reviewers reached an agreement about the depth-of-knowledge of the objectives based on a group discussion. These decisions were used as a baseline comparison to the assessment items to determine if the items were at or above the cognitive level in the objective.

The items within an assessment were then matched to the objectives within the standards and coded based on the depth-of-knowledge required by that item. Any match was called a "hit," however, one item could be matched to more than one objective. This increased the content and range alignment criteria areas, but proved to be an area of confusion for the reviewers. The reviewers also noted when items appeared to not match any objective. The results were aggregated to report by standard. The mean and standard deviation for each criterion were computed for each reviewer.

The results showed varied levels of alignment across grade levels and states. The strongest area of alignment was the categorical concurrence criterion. Three out of the four states fulfilled this criterion with at least six items measuring a standard, but in each state one-fourth or more of the standards were measured by fewer than six items. The balance-of-representation criterion was satisfied because the standards that were assessed had items evenly distributed among the objectives.

The weakest aspects of the alignment method were the depth-of-knowledge consistency and range-of-knowledge criteria. The results demonstrated that test items generally targeted a lower level of knowledge and did not sufficiently cover the range of knowledge laid out in the

standards. This finding lends some support to the common criticism that standardized testing does not test complex thinking and narrows the curriculum by testing a small part of the content domain. Armed with the results of this alignment research, these states could accurately address these issues in their assessment design. This study also demonstrated that each of the four criteria measured different aspects of alignment.

Webb (1999) noted that the reviewers could have benefited from more training at the beginning of the process. Some reviewers wanted to code near matches instead of exact matches and this confused the analysis. The reviewers needed more guidance about making distinctions relative to the depth-of-knowledge criteria and more explicit guidance about how to match an item to more than one standard based on the central content of an item. Webb also found that it could be helpful to put the standards in context so the reviewers know each state's purpose for the standards and how they were created. During the review process, the reviewers focused purely on the objective-item match and did not have an opportunity to critique the quality of each component. Webb concluded the reviewers were frustrated by this constraint. While it is important to stay focused on the task at hand, it could be helpful to gather this feedback throughout the process as a means to inform future standard or assessment development work.

Webb (1999) concluded that tradeoffs between these four alignment variables are realistic, but it is important to look at broader approaches to assessment to understand how other pieces (e.g., those discussed in the general Webb methodology but not specifically studied in his alignment process) complement the process. Unfortunately, these aspects are harder to measure and include in a formal study, and may involve validity issues that go beyond alignment per se. One limitation of Webb's method is that the range of knowledge criterion does not look at the breadth of the measured objective in terms of how many different ideas are combined under one objective. If an objective were very broadly stated, it was still considered assessed if it had an item matched to it, regardless of what else within that objective was not assessed. With objectives that combine many different ideas, possibly with different cognitive expectations, it was easier to satisfy the range-of-knowledge criterion, but this may result in a lower depth-of-knowledge as the complexity of the objective might have increased. The interplay between the alignment dimensions illustrates the benefit of using the alignment results to inform the development of both standards and assessments. Furthermore, the knowledge of these alignment criteria is being used to guide item development to ensure items meet a cognitive requirement and address a range of objectives within each standard.

Another limitation with the Webb methodology was that it did not capture the fact that assessments may purposefully contain items to measure standards from more than one grade. This misalignment by design should be carefully detailed in the alignment process. In looking at the alignment study process, Webb (1999) developed a number of recommendations. If the goal were to analyze standards from more than one state, Webb recommended starting with the most detailed state standards. It would be helpful to repeat the alignment study over time to capture the changing content of the assessment and how this may or may not impact the alignment results. Additionally, Webb recently noted (Webb, Herman, and Webb, 2006) that averaging reviewers' ratings across standards and objectives might mask the different views of what the item is truly measuring and inflate the degree of alignment across the four dimensions. Recent studies (Herman, Webb, and Zuniga, 2005; Webb, Herman, and Webb, 2006) have examined setting a minimum reviewer agreement requirement at the standard or objective level as to what the item is measuring, but this analysis is still ongoing.

The Webb alignment dimensions have also recently been applied to the issue of vertical scaling. Wise and Alt (2005) discussed the possible steps to vertically align content standards and then apply the Webb dimensions to examine how the standards address the skills across the grade levels. Wise, Zhang, Winter, Taylor, and Becker (2005) provide further in depth guidance about how the vertical alignment analysis could work in terms of types of judges, types of ratings, and how the ratings could be analyzed and reports. This is an interesting extension of the alignment discussion, but is still in its early stages.

Overall, the Webb model is comprehensive and provides a point of reference for the next two models reviewed. The strength of this model is its comprehensive analysis of the objective level detail, its view of alignment through four different dimensions, and the clear guidelines for what serves as acceptable levels of alignment. Another positive aspect of Webb's work is its recognition of a broader set of issues (e.g., articulation across grades, fairness, and pedagogical implications), even though measurement of these issues is not yet fully developed. Sample reports for the Webb methodology can be found in the Web Alignment Tool Training Manual (Webb, Alt, Ely, and Vesperman, 2005) and examples of these reports are produced in Appendix A. The results of a study using the Webb approach would illustrate the relationship between what is being asked of the students, how that is being assessed, and what trade-offs are made in the process.

This page left intentionally blank

Achieve Methodology

The Achieve methodology is detailed in an alignment protocol that is adapted to reflect the concerns of specific subject areas (English language arts, mathematics and science.) It yields both a quantitative and qualitative alignment comparison of a state's assessment to its related standards. Rothman, Slattery, Vranek, and Resnick (2002) laid out the components of the Achieve methodology, which is designed to judge the quality of the overall assessment, as well as the individual items that comprise the assessment. Since that time, Achieve's protocol has been further refined. The method is based on a team of carefully trained SMEs reaching consensus on the degree of match between the standards and the assessment based on specific criteria (dimensions). The five criteria in this methodology include: *content centrality*, *performance centrality*, *challenge*, *balance*, and *range*. In this methodology, "objectives" are defined as the most specific level of outcome (i.e., the smallest level of grain size used by a state in delineating its content standards).

Achieve Alignment Dimensions

Like the Webb methodology, the Achieve protocol compares individual items on an assessment to the related objectives and looks at the degree of content and performance match and the cognitive demand of the items, as compared to that stipulated in the objectives. This methodology also qualitatively considers how a set of items matched to an overarching standard (e.g., literary response or algebra) functions as a group. While potentially more time consuming than other approaches, these additional criteria provide a more thorough understanding of the degree of alignment.

The Achieve methodology is applied in two stages. The first stage is an item-by-item analysis to confirm the test blueprint, determine the content and performance "centrality" of each item compared to the objective to which it is matched, evaluate the source of challenge, and determine the level of cognitive demand. The second stage is a holistic evaluation of a set of items matched to an overarching standard in terms of the overall level of challenge, the balance and the range. The stages and steps within each stage are detailed below.

Stage one

Confirmation of the test blueprint. The first stage in the Achieve method focuses on item-level detail only and starts with a confirmation of the test blueprint. Items are compared to the objectives, defined at the most detailed level of outcome to ensure that every item is matched to at least one objective. A match between the test blueprint and the item requires only that the item address the *same* content; the level of cognitive demand or the associated objective is not considered. Items that are mapped inappropriately are reassigned to a more closely related objective, while items that do not match a standard or objective are eliminated from further analysis. When a state lacks a test blueprint or the blueprint does not allow for fruitful application of the protocol, Achieve constructs a blueprint. In these instances, Achieve provides a brief rationale and communicates the findings to the state. Achieve scrutinizes the test blueprint because of its importance in developing score reports. This level of analysis is missing in the Webb approach.

