AN ABSTRACT OF THE DISSERTATION OF

Jung-Mi Ha for the degree of Doctor of Philosophy in Science Education presented on December 14, 2006.
Title: The Use of NAEP Data in a State Context

Abstract approved: _____
Lawrence B. Flick

The National Assessment of Educational Progress (NAEP) has measured the condition and progress of education since 1969 to provide information on performance of American students in core subjects to policymakers, educators, and the public. Yet, there has been a perception that theses federal resources are under utilized. An analysis of the NAEP website revealed that NAEP data could be utilized to inform state-level education and policy decisions in multiple ways. However, there is a very limited empirical base on the actual uses of NAEP. Today, NAEP data and resources are easier to access through the Web and the NCLB's mandate that states participate in biennial state NAEP in reading and mathematics at grades 4 and 8 has drawn greater attention to state NAEP.

In this context, it was speculated that easier access to NAEP data and the NCLB requirement for states might facilitate the use of NAEP at the state level. This case study was intended to verify the assumption by investigating how state education personnel in US State perceive the usefulness of NAEP and how NAEP data are actually used in a state context. Data were collected through interviews, documents, and relevant websites.

This study found that in relation to NAEP's potential utility, the use of NAEP at the state level is limited. The use focused primarily on disseminating NAEP information in the state and considering NAEP frameworks when revising the state's standards.

Nonetheless, the findings of this study suggest that more NAEP data are currently used in US State than before and that the increased use has been more affected by NCLB than by the improved availability of NAEP data.

There appear to be several possible reasons for the limited use of NAEP by the state. Firstly, the state education personnel do not necessarily have an in-depth understanding of diverse aspects of the NAEP program and some of them might have limited knowledge of statistics. In addition, the National Center for Education Statistics sometimes does not provide enough of relevant information along with NAEP results in reporting to facilitate the use and correct interpretations of the results.

©Copyright by Jung-Mi Ha
December 14, 2006
All Rights Reserved

The Use of NAEP Data in a State Context

by
Jung-Mi Ha

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 14, 2006
Commencement June 2007

Doctor of Philosophy dissertation of Jung-Mi Ha presented on December 14, 2006.

APPROVED:

Major Professor, representing Science Education

Chair of the Department of Science and Mathematics Education

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Jung-Mi Ha, Author

ACKNOWLEDGEMENTS

My sincere thanks go first to my major professor, Dr. Lawrence Flick, for his continued inspiration and guidance. He has been a continuous and invaluable source of encouragement, support, and constructive criticism, providing the needed impetus to complete this project. Without his pushing me to be my best, I could not have made this long journey to the completion of my dissertation. I would also like to thank the other members of my committee (Dr. Susan Huggins, Dr. Karen Higgins, Dr. Maggie Niess, and Dr. Corinne Manogue) who willingly read various drafts and provided critical feedback. I also thank many other members of the department of Science and Mathematics Education for their support for my project.

I would like to give my special thanks to the participants of this study who volunteered to provide their perspectives and views on the issues related to my research. Thus, my thanks go to the seven members of the state education staff in US State and the two NAEP persons in the National Center for Education Statistics. In particular, I would like to thank the NAEP state coordinator in US State for generously and continuously providing valuable information.

Finally, my sincere thanks go to my family. My sister Jung-Nan and my brother Jung-Sik in Korea have given me their constant patience and moral support that were the basis for pursuing a Ph.D. degree here in the US. In particular, I express my deepest thanks to my parents in heaven.

TABLE OF CONTENTS

Page

CHAPTER I: THE PROBLEM 1

Introduction 1
The NAEP Website 6
Statement of the Problem 7
Theoretical Framework 10
Significance of the Study 11

CHAPTER II: REVIEW OF THE LITERATURE 13

Assessment, Accountability, & Achievement 14
The History of NAEP 17
The NAEP Program 21
Overview 23
Design 26
NAEP Frameworks 28
Scaling Procedures 30
Performance Standards 32
Administration 35
Accommodations 36
Background Questions 38
Alignment between Standards and NAEP Tests 39
The Psychometric Quality of NAEP 48
Utility of NAEP Data 60
Methodological Approaches Guiding This Study 83
Discussion and Conclusions 88

CHAPTER III: DESIGN AND METHOD 92

Method 93
Participants 97
State Education Personnel 98
NCES Staff 98
Data Collection 100
The NAEP Website 100
Interviews 101
Written Documents 110
State Education Agency's NAEP Website 111
Timeline for Data Collection 112
Data Analysis 113
Interviews 113
Written Documents 115
Websites 116

TABLE OF CONTENTS (Continued)

Page

The Researcher 118

CHAPTER IV: RESULTS 121

State Context 121
US State Assessment Program 123
US State Performance Standards 124
Nature of the NAEP Website 125
The Site Structure 126
The Site Management 141
Potential Utility of Information on the NAEP site 143
Summary 161
State Education Personnel's Perceptions of the Usefulness of NAEP 162
Assessment Purposes and the Appropriate Use of NAEP Data....164
Policy Relevance of NAEP Data 168
State-to-State Comparisons 175
NAEP under NCLB 182
NAEP Achievement Levels 186
Improvement of NAEP 192
NAEP Data Use by State Education Personnel 203
Provision of NAEP Information 206
Making Comparisons with Other States and the Nation....210
Revision of the State Standards 212
Relationships between Performance and Background Factors....215
Linking of State Test Results and NAEP Results 217
Use of the NAEP Website by State Education Personnel....219
The Website Use 220
Expectations for the Website 223
Summary 225

CHAPTER V: DISCUSSION AND IMPLICATIONS....228

Discussion....228
Alignment 229
Closing Achievement Gaps 232
Achievement Levels 235
State-to-State Comparisons 239
NAEP Data Use by State Education Personnel 244
Implications for Future Research....251
Limitations of the Study....253

REFERENCES 256

TABLE OF CONTENTS (Continued)

Page

APPENDICES 266
Appendix A. Informed Consent Form 1 267
Appendix B. Informed Consent Form 11 269
Appendix C. Interview Protocol (state education personnel) 271
Appendix D. Interview Protocol (NAEP webmaster) 276
Appendix E. Interview Protocol
(project officer for NAEP state coordinators).....279

LIST OF FIGURES

Figure Page

1.Study Design....96

2. Context of the Study 97

3. Homepage of the NAEP Website 125

4. The Structure of the Site 127

5. US State Profiles 129

6. Cross-State Comparison Map 130

7. Front Page of Questions Tool 132

8. Student Responses 132

9. Performance Data 133

10. Cross-State Data 134

11. 2005 Mathematics Grade 8 Item Maps 135

12. The Front Page of NDE 136

13. Variable Selections 137

14. Results Table....138

15. Statistical Significance Test 138

16. US State mathematics scale score percentiles, grade 4: 1996-2005 147

17. US State mathematics scale score percentiles, grade 8: 1990-2005 147

18. NAEP mathematics scores by National Lunch Program Eligibility
: Grade4 150

19. Average scores by National Lunch Program Eligibility 150
: Grade 8

20. Average mathematics scores by race/ethnicity: Grade 4 152

21. Average mathematics scores by race/ethnicity: Grade 8 153

LIST OF FIGURES (Continued)

Figure Page

22. NAEP mathematics scale score comparisons at grade 8 155

23. US State's trends in 8th- grade mathematics performance on state tests and on NAEP 157

The Use of NAEP Data in a State Context

CHAPTER I

THE PROBLEM

Introduction

The National Assessment of Educational Progress (NAEP), well known as *the Nation's Report Card,* has been in operation for more than three decades. It is the only nationally representative and continuing assessment measuring American students' academic achievement and their progress over time (Berends & Koretz, 1995/1996; Burstein, Koretz, Linn, Sugrue, Novak, Baker, & Harris, 1995/1996; Mullis, 2003). NAEP was originally created to serve as a barometer of student learning so that it could be used to help improve teaching and learning. The NAEP program is sponsored by the U.S. Department of Education and administered by the National Center for Education Statistics (NCES), the federal agency with fiscal and operational control over NAEP. Since 1989, NAEP policy has been determined by the National Assessment Governing Board (NAGB).

Since it was launched in 1969 following five years of intense preparation, NAEP has undergone a series of changes to reflect contemporary educational reforms and technological advances in measurement. For example, after the NAEP grant was awarded to the Educational Testing Service (ETS) in 1983, new features in design were introduced to respond to policy demands, such as Balanced Incomplete Block (BIB) design, item response theory (IRT) scaling, and grade-based data (Jones, 1996; Pellegrino, Jones, &

Mitchell, 1999). Particularly since 1990, the NAEP design has been statistically complex because of the pressures for more emphasis on constructed-response items (Glaser, Linn, & Bohrnstedt, 1996).

NAEP analysis and reporting methods also have changed accordingly. For instance, beginning in 1992, three separate scaling models were used: one for multiple-choice items, one for dichotomously-scored (right/wrong) constructed-response items, and one for partial-credit (extended constructed-response) items. As for reporting, before the early 1980s results were presented on a question-by-question basis, that is, reporting the percentages of students who answered each question correctly (Phillips, Mullis, Bourque, Williams, Hambleton, Owen, & Barton, 1993). Scaled scores were introduced later to report trends in achievement, but were found not meaningful to lay audiences because of the arbitrary nature of the scale values (Phillips et al., 1993). Finally, the current emphasis on standards-based reporting resulted with the hope that achievement levels presented with descriptions of varying detail and exemplar items would give scaled scores meaning to NAEP audiences (Jaeger, 2003; Linn, Koretz, & Baker, 1996; Phillips et al., 1993). Yet little research has been conducted on how NAEP users interpret and use the achievement-level results (Pellegrino et al., 1999).