Each item can have a primary and a secondary match to the objectives. The primary match is used in judging content and performance centrality, source of challenge, and level of cognitive demand (described below). The secondary match is taken into account in evaluating level of challenge, balance, and range. The use of a secondary match is similar to the Webb method in which items could be mapped to more than one objective, but this model is more explicit about the degree of match and how it can be used in the alignment process. After the test blueprint has been confirmed, the reviewers delve deeper into the actual content of the item and how it specifically relates to the identified objective.

Content centrality. To judge content centrality, SMEs rate each item based on the degree of content match between the item and the objective it is measuring. The rating system uses a four-point scale in which a “2” is a clearly consistent content match; “1A” is a match in which the degree of alignment is unclear (generally because the standard is too broad to conclude that the item is clearly consistent with the objective); “1B” is a somewhat consistent match in that the item assesses only part of a compound objective; and “0” signifies an inconsistent match. This rating dimension addresses a limitation of the Webb (1999) study in which a broadly stated objective may be considered adequately measured even if the item only addressed a part of the standard.

Performance centrality. In considering performance centrality, the Achieve protocol focuses on the quality of the match between the performance called for in the item and the performance described by the objective the item is intended to measure. This is similar to Webb’s (1997) method, but in the Webb approach the cognitive level of the objectives is coded in the beginning and the performance rating is made simultaneously with the content rating. The Webb method might be more efficient, but the Achieve method allows the reviewers to focus on each aspect of the process in isolation. The performance centrality rating process calls reviewers’ attention to the verbs in the objectives as compared to what the items actually demands of the student. The same 2, 1A, 1B, 0 scoring system is used for this dimension.

Source of challenge. Source of challenge is measured to ensure that items are fairly constructed and not designed to trick students. The items are reviewed to ensure they are not technically flawed (from a content perspective and by reviewing results from item analyses). For example, mathematical items are reviewed to ensure the reading level is appropriate for the grade level of the assessment and unnecessary reading is not required, while reading items are examined to ensure they measure comprehension and not prior knowledge. Reading passages are reviewed to ensure that the vocabulary, sentence structure, literary techniques, plot line, and organizational structure are all appropriate based on the grade level of the assessment. Writing prompts are similarly reviewed for accessibility, appropriate vocabulary, clarity of purpose and audience, and inclusion of basic criteria by which the sample will be scored. Each assessment item is scored as 1 for an appropriate source of challenge and 0 for an inappropriate source of challenge. If the item received a 0 for content and performance centrality then it would receive a 0 for source of challenge, as it is not a good measure of that standard. Webb recently included Source of Challenge as one of his alignment dimensions, although it is captured only through reviewer comments (Webb, Alt, Ely, and Vesperman, 2005).

Level of cognitive demand. Level of cognitive demand is concerned with the kind and level of thinking required by students to respond to an item. The level of demand can stem from the nature of the concept assessed (some concepts are more readily understood than others) or from the kind of thinking required to arrive at a response (an item may demand routine or concrete thinking as opposed to complex reasoning or abstract thinking.) Achieve has refined the way in which it tracks the level of cognitive demand of individual items to better inform the evaluation of overall level of challenge. (J. Slattery, personal communication, Dec. 15, 2006). SMEs formally rate each item on a scale ranging from Level 1 (recall or basic comprehension) to Level 4 (extended analysis, typically over an extended period of time). Level 4 items are not usually found on large-scale, on-demand tests. The next stage in the application of the Achieve protocol shifts from a focus on individual items aligned to objectives to sets of related items aligned to a larger standard.

Stage two

Level of challenge. Level of challenge is a global judgment (not item specific) that qualitatively captures whether the collection of items mapped to a given overarching standard appropriately challenges students in a given grade level. Ideally, items within each standard should range from simple to more complex. SMEs provide a brief written evaluation of the level of challenge for each set of items tied to a specific standard, describing how the “overall

demand” compares to that expressed in the standard. They base their judgment, in part, on the level of cognitive demand scores previously assigned to individual items in the set. SMEs look to see if a set of items are skewed toward one level of demand, if they are focused only on the more demanding or least demanding objectives within a standard and, where there are compound objectives, if the items are skewed toward the most or least demanding part of the overall standard. The next step of the Achieve methodology examines the balance and range of sets of items relative to the expectations expressed in the standards.

Balance. Balance, like level of challenge, is a holistic evaluation. It looks at a set of items mapped to a given standard to determine how closely the set of items measures the breadth and depth of the content and performances expressed in the related standard. The relative importance the test items give to content and skills should be proportionately similar to what is stated in the standards. The SMEs comment on objectives within a standard that are over or under- assessed, redundant items, and how the overall set of items measures content they think is important for that level. The analysis allows the experts to focus on how they view the balance of the assessment as compared to the standards (Rothman, 2003). Again, this is captured qualitatively and builds on the expert knowledge of the SMEs, which is similar to Webb’s (1997) balance criterion, although that measure is quantitative. Webb’s balance calculation only determines if the objectives are equally represented, but that might not be meaningful if one area of the standards should be emphasized more through the assessment (Rothman, 2003). The quantitative measure facilitates comparison across states or districts, while the qualitative measure provides information more informative to the standards or assessment revision process.

Range. The range criterion also considers a set of items matched to a standard, but it measures the standard coverage. Range is a quantitative measure of the proportion of the objectives within a standard that are measured by at least one item. Ranges between 0.50 and 0.66 are acceptable and above 0.67 is considered good coverage. This is similar to Webb’s (1997) range calculation although his methodology uses 50 percent coverage criterion. It is possible for a test to be well balanced, but have low coverage (and vice versa) and so it is important to consider both of these criteria.

At the close of the alignment review, SMEs look across all of the overarching standards (i.e., at the assessment as a whole) to determine the overall rigor of the assessment and how closely it succeeds in measuring the content and performances described by the standards. When Achieve analyzes state assessments at multiple grade levels, SMEs comment on the comparative strengths and weaknesses of the assessment system taken as a whole. It is also the case that application of the Achieve protocol provides SMEs with insights regarding the quality of a state’s standards. For example, if a great many items are scored a “1A” for content centrality, it signals that many standards are written at too high a level of generality. Achieve transmits all its findings in a comprehensive, technical report to the state that is kept secure because it contains detailed commentary on actual test items. Achieve also produces a policy level report meant for the state to release publicly. Sample policy alignment reports can be found at www.Achieve.org.

An Application of the Achieve Model

Rothman et al. (2002) applied the Achieve methodology to the evaluation of assessments in five states. The process began with a training of expert reviewers. The reviewers represented a diversity of viewpoints and included classroom teachers, curriculum specialists, and content and assessment experts. They were trained through the use of carefully selected items to illustrate each of the rating criteria in the Achieve protocol.

Rothman et al. (2002) found that states with standards written in global terms received low ratings because it was more difficult to determine accurate item-standard matches. Overall, they found that items were well matched to content and performance standards. Most states also fared well with respect to the source of challenge criterion. However, they found that the states were not doing a sufficient job of assessing the full range of standards and objectives, and that the most challenging standards and objectives were under-sampled or omitted (similar to Webb,

1999). With respect to balance, they found that the sets of items were too focused on the less important standards, a finding that was also supported by the level of challenge results.

Rothman et al. (2002) emphasized the need to focus on the issues of balance and challenge in the design and selection of state assessments. Their study illustrated both the drawbacks and strengths of the Achieve alignment method—the process can be time consuming and expensive to undertake, but it can result in a thorough understanding of the strengths and weaknesses of a state’s assessment system.

Surveys of Enacted Curriculum (SEC) Methodology

While many teachers may think they are assessing what is taught and vice versa, assessments present different stimulus conditions than those used in the classroom, and teaching and assessing are often “institutionally dichotomized” (Cohen, 1987). Porter and Smithson (2001) developed the SEC alignment methodology to help people involved in the education process see the connection between what is taught in the classroom and what is assessed, and they applied it in 11 states and four urban districts. Development and application of this model were supported by the Council of Chief State School Officers (CCSSO) through grants from the National Science Foundation and a state collaborative project. This methodology was developed to quantitatively compare degrees of alignment for standards, assessments, and instruction across schools and states. The SEC methodology builds on a content validity approach but also measures the instructional content purportedly taught and captures this information at both a detailed and more general level of analysis.

SEC Alignment Dimensions

The SEC alignment methodology comprises alignment analyses of standards, assessments, and instruction by use of a common content matrix or template that allows comparison across schools, districts or states. The methodology begins in which a coding process in which the content and cognitive levels are determined for the standards, the assessment items, and the instructional focus. The frameworks are coded at the smallest unit possible. Coding at the objective level is similar to the Webb and Achieve methods as the results can be aggregated and reported at the strand level. The assessments are coded at the individual item level. Content experts, teachers, and people familiar with the frameworks code both the standards and the assessments.

There are three main alignment dimensions in the SEC methodology: *content match*, *expectations for student performance*, and *instructional content*. These dimensions are discussed below, as is an application of the SEC methodology.

Content match

The SEC method employs a content matrix of two dimensions: content topic and cognitive complexity/demand (CCSSO, 2002). The task for SMEs is to review items and match them to the topic and complexity cells in the matrix. An example of a content matrix is presented in Figure 1 (see p. 6-44). In the SEC content matrix for mathematics there are 57 topic descriptors at the elementary level, 90 at the middle school, and 160 the high school level. One area of criticism of this method is that the number of content areas can be difficult to manage. However, the benefit is an exhaustive common view of all the content in each area of the educational process. The topics can also be reported at a fine or coarse grain level. The fine grain level displays all of the topics by cognitive area and the coarse grain level rolls up the results to the six broad topic areas, which are similar to strands of content (e.g. Number Sense and Patterns). Thus, the method provides information similar to that gained from traditional content validity studies, but also provides information at a more micro level, which is more likely to better inform instructional and curricular changes (Porter and Smithson, 2002).

Expectations for student performance

The items, standards, and instruction are also coded based on expectations for student performance. This measure is similar to Webb's depth criterion and Achieve's performance centrality measure. The SEC method utilizes six levels of cognitive demand or expectations for student performance. These are: memorize facts, perform procedures, demonstrate understanding, conjecture; generalize; prove, solve nonroutine problems, and make connections. These terms were chosen to be more behaviorally oriented and indicate knowledge and skills required of students as a way to help teachers describe the cognitive expectations they hold for students (Porter and Smithson, 2001). Porter and Smithson recommend using the same cognitive levels for each area of analysis as a means to accurately make comparisons across the instructional content, standards, and assessments.

While the terms and their definitions differ across the Webb, Achieve, and SEC methods, all three approaches highlight the difficulty in training the expert reviewers to understand the distinctions between the cognitive levels. The cognitive areas, however are an important part of the alignment process to address the criticism that standardized tests "dumb down" the curriculum. Through an evaluation of the match between the cognitive demands of each of the educational components (assessment items, standards, instruction), the alignment measure can accurately reflect where differences occur to address the issue of less challenging curricula. The common mapping language allows alignment results to illustrate comparisons of classroom practice to standards and assessments, as well as comparisons among states, districts, and individual teachers.

Instructional content

Unlike the other two alignment methods, the SEC method includes a measure of instructional content. Porter and Smithson (2002) emphasized the importance of including an instructional content component because it serves as an intervening variable when looking at student achievement gains due to standards-based reform. Through surveys, teachers code the instructional content as they think about a preselected target class over a specified period of time. Then, the teachers estimate the emphasis allotted to that topic for each of the cognitive areas. This is then summed to determine the proportion of each topic relative to the total instructional time (Porter, 2002).

The SEC methodology provides a snapshot of practice over a period of time, which is useful in determining the extent to which teaching reflects standards and assessments (Blank, Porter, and Smithson, 2001). This is a critical question that is not directly addressed by the two other alignment approaches. The benefit of the survey approach is that it allows data collection from a large number of respondents and is relatively inexpensive. Other data collection approaches such as daily logs or classroom observations will be more expensive, time consuming, and intrusive on the classroom. Porter (2002) acknowledged the weaknesses of the SEC approach in that the findings are limited to what is asked, it can be subject to self-report bias, and it may be difficult to capture the complexity of instructional practice. Nevertheless, the survey tool has been piloted in multiple settings (Blank et al., 2001) and has proven useful to address the many questions educators and policymakers have about patterns and differences curriculum and instructional practices across classrooms, schools, districts, and states.

The result of the SEC coding across standards, assessments, and instructional content is that each cell in the two dimensional matrix (content by performance expectations) represents the proportion of content, assessment, or standards in that cell and these three pieces can then be compared to determine the degree of alignment. Each area matrix is compared to another to determine the degree of alignment. This resulting alignment index is:

$$1 - [(\sum|X - Y|)/2] \quad [2]$$

where X represents the cell proportions in one matrix and Y represents the cell proportions in the other (Porter, 2002). The values range from 0.0 to 1.0. The results are presented on topographical map layouts to show the relative areas of concentration and facilitate easier comparisons. An example of a topographical map is presented in Figure 2 (see page 6-45). The results of an SEC alignment analysis illustrate gaps in the assessment, the curriculum, or the instruction, which can then be used to guide additional discussions about what, if any, steps need to be taken to address these gaps.

An Application of the SEC Methodology

Blank et al., (2001) studied the degree of alignment between instruction and assessments across six states using the SEC approach. As with other alignment approaches, the reviewer role was crucial to this process. Specialists were brought together for a two-day workshop to code the assessment items and standards. At least four raters independently coded each test. Because one assessment item could potentially assess different areas of content, this procedure limited raters to matching each item with up to three topic areas by student expectation combinations. To capture the instructional content piece, 600 teachers from 200 schools across six states completed the surveys in eighth-grade mathematics.

The results indicated that the alignment of assessment and instruction within a state was similar to the alignment of assessments across states. That is, the alignment indexes derived from cross-state comparisons of tests and standards were similar to those indices derived for comparisons of tests and standards within a state. Alignment of the state assessments to NAEP Grade 8 math and reading assessments were also conducted, and they found there was slightly higher alignment between state assessments and instruction within the state than there was between instruction within the state and NAEP. On the 0 to 1 alignment index scale, across the six states the average alignment among state instruction and state assessment ranged from .23 (grade 8 science) to .42 (grade 4 math), and the average alignment between state instruction and the NAEP assessment ranged from .14 (grade 8 science) to .41 (grade 4 math). However, it should be noted that this study was conducted pre-*NCLB* and none of the states studied had high-stakes attached to the assessments (which would probably influence the degree to which the assessments influence classroom instruction). Nevertheless, the study is a good illustration of how NAEP assessments can be considered in alignment research.

The involvement of teachers in the data collection process for the SEC methodology illustrates how the alignment process and results can directly impact teachers and their instruction. The SEC methodology is one way to get inside the “black box” of classroom instruction and examine these practices in the context of a large-scale study, which is necessary to evaluate the effectiveness of any reform initiative (Blank et al., 2001). To gain teachers’ participation in SEC studies it is imperative that it be voluntary and the results not be tied to any accountability measures. Additionally, teachers should be given individualized results and provided with training about how to use the results (Blank et al., 2001). Results of SEC studies have been used as the basis for professional development opportunities using the in-depth curriculum data for improving instruction in math and science (Blank, 2004).

Porter (2002) summarized the multiple benefits of implementing an SEC approach to alignment. It is an efficient process, once the coders of the assessment and standards and the teachers being surveyed are trained, and the process allows for an objective evaluation of the

alignment goals. It also provides a quantitative measure of alignment that can be used to examine the effect of reform policies over time. Because this approach maps the education pieces to a common language and then compares the results, the process can be used to compare findings across schools, districts, and states, and so it could be used to evaluate NAEP-state alignment across the nation.