In addition to NAEP's major purpose, other functions have been added along the way (Jones, 1996; NCES, 1999b). For example, the 1988 reauthorization of NAEP provided the option to individual states for Trial State Assessments (TSA) in mathematics, establishing a new direction for NAEP. Clearly, NAEP was redirected to be a policy-relevant instrument by enabling state-to-state comparisons, rather than a neutral statistical instrument originally designed by Tyler (Epstein, 1998). Mullis (2003) and Phillips (as

cited in Jones, 1996, p. 18) argue that state NAEP could be used as a tool for allowing state achievement to be validly compared to national performance and to other states, thus leading to fostering educational improvements. In contrast, Koretz (1991, 1992) asserted that state NAEP simply confirms what is already known from other sources and that state NAEP is not worth the costs where costs include not only dollars expended but also risks of corruption of NAEP by teaching to the test. This early debate led to a question: "how do state education personnel view state comparisons resulting from state NAEP?" Finally, following an evaluation of TSA, a panel of the National Academy of Education recommended that state NAEP be continued with ongoing evaluation and Congressional oversight, thus resulting in Congressional authorization of State NAEP in 1994 (Jones, 1996).

Recently, under the No Child Left Behind (NCLB) legislation of 2001, NAEP seems to be entering a new phase in which it might be used for state accountability (Rust, 2004). NCLB requires states to set Adequate Yearly Progress (AYP) objectives with the goal of having all students at the proficient level on state assessments within 12 years, say, by the 2013-2014 school year. AYP is a measure of year-to-year student achievement on the statewide tests. NAEP has always been a low-stakes assessment, but NCLB changes that. For example, the NCLB made it mandatory for states to participate in State NAEP at grades 4 and 8 in both reading and mathematics every other year. If states and school districts do not participate, they will not be eligible to receive Title 1 funds under the Elementary and Secondary Education Act (ESEA) of 1994. At the same time, NCLB requires that results in these subjects be released within six months of data collection, a far faster schedule than ever before. Further, NCLB appears to encourage states to verify

state gains with trends on NAEP (Linn, Baker, & Betebenner, 2002; Wise, 2003), which raises a question: “how do state education personnel perceive the possible NAEP’s new role as confirmation?” This issue was addressed in this study and is discussed in the following sections.

NAEP data and resources might be used in diverse and meaningful ways to inform education and policy issues and reform efforts at the state level. Sebring and Boruch (1983) found NAEP could be used in three ways: policy use, professional use, and research use (the categories might overlap). *Professional decisions* by the states involve employing NAEP data, methods, or materials to improve educational programs and instruction. For example, NAEP might be utilized: (1) to develop or revise their content standards; (2) to learn more about the relative strengths and weaknesses of student performance; and (3) develop in-service programs. *Policy decisions* at the state level might refer to the use of NAEP data: (1) to document education inequities and take appropriate actions; and (2) to revise methods of assessing student performance. For *research*, NAEP data can be used to understand the relationship between student and school variables and achievement.

The utility issue is critical to the continued success of the NAEP program since promoting use will enhance the participation (Forbes, 1977; Mullis, 2003). Quality and utility reflect the very basis on which the success of the NAEP program must be judged (Glaser et al., 1996). Yet, NAEP’s utility is often unclear (Sebring & Boruch, 1983) and there has been a persistent perception that NAEP data are under utilized (Glaser et al., 1996; Mullis, 2003; Pellegrino et al., 1999). Wilson and Blank (1999) argue that educators and researchers have found difficulties in NAEP data use, because of: 1) high

degree of complexity of the NAEP program; 2) methods of scoring and reporting NAEP results that differ from those used with state and local tests; 3) overreliance on composite ratings and rankings of states; and 4) not providing disaggregated data for analyzing issues of concern to teachers and schools (e.g., district- and school-level data).

Utility of NAEP is related to the extent to which its results are accessible and adequately disseminated to its diverse users (Glaser et al., 1996). During the period between the mid-1990s and the early 21st century, the development of new and powerful computing technologies, especially the Internet, led to innovations in dissemination of NAEP (Lazer, 2004). To make NAEP reports more accessible, NCES has posted diverse NAEP reports on its Web site concurrently with the public release of printed NAEP reports since 1996 (Lazer, 2004; NCES, 2005). The NAEP website has helped to expand and simplify the dissemination of NAEP information (Lazer, 2004). In particular, a useful function of the site is its ability to provide data access tools that allow examination of NAEP results in detail.

Further, the federally funded position of a NAEP state coordinator was established in each State Education Agency (SEA) in 2002 to support the implementation of the mandatory biennial reading and mathematics assessments at grades 4 and 8 under NCLB, according to a project officer for NAEP state coordinators at NCES. In the same year, a NAEP State Service Center was also established to provide ongoing support and training for NAEP State Coordinators (NCES, 2003). In general, the NAEP state coordinator (a) serves as the liaison between the SEA and NCES; (b) serves as the state's representative to review NAEP assessment items and processes; (c) coordinates the NAEP

administration in the state; (d) analyzes and reports NAEP data; and (e) coordinates the use of NAEP results for policy and program planning (NCES, 2003).

Evidently, within this context SEA's are in a better position to utilize NAEP data comprehensively for policy decision-making. Further, the NCLB requires states to participate in state NAEP at grades 4 and 8 in both reading and mathematics every two years. Therefore, it was assumed that NAEP data are currently used more widely within the context of a state. This case study was intended to empirically validate the assumption.

The NAEP Website

The NAEP Website (http://nces.ed.gov/nationsreportcard) is included in the NCES Website that contains a large amount of easily accessible education-related information. A full set of state and national results is available in an interactive database on the NAEP web site. The website is a useful location for viewing NAEP products and for using some web tools to search for, locate, and retrieve information on NAEP.

The structure of the website is subject-focused, and the site provides results for core subjects including reading, mathematics, science, writing, U.S. history, civics, geography, and the arts. The NAEP website consists of several sections including NAEP results, special tools, NAEP publications, and other resources, and these sections are discussed in detail in Chapter 4

Statement of the Problem

NAEP has become an integral part of the nation's evaluation of the condition and progress of education, by providing information on academic performance of groups and subgroups of students in specific content areas to policymakers, educators, and the general public at the national level (NCES, 2005). In particular, the expansion of NAEP to gather and report student achievement for sub-national units responsible for schooling such as states has increased its capacity for influencing state and local decisions on educational policies (Epstein, 1998).

From a state perspective, states need to seek multiple indicators to increase the validity of inferences drawn from student performance in achievement. Multiple measures of student progress over time provide much more reliable information. Researchers raised concerns that statewide tests used for accountability purposes cannot be validly used to provide information on the status of the educational system (McDonnell, 1994; Linn, 2000). In this sense, NAEP could provide one source of comparative data to validate performance trends on state assessments.

NAEP data and resources provided by the federal government have been made available to states, which could be used meaningfully to inform educational decisions in several ways. Unfortunately, it has been widely perceived that NAEP data are under utilized (Glaser et al., 1996; Mullis, 2003; Pellegrino et al., 1999). To date two research studies (Bullock & DeStefano, 1998; Sebring & Boruch, 1983) were completely dedicated to this issue. Sebring and Boruch (1983) examined NAEP data use at the federal, state, and local level and found that NAEP data were most useful to audiences with a national perspective. Bullock and DeStefano (1998) explored the usefulness of the

1992 TSA in reading through interviews with state directors of assessment. They found that the NAEP results were primarily used to inform policymakers at state and district levels and to provide information to the general public, teachers, and administrators. Yet, little research has been conducted in this field since the publication of those studies.

The state assessment component of NAEP fulfills a unique role in the U.S. educational system where there is a long tradition of local control over schools. Furthermore, the recent NCLB Act of 2001 forces NAEP to serve accountability purposes for states. For example, NAEP required states to participate in biennial State NAEP at grades 4 and 8 in reading and mathematics. The NCLB relies on assessment and accountability requirements as a major mechanism for improving achievement and enlarges the responsibility of state education agencies in implementing a standards-based accountability system (Linn, et al., 2002). It calls for NAEP to be used as a discussion tool to provide a context for state assessment results, but the specific role of NAEP in the NCLB accountability system has not been determined (Hombo, 2003; Lazer, 2004; Linn et al., 2002). Clearly, NCLB has brought new attention and scrutiny to NAEP results, requiring changes in the practices of many states (Lazer, 2004; Linn et al., 2002). It appears evident that under NCLB, NAEP's role becomes increasingly important and that state NAEP results are the focus of more attention and discussion (Hombo, 2003). Thus, it was speculated that the NCLB might facilitate the use of NAEP at the state level.

In addition, the Internet created the possibility for broadening dissemination of NAEP information beyond standard publishing taking advantage of the web's capabilities as the mode of releasing its results (Lazer, 2004). NAEP reports and data almanacs have been posted on the NAEP website to provide easier public access, and relevant data

access/analysis tools are also available. This system appears to represent an innovation in NAEP data dissemination and help facilitate use of NAEP data by NAEP audiences more than ever before.

In this context, this study was intended to investigate the use of NAEP data by state education personnel in terms of informing state educational decisions in the recent NCLB accountability system. The study started with examination of the potential utility of NAEP information provided on the NAEP website from the state perspective. The study focus was then given to exploring how state education personnel perceive the usefulness of NAEP and how NAEP data are actually used in a state context. In this study, the term "NAEP data" included not only the results of the NAEP assessment but also NAEP methodology and materials. Interviews and content analyses of institutional documents were conducted as research strategies for collecting information on perceived utility by state education personnel and actual evidence of NAEP data use.

In summary, this study first examined the nature of the NAEP website as a source of NAEP data that could be used to inform state policy and then conducted a case study of how NAEP data are currently utilized to inform policy decisions and program planning in an SEA. A case study approach provided thick description and detail of the case under study. The primary research questions guiding this study included:

1. What is the nature of the NCES website in terms of NAEP as sources of data to inform state educational decisions?
2. What are the state education personnel's perceptions of the use of NAEP in making informed educational decisions?