SEC limitations

The SEC approach has similar limitations to the other alignment approaches. The process begins with the state standards; however, the tests will only measure a sample of the content domain, while the standards represent the entire domain (Porter, 2002). Additionally, if the standards are not specific enough it will not be possible to tightly align the assessments (Porter, 2002). This methodology does not include the more detailed criteria beyond content and depth match, which are found in the Webb and Achieve models, and so the methodology is unable to quantify the detailed reasons behind limited alignment. Also, research is needed to understand the degree to which teachers and policymakers understand the concept maps that characterize instructional coverage.

The survey process can also be somewhat complex for teachers given the multiple ways they code their instruction (Anderson, 2002). Although the two studies applying this approach had a 75 percent response rate (Porter, 2002), the survey response rates can be dependent on how the survey is administered. Blank et al. (2001) found that the worst response rates were seen in those schools in which teachers were given the surveys to complete on their own at their convenience and the best response rates came from those schools in which the teachers gathered as a group to complete the surveys. Response rates were also higher where teachers were compensated or given professional development credit for the time it took to complete the survey. Blank et al. (2001) concluded that teachers must perceive some personal value to the information they provide. It was important that the information was confidential and that teachers were provided with individual reports if requested, while ensuring the results would not be used for teacher accountability.

Summary of Alignment Methodologies

Bhola et al. (2003) provided a comprehensive overview of different alignment approaches and classified each according to the degree of complexity entailed in the model. *Low complexity* models defined alignment as the extent to which the items in a test match relevant content standards (or test specifications) as judged by content experts rating the degree of match with Likert scale ratings. This is the approach taken in more traditional content validity-type studies (e.g., Sireci, 1998a; Buckendahl et al., 2000). In *moderate complexity* models, content experts decide matches both from content and cognitive perspectives and the result may be a reduction in the number of matches because of this additional constraint. This is the approach used in SEC in which the standards, assessments, and instruction are aligned. *High complexity* models tie in additional criteria to give a broader view of alignment. Webb's (1999) approach and the Achieve approach (Rothman et al., 2002) are examples of this level of detail.

Similarities and Differences Across Methods

The Webb, Achieve, and SEC alignment methods have not yet all been applied in a single study and so the differential utility of the results they provide cannot be accurately described. However, in Table 1 (see page 6-37), we provide a description of the major aspects of each method, organized by four generic dimensions: content, cognitive, distribution, and item quality.

The Webb approach provides the most detailed quantitative results. Based on the four criteria applied, one can see which aspects of alignment are strong or weak. The Achieve methodology builds on the Webb methodology, with the addition of the source and level of challenge dimensions. These dimensions are a means to capture item quality, which was a limitation of Webb's method. However, the most recent applications of Webb's methodology now include a Source of Challenge criterion (Webb, Alt, Ely, and Vesperman, 2005). The Achieve methodology also provides more qualitative information about overall alignment and the quality of the matches. This latter point is missing in the Webb approach where an item-objective match does not convey if the objective is only partially assessed or too vague to be assessed. In this way the specific coding in the Achieve methodology provides a bit more helpful information in terms of possible changes a state might undertake. The broader qualitative results from the Achieve method are very helpful for a specific state application, but might become cumbersome if used for comparison purposes among states. The SEC methodology is the only method that considers the instructional piece of the educational process, which allows for easy comparison of assessments, standards, and instruction across states, districts, and schools. It may also be particularly useful for studying the consequences of a testing program, if comparisons are conducted and compared over time. However, this approach does not probe as deeply as the other two into the quality of the alignment. Thus, these alignment methods have different focuses and each has strengths and limitations in specific situations.

This page left intentionally blank

Importance of Subject Matter Experts (SMEs)

All of the alignment methods depend on SMEs to rate the different components of alignment. In selecting these expert reviewers, all approaches emphasize the importance of knowledgeable SMEs who are familiar with the standards, assessments, and instructional components. It is also critical that the SMEs are familiar with the knowledge and skill levels of the tested population (Sireci, 1998a).

Using expert reviewers is an important part of the process as studies have shown test publisher ratings may differ significantly from expert reviewers (Buckendahl, Plake, Impara, and Irwin, 2000). Additionally, SMEs may be influenced by the fact that they are told the categories that the items, standards, or instructional content must fit into and are constrained by these definitions. Furthermore, they can be influenced by social desirability of what they think is expected, leniency to find a match, and guessing (Sireci, 1998a, 1998b).

Regardless of the alignment method employed, it is important that the level of SME agreement is reported. Rothman (2003) discusses the varying levels of reviewer agreement among the different types of studies. While Achieve uses SMEs that are highly trained in the Achieve methodology, the Webb and SEC methods appear to have more limited training. However, the Webb and SEC alignment results quantify the levels of reviewer agreement. The Webb methodology provides explicit details about the calculations used to capture the reliability of both the cognitive level coding and the item-objective matches (Webb, Alt, Ely, and Vesperman, 2005). The SEC method also computes inter-rater agreement levels. Webb et al. (Webb, Alt, Ely, Cormier, and Vesperman, 2005) noted the importance of having an adequate number of reviewers to ensure the reliability of the coding. Earlier iterations of the Webb methodology recommended three to eight reviewers, but Webb now finds that ideally more than six, but anywhere from five to 12 reviewers, is better to ensure a greater degree of reliability in the coding. While some guidance is provided as to acceptable levels of agreement, this calculation in general serves as a check as to the reliability of human judgments.

Challenges in Evaluating Alignment

Alignment research can be difficult to conduct for six main reasons. First, not everything that is in the standards can be assessed through large-scale standardized assessments. Webb supported broadly defined assessments to include classroom, district, and statewide assessments so as to capture a broader view (Webb, 1997). However, in the alignment studies we reviewed this does not seem practical. All of the alignment studies used statewide, standardized assessments as their comparison, which is most in line with the expectations laid out in *NCLB*. Second, standards may be written at multiple levels and tests may be written to align with standards at the highest level, but the alignment study may use a more detailed level for the standard comparison (Ananda, 2003a). Third, standards may be written to different levels of specificity and may be written so generally that many different types of content are incorporated so that determining a match is difficult (Rothman et al., 2002). Fourth, the terms within the standards may have multiple meanings to different people. Webb (1997) provided an example with the phrase “demonstrate a range of strategies” and discussed how this was difficult to interpret and therefore assess. This point can be addressed in the training of the expert reviewers by determining a set protocol about the level and types of matches that are acceptable. Fifth, items may measure multiple content standards, which can result in error among expert judgments (Le Marca et al., 2000). Sixth, some standards may not be easily assessed and may be redundant within a level, or tests may be designed to assess multiple grade levels. For these reasons perfect alignment is never expected (Ananda, 2003a).

Given the range of criteria used in an alignment study, states need to be clear about their alignment goals. For example, some states might not value the goal of the assessments having a balanced distribution of items across objectives within a standard and may want greater emphasis within specific areas (Ananda, 2003b). Most states will want to ensure their tests adequately measure the intended strands or objectives, and so a traditional content validity study that focuses on this congruence, or the dimensions of alignment models that look at this congruence, may suffice.

Alignment Methods Used by States

In chapters on the Webb Methodology, Achieve Methodology, and the Survey of Enacted Curriculum Methodology, we described the three most popular methods for evaluating test-curriculum alignment and discussed the relative strengths and limitations of each approach. In this section, we discuss the methods being used by states.

Although *NCLB* requires that state tests be aligned with state curriculum frameworks, we found it extremely difficult to find out what methods states were using to ensure or evaluate such alignment. We concluded either such studies are buried in contractor's technical reports that were not available on the Web, or that some states have not completed formal alignment studies. It was easier to find district reports of alignment studies on the Web as well as lists of states and districts with which various alignment service providers had worked. It is likely that some states have conducted alignment analyses but have not made the results public. However, now that states have submitted their assessments for *NCLB* approval, it is possible that more alignment studies will be publicly released and that more studies will actually be conducted (particularly for those states that were not approved due to a lack of alignment evidence!).