3. How are NAEP data used in supporting the state in responding to current issues in education?

Theoretical Framework

This case study is grounded in both constructivist and interpretivist paradigms. These two approaches share the notion that to understand the world of human action, an inquirer should capture the process of meaning construction of social actors s/he studies and clarify what and how meanings are embodied in their language and actions (Schwandt, 1994).

According to Patton (2002), the constructivist perspective begins with the premise that the world of human perception is shaped by cultural and linguistic constructs. “Social constructionism refers to constructing knowledge about reality, not constructing reality itself” (as cited in Patton, 2002, p. 96) and all reality is socially constructed (Crotty, 1998). Worlds in which different people live are distinct worlds, and thus there are as many world realities as there are conceptualizations of it (Sears, 1992). In the constructionist view, it is evident that different people may construct meaning in different ways even in relation to the same phenomenon (Crotty, 1998; Patton, 2002). The interpretive perspective holds that human behavior is purposive and that social actors construct and interpret their own behavior and that of their fellows (Schwandt, 1994). Further, Schwandt maintains that interpretivists themselves construct a reading of the meaning-making process of the people they study.

The two perspectives, therefore, enabled the researcher to grasp and further seek explanations of, the meanings and perceptions of NAEP information that different people have constructed within the context of their particular language and culture in an SEA.

Culture was defined here as a specific context within which social events, behaviors, or processes could be captured and interpreted (Schwandt, 1994). In this context, these perspectives provided a framework for this qualitative inquiry within which to collect and analyze the perceptions and perspectives of state education personnel about the usefulness of NAEP. Further, these perspectives allowed the researcher to acquire an in-depth understanding of how their experiences have subsequently led to the way they use NAEP data in a state context.

Significance of the Study

NAEP's utility is essential to the continued success of the NAEP program. Accordingly, NAEP has endeavored to report and disseminate its findings in the ways that could reach its audiences fully and be useful to them. However, there exists very limited empirical evidence that this challenging task has been successful. This study addresses this issue by examining state education personnel's perceptions of the usefulness of NAEP and their use of NAEP data.

Since its first administration of State NAEP in1990, NAEP has been speculated to be used for educational policy determinations particularly in a state context. Bullock and DeStefano (1998) explored the usefulness of the 1992 TSA in reading through interviews with state directors of assessment and revealed that the NAEP results were primarily used to inform policymakers at state levels. However, they did not specifically examine the ways in which NAEP data were used to inform state policy decisions. This study

investigated how specifically NAEP results and resources were utilized for educational policy decisions by state education personnel.

In these years, although NAEP's new role under the NLCB did not change the stakes for students at all, it further changed the nature of NAEP and created the impression that the assessment is now high stakes for states (Bourque, 2004; Hombo, 2003). The NCLB's mandate for states to participate in biennial state NAEP reading and mathematics assessments at grades 4 and 8 may facilitate the use of NAEP in a state context. In addition, dissemination of NAEP data through the Web along with special web-based tools might further foster the use of NAEP. However, no empirical evidence existed on NAEP data use under the NCLB. This study explored how useful state education staff consider NAEP and how NAEP data are used in the NCLB accountability system at the state level.

The present study was intended to investigate the actual use of NAEP by an SEA under NCLB in relation to the potential utility of NAEP information placed on the NAEP website. The basic need for this study stemmed primarily from the aforementioned significance of effectively communicating the complex NAEP data to a variety of intended audiences. In this sense, the current study was intended to provide some empirical evidence that could support or refute the claim that NAEP data would be more utilized if they were easier to access and made more useful. Simultaneously, the focus of this study was on examining how the NCLB's demand on NAEP for states might affect the use of NAEP by the SEA. The findings of this study are expected to provide some insight into how to narrow the gap between potential and actual uses of NAEP.

CHAPTER II

REVIEW OF THE LITERATURE

In order to communicate dependable information about performance of American students to the nation, NAEP has made a substantial investment in reliably and validly measuring student achievement against a common set of challenging, consensus-based standards. Simultaneously, NAEP has placed continuous effort into improving the utility of its results by potential users, but the task remains challenging. Beginning with 1994 assessments, NAEP posted NAEP reports on its website and further web-based data analysis tools were placed on the site to speed access to needed information. In addition, the recent NCLB legislation has increasingly forced states to focus more on the performance of their students on state NAEP. Within this context, it was assumed that NAEP data are currently used more than ever in the state level.

This literature review focused on NAEP's history, the characteristics of the NAEP program such as assessment content, conduct, and reporting, the use of NAEP data, and methodological approaches to analyzing qualitative data, in an effort to help readers better understand the issues surrounding the research problems being addressed in this study. This literature review chapter consists of six sections as follows:

- Assessment, accountability, and achievement
- The history of NAEP
- The NAEP program
- Utility of NAEP data
- Methodological approaches guiding this study

- Discussion and conclusions

Assessment, Accountability, & Achievement

Tests and assessment have been used as key elements in many reform efforts, although the nature of tests has changed. They might be administered in many different forms and be used in diverse ways in accountability systems aiming to improve education. According to Linn (1998), there are several reasons for the use of assessment as an agent of reform by policy makers: (1) tests and assessments are relatively inexpensive; (2) assessment can be externally mandated; (3) testing and assessment changes can be rapidly implemented; and (4) test results are visible.