With respect to test-curriculum alignment at the state level, given that state alignment studies were difficult to locate, we relied primarily on personal communications. We first noted that 21 states are working with CCSSO as part of the Technical Issues in Large Scale Assessment (TILSA) consortium, to develop and provide alignment resources (see http://www.ccsso.org/projects/Alignment_Analysis/). This group developed resource material for applying the Webb and SEC methods, and the aforementioned Web site presents documentation describing the Webb, SEC, and Achieve methods, as well as links to reports on applications of these methods. A representative of the TILSA consortium informally told us that virtually all states with which he was familiar were using a Webb methodology or something very similar.

We checked that impression by calling or e-mailing the chief state school officer in each state to find out what method was used for state test-state curriculum alignment. We were partially successful in that 24 states responded to our request. Of these states, 19 used the Webb method, one other state was about to conduct a study using the Webb method, two states used the Achieve method, one state said an alignment study was conducted but was not sure of the method, and one state did not use a statewide assessment system. Thus, these results suggest the Webb model is currently the most widely used. However, although these applications referred to the Webb and Achieve methods, we realize they may have been modified or adjusted by contractors to best fit the needs of a particular state, and so a "Webb-based" model may be a better description of actual implementations at the state level. Note also that TILSA provided training to states on using Web Alignment Tool, which may explain its current popularity for helping states provide evidence of the alignment of their assessments to state standards in accordance with *NCLB*.

It is interesting that none of the responding states mentioned use of the SEC model, but that is not surprising given the intensity of that model with respect to data collection. However, with respect to alignment studies conducted at the school and district levels, it appears SEC is quite popular. Below the level of state test/state curriculum framework alignment, we found references to applications of the SEC model in 24 states, applications of the Webb model in 17 states, and applications of the Achieve model in 14 states. Thus, the three alignment methods that are most cited in the literature appear to have at least moderate levels of application.

NAEP-State Alignment Studies

We are aware that the National Center for Education Statistics has supported some research into the alignment of NAEP and state assessments. For example the NAEP State Service Center compared the frameworks, assessments, and results between NAEP and assessments in three states, but the states were anonymous and a report has not yet been released. Thus, at this juncture, there does not seem to be a comprehensive effort at the federal level to do a formal analysis of the alignment of state tests to NAEP, although there seems to be some interest in this topic.

WestEd (2002) compared the alignment of NAEP Reading, Writing, and Mathematics frameworks to the content standards for four states—Arizona, California, Nevada, and Utah. They focused on tests at grades 4 and 8. They stressed the importance of the study in the context of *NCLB* by stating:

Although states will not be sanctioned for failure to demonstrate NAEP student performance improvement, NAEP data will provide an external accountability benchmark and serve to verify student achievement on state assessments. Given the elevated status of NAEP to a de facto national benchmark, states naturally want to know how well their standards align to NAEP so that they can make informed decisions about possible changes to their own standards and assessment systems. (p. 1)

Given that the purpose of this study was to evaluate the congruence between NAEP content specifications and state content standards across four states, they did not implement a comprehensive alignment model. Instead, they used a three-point rating scale to develop a crosswalk between the NAEP and state frameworks. For each NAEP concept, the rating scale was used to determine whether the state standard (a) fully addressed or exceeded the NAEP concept, (b) partially addressed the NAEP concept, or (c) did not address the NAEP concept. Using this approach, they concluded that there was “strong” correspondence between all four states and NAEP reading at both grades. For writing, the correspondence was “moderate” for two of the states and “strong” for the other two. For grade 8 math, the congruence was “moderate” for three states and “strong” for the other. Grade 4 math exhibited the lowest level of congruence, with three states classified as “partial” and the other classified as “moderate.”

An interesting feature of this study is “partial” matches were concluded by looking at state standards assessed below and above the NAEP grade levels (not all states had approved standards at the NAEP grade levels). For example, some grade 8 NAEP concepts could be assessed at grades 6, 7, 9, or 10 on a state assessment. Looking for standards at grade levels different from NAEP is likely to be important in future studies of NAEP-state alignment, if the similarities and differences in what is being measured are to be uncovered and explained. Another interesting feature of the WestEd (2002) study is that the coding scheme allowed reviewers to insert comments into the crosswalk whenever discrepancies arose or clarifications were needed. Thus, although it is a bit dated, this study indicates how the alignment between NAEP and state assessments could be conducted across multiple states.

Conclusion

Alignment is a means for understanding the degree to which different components of an educational system work together to support a common goal. In this age of accountability, it is important that state organizations, districts, and schools support each other to send a consistent message to teachers and students about what is required. Alignment research is one method to demonstrate this consistency of message or to understand what changes need to be addressed to ensure every student has the opportunity to learn the content on which they are assessed, and to demonstrate his or her proficiency. Furthermore, to meet the expectations of alignment under *NCLB*, states will need to conduct independent analyses of the alignment between their tests and curriculum frameworks, and if any gaps are discovered, they will need to take corrective action. All three of the methodologies we reviewed start with the basic evaluation of the alignment of the content and cognitive complexity of standards and assessments. The SEC methodology also includes an instructional component. On to this foundation the Webb and Achieve methodologies layer additional criteria to better understand the breadth and range of comparison between the standards and the assessments. The Achieve methodology also includes an overarching view of the sets of items to look at the broader quality of an assessment relative to the standards on which it is based.

When deciding between these three alignment approaches, it is important to understand the financial, time, and personnel resources available, as well as the ultimate goals of the research. However it is accomplished, alignment research should be viewed as an ongoing process to continually understand how the assessment, the standards, and the instruction support each other to deliver a consistent message to students about what is expected.

Through *NCLB*, student assessments have become a dominant feature of the educational process. An important component of the effectiveness of *NCLB* is the use of assessments to improve instruction. Teachers need to understand the value of the assessments, how the assessments relate to what they should be teaching, and how to make changes in their approach based on the results they see. Teachers' involvement in alignment research is one way to help teachers become more familiar with the assessments and the standards on which they are based. Alignment research that incorporates the findings about effective forms of professional development studies can ensure teachers apply what they are learning through the alignment process to their classroom.

With respect to NAEP, the alignment methods reviewed in this paper could be used to evaluate (a) the degree to which NAEP tests are congruent with content and cognitive frameworks (Sireci, Robin, et al., 2000), and (b) the degree to which different state assessments are congruent with NAEP assessments and with each other. Each alignment approach allows for a useful summarization of the congruence among specific aspects of an assessment system. Alignment studies for NAEP exams, or for NAEP-state comparisons, that focus on the most general level of alignment, such as the study conducted by WestEd (2002) could provide valuable information for understanding discrepancies in NAEP and state test results, and for determining the unique features of state curriculum frameworks, relative to NAEP frameworks. Studies using the SEC methodology could be particularly helpful as that approach steps away from the specific language in the standards and uses a more generally applicable topic-focused language.

Theoretically, there are at least four ways in which NAEP-state alignment could be evaluated: (a) comparing NAEP frameworks with state curriculum frameworks (content standards), (b) comparing NAEP and state assessments, (c) comparing NAEP frameworks with state tests, and (d) comparing state frameworks and standards with NAEP tests. However, direct comparisons with NAEP assessments are problematic due to the confidentiality of NAEP items and the complexities of the balanced, incomplete block design used to partition the NAEP item pool across samples of students. Therefore, approaches that compare NAEP and state frameworks are most practical (see Gatti, 2004; Smithson, 2004; and WestEd, 2002, for examples). Nevertheless, comparing state assessments with NAEP frameworks is also possible

and is recommended for a more complete analysis of NAEP-state alignment. Given that many states use sources for deriving test specifications that are similar to those used to develop NAEP test specifications (e.g., National Council of Teachers of Mathematics), the results of NAEP-state alignment studies may show substantial overlap. However, that is simply a hypothesis to be tested.

Alignment research represents an exciting and powerful means for bringing different parts of the educational system together in a systematic and efficient way. While the process may be costly, as it is dependent on expert reviewers and takes time, the results send a powerful message about the state of these educational components, assessments, standards, and instruction, and what might need to be addressed going forward.

References

- Ananda, S. (2003a). Achieving alignment. *Leadership*, 33(1), 18–22.
- Ananda, S. (2003b). *Rethinking issues of alignment under "No Child Left Behind."* San Francisco, Calif.: WestEd.
- Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory into Practice*, 41(4).
- Barth, P. (March, 2006). Score wars: Comparing the National Assessment of Educational Progress with state assessments. Alexandria, Va.: Center for Public Education. Downloaded from http://www.centerforpubliceducation.org/site/c.kjJXJ5MPIwE/b.1577019/k.A07C/Score_wars_Comparing_the_National_Assessment_of_Educational_Progress_with_state_assessments.htm, on May 23, 2006.
- Bhola, D. S., Impara, J. C., and Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21–29.
- Blank, R. K. (April, 2004). Findings on alignment of instruction using enacted curriculum data: Results from urban schools. Paper presented at the Annual meeting of American Educational Research Association, San Diego, Calif.
- Blank, R. K., Porter, A. C., and Smithson, J. L. (2001). *New tools for analyzing teaching, curriculum and standards in Mathematics and Science*. Washington, D.C.: Council of Chief State Schools Officers.
- Buckendahl, C. W., Plake, B. S., Impara, J. C., and Irwin, P. M. (April, 2000). Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, La.
- CCSSO. (2002). *Models for alignment analysis and assistance to states*. Retrieved Aug. 28, 2005, from www.ccsso.org/content/pdfs/AlignmentModels.pdf
- CCSSO. (2005). *Models*. Retrieved Aug. 28, 2005, from <http://www.ecs.org/html/offsite.asp?document=http%3A%2F%2Fwww%2Eccsso%2Eorg%2Fprojects%2FAlignment%5FAnalysis%2FModels%2F>.
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19–27.
- Cohen, J., Seburn, M., Gushta, M., Chan, T., and Jiang, T. (2005). What can NAEP and state assessments learn from each other about measuring progress? Unpublished report. Washington, D.C.: American Institutes for Research.
- Cohen, D. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, 12(3), 3,111–3,344.
- Cohen, S. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher*, 16(8), 16–20.
- Crocker, L. M. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement, Issues and Practice*, 22(3), 5–11.
- Crocker, L. M., Miller, D., and Franks E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2, 179–194.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, D.C.: American Council on Education.
- Cross, R. W., Rebarber, T., and Torres, J. (eds.) (2004). *Grading the systems: The guide to state standards, tests, and accountability policies*. Washington, D.C.: Thomas B. Fordham Foundation.
- de Vise, D. (2005). State gains not echoed in federal testing. *Washington Post*, Oct. 24, 2005, p. B01
- Dillon, S. (2005). Students ace state tests, but earn D's from U.S. *New York Times*, Nov. 26, 2006, p. A1.

- Fuller, B., Gesicki, K., Kang, E., and Wright, J. (2006). Is the No Child Left Behind Act working? The reliability of how states track achievement. Working paper 06-1. Berkeley, Calif.: Policy Analysis for California Education, University of California, Berkeley.
- Gatti, G. G. (October, 2004). Alignment of state-to-NAEP content standards in 4th grade Scott Foresman and 8th grade Prentice Hall mathematics textbooks. Paper presented at the Northern Rocky Mountain Educational Research Association, Custer, S.D.
- Herman, J., Webb, N., and Zuniga, S. (2005). Measurement issues in the alignment of standards and assessments: A case study (No. CSE Report 653). Los Angeles, Calif.: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Le Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *ERIC Digest* (No. ED458288): ERIC Development Team.
- Le Marca, P. M., Redfield, D., Winter, P. C., and Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessments*. Washington, D.C.: Council of Chief State Schools Officers.
- Linn, R. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4–16.
- Linn, R. L. (June, 2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, 13(33). Available at <http://epaa.asu.edu/epaa/v13n33/>.
- McGehee, J. J., and Griffith, L. K. (2001). Large-scale assessments combined with curriculum alignment: Agents of change. *Theory into Practice*, 40(2).
- Phillips, S. E. (2000). GI Forum v. Texas Education Agency: Psychometric evidence. *Applied Measurement in Education*, 13, 343–385.
- Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education*, 5, 285–301.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31(7), 3–14.
- Porter, A. C., and Smithson, J. L. (2001). *Defining, Developing, and Using Curriculum Indicators. CPRE Research Report Series* (No. RR-048). Philadelphia, Pa.: Consortium for Policy Research in Education.
- Porter, A. C., and Smithson, J. L. (2002). Alignment of assessments, standards and instruction using curriculum indicator data. Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, La.
- Roach, A. T., Elliott, S. N., and Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin Alternate Assessment. *Journal of Special Education*, 38(4), 218–231.
- Rothman, R. (2003). Imperfect matches: The alignment of standards and tests: Commissioned paper, National Research Council, Washington, D.C.
- Rothman, R., Slattery, J. B., Vranek, J. L., and Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing. CSE Technical Report* (No. CSE-TR-566). Los Angeles, Calif.: National Center for Research on Evaluation, Standards, and Student Testing.
- School Matters (Fall, 2005). *The National Assessment of Educational Progress and state assessments: What do differing student proficiency rates tell us?*. Retrieved May 23, 2006 from http://www.schoolmatters.com/pdf/naep_schoolmatters.pdf.
- Sireci, S. G. (1998a). Gathering and analyzing content validity data. *Educational Assessment*, 5(4), 299–321.
- Sireci, S. G. (1998b). The construct of content validity. *Social Indicators Research*, 45, 83–117.
- Sireci, S. G., Lewis, C., and Martone, D. (2006). Principals' perceptions of the effects of standardized testing on instruction. *Center for Educational Assessment Research Report No. 595*. Amherst, Mass.: Center for Educational Assessment, University of Massachusetts Amherst.

- Sireci, S. G., Robin, F., Meara, K., Rogers, H. J., and Swaminathan, H. (2000). An external evaluation of the 1996 Grade 8 NAEP Science Framework. In N. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell and L.R. Jones (eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 74–100). Washington, D.C.: National Academy Press.
- Smith, M., and O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman and B. Malen (eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Bristol, Pa.: Taylor and Francis.
- Smith, M., and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 4(10), 7–11.
- Smithson, J. L. (September, 2004). Summary report on alignment analyses of Prentice Hall mathematics test forms to gr. 8 NAEP benchmarks and state mathematics standards in five states. Unpublished research report, University of Wisconsin-Madison: Madison, Wisc.
- Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.
- Webb, N. L. (1997). *Research monograph no. 6: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, D.C.: Council of Chief State Schools Officers.
- Webb, N. L. (1999). *Research monograph no. 18: Alignment of Science and Mathematics standards and assessments in four states*. Washington, D.C.: Council of Chief State Schools Officers.
- Webb, N. L., Alt, M., Ely, R., Cormier, M., and Vesperman, B. (2005). *The WEB alignment tool: Development, refinement, and dissemination*. Washington, D.C.: Council of Chief State School Officers.
- Webb, N. L., Alt, M., Ely, R., and Vesperman, B. (2005). *Web alignment tool (WAT): Training manual draft 1.1*. Retrieved March 17, 2006, from <http://www.wcer.wisc.edu/WAT/Training%20Manual%202.1%20Draft%20091205.doc>
- Webb, N., Herman, J., and Webb, N. L. (2006). *Alignment of mathematics state-level standards and assessments: The role of reviewer agreement* (No. CSE Report 685). Los Angeles, Calif.: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- WestEd (April, 2002). *A comparison of NAEP content to reading, writing, and mathematics content standards for Arizona, California, Nevada, and Utah*. San Francisco: Author.
- Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education*, 62, 288–310.
- Wise, L. L., and Alt, M. (2005). Assessing vertical alignment. Washington, D.C.: Council of Chief State School Officers.
- Wise, L. L., Zhang, L., Winter, P., Taylor, L., and Becker, D. E. (2005). *Vertical alignment of grade-level expectations for student achievement: Report of a pilot study*. Washington, D.C.: Council of Chief State School Officers.