The national standards-based reform movement, which emerged in the1990s, differs from the earlier efforts in at least three significant ways (Linn, 1994): (1) the federal government is playing a more active role; (2) there is a greater emphasis on "high standards for all students" (content, performance, and opportunity-to-learn); and (3) performance-based assessments are emphasized. The prominence of these three characteristics is exemplified by the provisions of America 2000 (U.S. Department of Education, 1991) and Goals 2000: Educate America Act of 1994. In particular, the third goal, "Students will leave fourth-, eighth-, and twelfth-grades having shown competency in English, mathematics, science, …", has been interpreted as requiring a national system of assessments (McLaughlin & Shepard, 1995).

Large-scale assessments are designed to meet certain purposes under constraints that include: (a) providing reliable scores for individuals and groups; (b) sampling a broad set

of content standards within a limited testing time; (c) determining whether students are meeting the standards set for their achievement; (d) using this feedback to improve teaching and learning practices; and (e) offering cost efficiency in terms of development, scoring, and administration (Chudowsky & Pellegrino, 2003; Herman, 1997; Ungerleider, 2003). For more than 50 years, large-scale assessment has been used to promote system accountability and improvement for educational systems historically committed to local control.

Likewise, educational stakeholders including policymakers, educators, the general public expect large-scale assessments to serve a variety of purposes including (1) measuring student learning; (2) holding education systems accountable; (3) signaling worthy goals for students and teachers to work toward; and (4) providing useful feedback for instructional decision making (Chudowsky & Pellegrino, 2003). To meet these demands placed on large-scale assessments, it is necessary to explore and implement alternative approaches to these assessments focusing on capturing the complexity of cognition and learning beyond the traditional paper-and-pencil format (Chudowsky & Pellegrino, 2003; Herman, 1997).

In this era of educational accountability, assessments might be used in a variety of ways to improve education. However, accountability systems relying heavily on a single test might decrease the validity of inferences based on observed gains in achievement (Linn, 1998). It is important to evaluate the validity of gains obtained, particularly if high-stakes are attached to tests (Koretz, McCaffrey, & Hamilton, 2001; Linn, 1998).

There have been concerns that high-stakes test results yield inflated impressions of student achievement and that assessments in high-stakes conditions lead to a narrowing

of the curriculum and teaching to the test (as cited in Smith & Fey, 2000, p. 339; Linn, 1998; Madaus, 1985). For example, a physician, John Cannell, brought to public attention the incredible finding, known as the Lake Wobegon effect, that essentially all states were reporting that their students were scoring above the national norm (Koretz, 1988). Further evidence of lack of generalizability of accountability test results is provided by Koretz et al.'s study (2001). They found that in the first year a new test was administered, there was a sharp drop in scores but that the drop was followed by gains on administrations in subsequent years. The findings of their study suggest that high-stakes assessments might distort test scores.

The NAEP's new role in NCLB, which mandates states' participation in biennial reading and mathematics State NAEP at grades 4 and 8, does not change the stakes for students, but NAEP becomes increasingly higher stakes for states than before. As noted previously, studies have indicated that when a test has high-stakes attached to its results, the school curriculum narrows and teachers tend to teach to the test (Madaus, 1985). These findings might raise some concern that a potential federally funded high-stakes national assessment under NCLB could have a negative influence on curriculum.

The NCLB indicates that NAEP might be used to confirm AYP measures on state test results. However, previous studies suggest that relying on NAEP to determine performance levels and AYP objectives might be problematic since the performance levels of NAEP and state assessments are not comparable (Linn et al., 2002; NAGB, 2002b).

The History of NAEP

To gain historical insights and understand some challenges that encounter NAEP today, it appears necessary to examine the early forces that helped shape and direct NAEP. In 1963 Francis Keppel, U.S. Commissioner of Education, appointed a committee to explore options for assessing the condition and progress of American education (Jones, 1996; Pellegrino et al., 1999). In 1966 Keppel and Ralph Tyler who was the committee's chair proposed a national assessment that would provide information about student achievement across the nation (Epstein, 1998; Fitzharris, 1993; Jones, 1996). The concept of a national assessment raised key issues associated with the purpose of the assessment, the role of the federal vs. the state government in education, and the source of funding (Fitzharris, 1993; Jones, 1996). Some of the issues resurfaced during different periods in the history of the NAEP.

The major issue was the lack of a clear understanding or consensus about the purpose of the national assessment (Fitzharris, 1993). It is interesting to note that Tyler and Keppel initially had different but compatible purposes for a national assessment (Epstein, 1998; Fitzharris, 1993). Tyler expected a national assessment to serve as a barometer of student learning over time, while Keppel desired national data to meet the intent of the legislation that had created the United States Office of Education and to support his educational initiatives in Congress. Many professional educators opposed a national test since they feared that the test would be the genesis of a centrally controlled national curriculum. In fact, opposition made public by professional educators during NAEP's

developmental years helped shape the assessment not to be designed as a national test substantially influenced by the federal government (Epstein, 1998).