This page left intentionally blank

Appendix A: Sample Webb Reports

Table 1. Categorical Concurrence Table

9 Reviewers, 51 Assessment Items

Standards			Level by Objective			Hits		Cat. Concurr.
Title	Goals #	Objs #	Level	# of objs by Level	% w/in std by Level	Mean	S.D.	
I - Patterns, Relationships and Functions	2	11.11	2 3	5 6	45 54	10.44	3.06	YES
II - Geometry and Measurement	3	18	1 2 3	4 11 3	22 61 16	13	2.00	YES
III - Data Analysis and Statistics	3	14.22	2 3 4	3 8 3	21 57 21	13.44	2.22	YES
IV - Number Sense and Numeration	3	14.33	1 2	4 10	28 71	2.78	1.69	NO

The *Categorical Concurrence* criterion is one of the five main alignment criteria. It measures whether the same or consistent categories of content appear in the standards and the assessments. The criterion is met for a given standard if there are at least six assessment items targeting objectives falling under the standard.

Goals #	Number of objectives plus one for a generic objective for each standard.
Objectives #	Average number of objectives for reviewers. If the number is greater than the actual number in the standard, then at least one reviewer coded an item for the goal or objective but did not find any objective in the goal that corresponded to the item.
Level	The Depth-of-Knowledge level coded by the reviewers for the objectives for each standard.
# of objs. by Level	The number of objectives coded at each level
% w/in std by Level	The percent of objectives coded at each level
Hits—Mean & SD	Mean and standard deviation number of items reviewers coded as corresponding to standard. The total is the total number of coded hits.
Cat. Conc. Accept.	“Yes” indicates that the standard met the acceptable level for criterion. “Yes” if mean is six or more. “Weak” if mean is five to six. “No” if mean is less than five.

Table 2. Depth-of-Knowledge Consistency Table

9 Reviewers, 51 Assessment Items

Standards			Hits		Level of Item w.r.t. Standard						DOK Consistency
Title	Goals #	Objs #	M	S.D.	% Under		% At		% Above		
					M	S.D.	M	S.D.	M	S.D.	M
I - Patterns, Relationships and Functions	2	11.11	10.44	3.06	83	37	17	37	0	0	NO
II - Geometry and Measurement	3	18	13	2.00	20	38	51	46	29	43	YES
III - Data Analysis and Statistics	3	14.22	13.44	2.22	58	41	40	40	2	12	WEAK
IV - Number Sense and Numeration	3	14.33	2.78	1.69	25	42	61	48	14	34	YES

The *Depth-of-Knowledge Consistency* is another of the main alignment criterion. This criterion between standards and assessment measures the degree to which the knowledge elicited from students on the assessment is as demanding and complex within the context area cognitively as what students are expected to know and do as stated in the standards. To find the percent Under for a standard (for one reviewer), the percent Under is calculated for each objective (percent of items targeting that objective at too low of a DOK level), then this result is averaged across all the (hit) objectives in the standard. This is then averaged across all reviewers. If the combined percent At and percent Above is 50 or higher, the criterion is fully met. If they add to 41–49, the criterion is weakly met. In the case of a balanced standard, this amounts to the criterion being fully met if 50 percent of the assessment items are at as high a DOK level as the objectives that they target.

The first five columns repeat columns from Table 1 (Categorical Concurrence).

Level of Item w.r.t. Stand Mean percent and standard deviation of items coded as “under”, “at”, and “above” the Depth-of-Knowledge level of the corresponding objective. See explanation above.

Depth-of-Know.
Consistency Accept.

For a balanced standard, “Yes” indicates that 50 percent or more of the items were rated as “at” or “above” the Depth-of-Knowledge level of the corresponding objectives. “Weak” indicates that 40 percent to 50 percent of the items were rated as “at” or “above” the Depth-of-Knowledge level of the corresponding objectives. “No” indicates that less than 40 percent items were rated as “at” or “above” the Depth-of-Knowledge level of the corresponding objectives.

Table 3. Range-of-Knowledge Correspondence and Balance of Representation (Continued)

Note: BALANCE INDEX

$$1 - \frac{\left(\sum_{k=1}^0 \left| \frac{1}{O} - \frac{I_k}{H} \right| \right)}{2}$$

Where O = Total number of objectives hit for the standard
 $I_{(k)}$ = Number of items hit corresponding to objective k
 H = Total number of items hit for the standard

Bal. of Rep Accept.

“Yes” indicates that the Balance Index was .7 or above (items evenly distributed among objectives).

“Weak” indicates that the Balance Index was .6 to .7 (a high percentage of items coded as corresponding to two or three objectives).

“No” indicates that the Balance Index was .6 or less (a high percentage of items coded as corresponding to one objective.)

Table 4. Summary of Attainment of Alignment Criteria

9 Reviewers, 51 Assessment Items

Standards	Alignment Criteria			
	Categorical Concurrence	Depth-of- Knowledge Consistency	Range of Knowledge	Balance of Representation
I - Patterns, Relationships and Functions	YES	NO	NO	YES
II - Geometry and Measurement	YES	YES	NO	YES
III - Data Analysis and Statistics	YES	WEAK	NO	YES
IV - Number Sense and Numeration	NO	YES	NO	YES

Table 5: A Comparison of the Three Most Popular Alignment Approaches

Dimension	Webb	Achieve	SEC
Content	<p>Categorical concurrence: compare standards and assessments.</p> <p>Goal: 6 items per broad content standard</p>	<p>Confirm test blueprint then analyze content centrality</p> <p>Able to capture standards that are too broadly written to be completely assessed</p>	<p>Topic coding— assessment items, standards, and instructional content are mapped to a common content language, organized into logical groupings of topics. Allows for comparison of the instructional content emphasized in standards, assessments, and instruction.</p>
Cognitive levels	<p>Depth-of-knowledge consistency: Cognitive demand comparison between objectives and tests</p> <p>Cognitive levels: recall, skill or concept, strategic thinking, extended thinking</p> <p>Goal: At least 50 percent of the items matched to an objective at or above the cognitive level of that objective</p>	<p>Performance centrality: Cognitive demand comparison between objectives and tests, coded after content match; focuses on the verbs used in the standard vs. what the item requires— e.g. select, identify, compare, analyze, represent, use</p> <p>Able to capture standards that are too broadly written to be completely assessed</p> <p>Cognitive levels: Assigns a level of demand, ranging from 1–4, to each item</p> <p>Level of challenge: captures whether the collection of items mapped to a given standard are appropriately challenging Similar to cognitive comparison but adds a more descriptive piece</p>	<p>Expectations for student performance: Cognitive demand comparison of items, standards, and instructional focus</p> <p>Cognitive levels: memorize facts, perform procedures, demonstrate understanding, conjecture generalize prove, solve non-routine problems, and make connections</p>

Continues next page

Table 5. A Comparison of the Three Most Popular Alignment Approaches (Continued)