The lack of a clearly defined purpose made it difficult to gain political and financial support for developing the idea of a national assessment (Fitzharris, 1993). When Keppel proposed the idea of a national assessment, he was not able to obtain funding from the federal government and thus he sought financial support from the Carnegie Corporation and later from the Ford Foundation. Private funds supported the development efforts for the first three and one-half years. The Carnegie foundation sponsored the Exploratory Committee on Assessing the Progress of Education (ECAPE) with additional support from the Ford Foundation. The ECAPE with Ralph Tyler as chairman was formed to design a system to monitor the progress of the nation's education (Jones, 1996; Lehmann, 2004). The Technical Advisory Committee (TAC) proposed instrumentation, sampling, exercise development, data analysis, and reporting (Jones, 1996; Lehmann, 2004). Despite strong opposition from professional associations, the development of NAEP finally led to the first NAEP assessment in 1969.

Periodic reauthorization of NAEP has generated debates about the purpose of the national assessment and appropriate levels of funding (Fitzharris, 1993). NAEP has evolved over three decades as the only national measure of student progress over time, but has been "pulled and pushed" by different political and social needs (Fitzharris, 1993, p. 5). The proposed NAEP was an objective-referenced assessment, very similar to criterion-referenced assessment, intended to be distinct from traditional norm-referenced cognitive testing. Many innovative features were recommended by TAC and endorsed by ECAPE (Jones, 1996; Lehmann, 2004). NAEP's design and methodology have changed

to reflect technological advances in measurement and contemporary educational reforms (Berends & Koretz, 1995/1996; Burstein et al., 1995/1996; Mullis, 2003; Phillips et al., 1993). The addition of state NAEP since 1990 has provided a new dimension for NAEP and has expanded the scope of the national effort (Epstein, 1998; Fitzharris, 1993; Jones, 1996).

In 1968, ECAPE was promoted to the Committee on Assessing the Progress of Education (CAPE), and an administrative home for NAEP was established at the Education Commission of the States (ECS) in Denver when no other agencies were unwilling to accept the controversial project. Because ECS represented the interests of the states, it was the linchpin to gaining the needed cooperation of the project and "this structure appeared to be a perfect blend of state and federal efforts" (Fitzharris, 1993, p. 50). TAC became ANAC (Analysis Advisory Committee) and its focus shifted from assessment design to data analysis and reporting.

In 1972, oversight of NAEP was transferred to NCES from the National Center for Educational Research. However, the struggle for power over NAEP between federal officials and NAEP staff ensued. In 1983, the Educational Testing Service (ETS) replaced ECS as the primary contractor for NAEP, and significant changes resulted. The takeover resulted in NAEP's expansion and transformation, placing it in a position that might inform educational policy and curricular reforms (Epstein, 1998). In 1988, Congress authorized a Trial State Assessment (TSA) and established a National Assessment Governing Board (NAGB) to assume responsibilities for setting policy on NAEP independently of any federal agency. NAGB made efforts to improve the

interpretation of results by establishing performance standards designed to represent basic, proficient, or advanced levels of achievement.

As ETS took over the assessment management responsibilities in 1983, significant changes in the NAEP design were made. New features included BIB-spiraling, item response theory (IRT) scaling, grade-based data, and collection of richer background information. These changes had some implications for sample design, weighting, variance estimation, and field operations (Elliott & Phillips, 2004). In its contract proposal, ETS asserted, "We are concerned not only with improving the meaning and interpretability of the assessment results but also with enhancing their utilization in affecting educational policy and practice" (as cited in Epstein, 1998, p. 5).

In 1987, the Alexander-James study group formed by Secretary of Education William Bennett recommended that NAEP collect representative data on achievement in each of the fifty states thereby providing state-to-state comparisons (Elliott & Phillips, 2004; Epstein, 1998). The new NAEP design made it possible for NAEP to be redirected toward state-to-state comparisons of student achievement (Epstein, 1998). In 1988, acting on the report by Alexander and James, Congress enacted legislation authorizing state NAEP assessments (a Trial State Assessment) and established a National Assessment Governing Board (NAGB). Permitting comparisons of NAEP results below national and regional levels was against the very protections originally built into the National Assessment (Epstein, 1998).

A Nation at Risk, the 1983 report from President Reagan's National Commission on Excellence in Education, created a strong incentive for policymakers to pay attention to education. At the 1989 education summit, which was a reaction to a perceived national

crisis, President George Bush and fifty state governors agreed to establish clear national performance goals. The third of the original six goals that American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter implied the need for setting national performance standards.