Dimension	Webb	Achieve	SEC
Distribution	<p>Balance of representation (how evenly assessment items are distributed across objectives within a standard)</p> <p>Goal: All objectives should be measured by at least two items</p>	<p>Balance: relative importance the test items give to content and performances should be proportionately similar to what is stated in the standards</p> <p>Qualitatively captures which objectives within a standard seem to be over- or underassessed, redundant items, and how the set of items measures what content reviewers think is important for that level</p>	NA
Item quality	NA	<p>Source of challenge: ensure items are fairly constructed and are not designed to trick students; Also examines reading passages and prompts, rubrics and anchor papers for writing samples for grade level appropriateness</p>	NA

Figure 1. Example of SEC Content Matrixes

From (Porter and Smithson, 2002)

Example of matrices

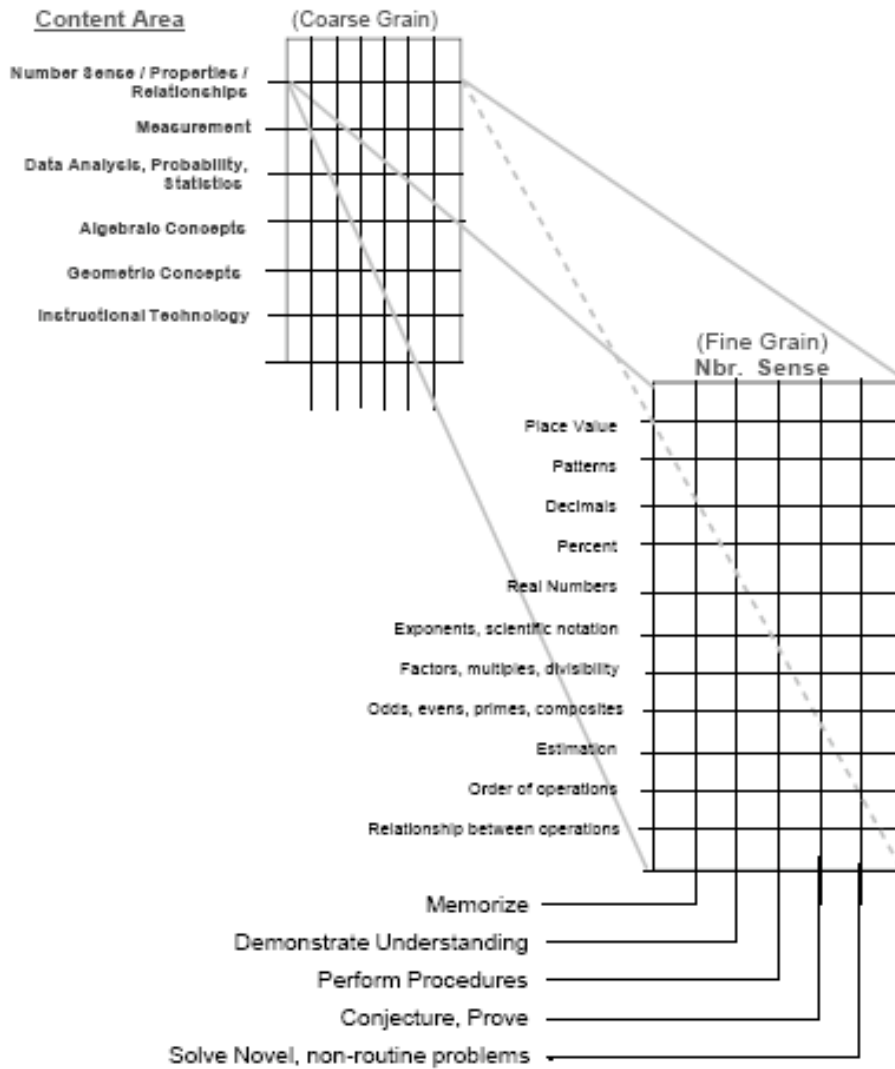
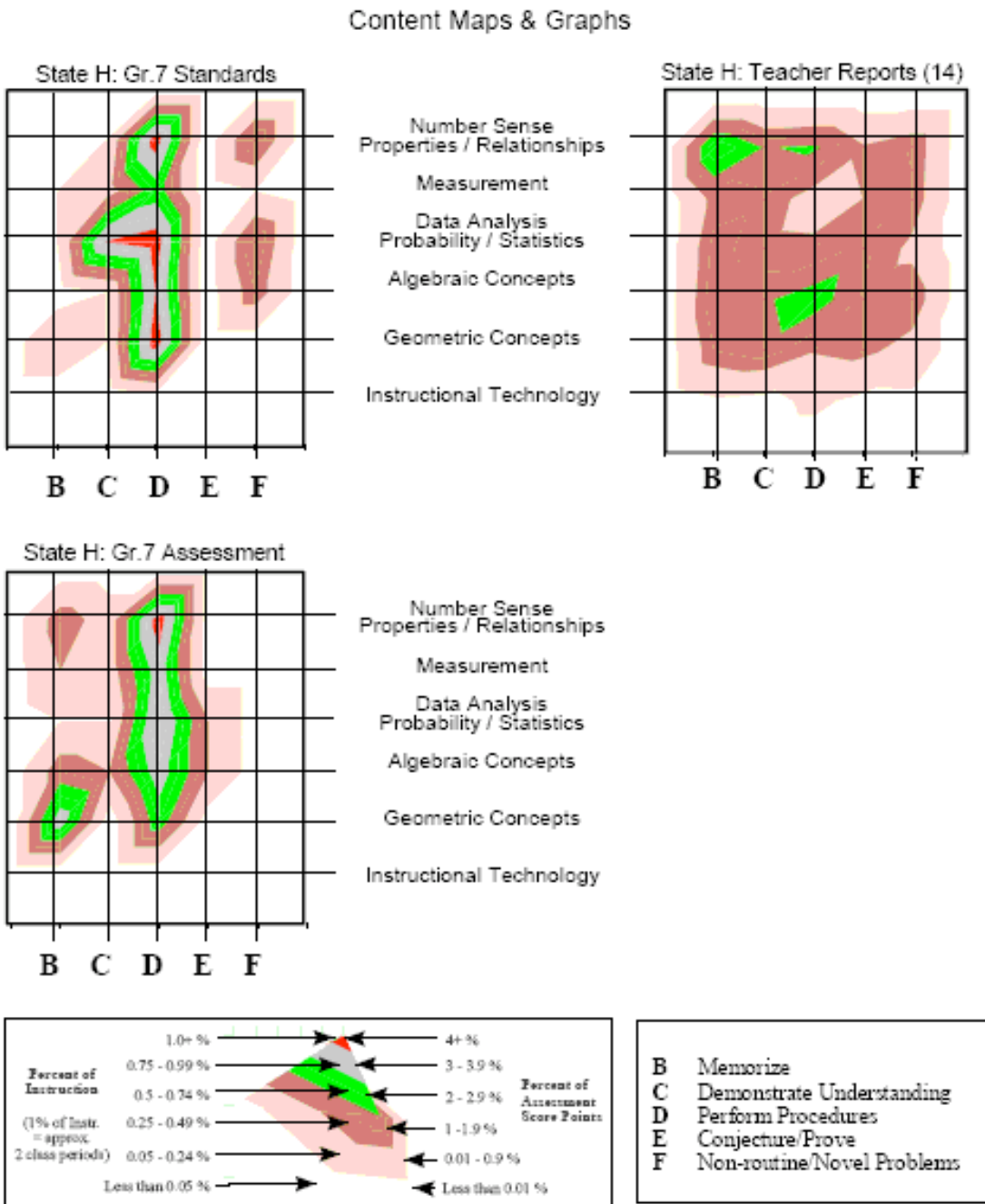


Figure 2. Example of an SEC “Topographical” (Content) Map

From Porter and Smithson, 2002



This page left intentionally blank

*The Department of Education's mission is to promote student achievement
and preparation for global competitiveness by fostering educational
excellence and ensuring equal access.*

www.ed.gov