Soon after the summit, the NAGB, NAEP's policymaking body, established performance standards for NAEP and student achievement has been reported using the NAGB achievement levels since 1990 (Epstein, 1998). Again, the national education movement of the 1980s, the Governors' Education Summit in 1989, and in particular, the expansion of NAEP to include state-level assessments in 1990 drew extensive attention, by stakeholders and the general public, to NAEP (Linn & Dunbar, 1992; Resnick, 1999). Further, the NCLB act of 2001 has brought more attention to state NAEP results. NCLB requires states to participate every other year in state-level administrations of the NAEP in reading and mathematics at grades 4 and 8.

The NAEP Program

The National Assessment Governing Board (NAGB), appointed by the Secretary of Education but independent of the department, governs and oversees the program. The board, whose members represent NAEP's varied audiences, sets policy for NAEP. For example, NAGB selects the subject matters to be assessed, establishes the frameworks, develops guidelines for reporting, and gives direction to NCES. The NCES has coordinated the assessment since 1983 through a series of contracts, grants, and cooperative agreements with ETS and other contractors.

For 1996-2002, ETS was responsible to develop the assessment instruments, to score student responses, to analyze the data, and to report the results. Westat was responsible for sample design and selection, data collection and field operations, and survey weighting. Westat has either been under subcontract to ETS or contracted directly with NCES. National Computer Systems (NCS) as a subcontractor to ETS was responsible for printing and distributing the assessment materials and for scanning and scoring students responses. American Institutes for Research (AIR) as a subcontractor to ETS was responsible to develop the background questionnaires.

For the 2003-2006 assessments, ETS is still responsible for ensuring coordination among the contractors on the team and maintaining data for tracking the program progress. ETS was also awarded the contract for design, analysis, and reporting. AIR was awarded the contract for item development. Materials preparation, distribution, and scoring will be conducted by Pearson. Westat, Inc., was awarded the contract for sampling and data collection for both national and state NAEP assessments.

NAEP was an annual assessment between 1969 and 1979, but was administered every two years between 1980 and 1996. In 1997, NAEP returned to annual assessments. After the prohibition on district and school assessments was lifted in the 1994 reauthorization of NAEP, the Trial Urban District Assessment (TUDA) has been conducted since 2002 on an experimental basis. What follows is a brief description of diverse aspects of the NAEP program, the review of the research studies on alignment between standards and NAEP assessments, the psychometric quality of NAEP, utility of NAEP data, and methodological approaches to guiding this study, as well as discussion and conclusions.

Overview

National NAEP National NAEP reports information about student achievement at grades 4, 8, and 12 for the nation, specific geographic regions of the country (prior to 2003: Northeast, Southeast, Central, and West; since 2003: Northeast, South, Midwest, and West) and specific subgroups of the population (by gender, race or ethnicity, highest level of parental education, and type of school). Before 1980, it was an annual assessment, but from 1980 through 1996 it was administered every two years. In 1997 it returned to annual assessments. Students are drawn from both public and nonpublic schools.

These tests are constructed based on the frameworks that are developed by the NAGB and updated on a regular basis to reflect changes in curriculum and pedagogical thought (Beaton & Zwick, 1992; Johnson, 1992; Linn & Dunbar, 1992). Thus, the content and nature of the NAEP instrument evolve to reflect current instructional practice, new educational issues, and new assessment technology, thus greatly reducing the ability of the assessment to measure change over time (Johnson, 1992; NCES, 1999b). However, short-term trends can be measured in many of the NAEP subjects (e.g., mathematics and reading). NAEP assessments include a large percentage of constructed-response questions and questions that require the use of calculators and other materials in addition to multiple-choice items (Beaton & Zwick, 1992; Silver & Kenney, 1993a; NAGB, 2002a). Innovative types of questions have been used in assessments such as the arts and science to measure students' ability to perform hands-on tasks.

State NAEP NAEP has reported results for student performance of public schools at grades 4 and 8 at the state level since 1990, because national samples are not designed to support the reporting of accurate and representative state-level results. States can monitor their own progress over time in selected subject areas, and compare student achievement with other states and the nation. State assessments in content are identical to the main assessment, but separate representative samples of students in each participating state are selected (Beaton & Zwick, 1992).

Since 2002, a combined sample of public schools has been selected for both state and national NAEP to reduce state burden by decreasing the total number of schools participating in state and national NAEP (Mullis, 2003). Thus, the national sample is a subset of the combined sample of all participating states, plus an additional sample from the states that do not participate in the state NAEP and a representative sample of private schools. NCLB requires states that receive Title 1 funding to participate in state NAEP in reading and mathematics every two years. State participation in other state NAEP subjects, science and writing, remains voluntary.

Long-Term Trend NAEP Long-term trend assessments give information on changes in achievement over extended time periods. They were administered to students at ages 9, 13, and 17 in mathematics, reading, and science and grades 4, 8, and 11 in writing, but since 2004 they are administered nationally every four years and report student performance at ages 9, 13, and 17 only in mathematics and reading. Results are reported as performance levels using scale scores, not in terms of achievement levels. For each of the subject area scales, performance levels were set at 50-point increments from

150 through 350 (called “scale anchoring”). The five performance levels (150, 200, 250, 300, and 350) were then described in terms of the knowledge and skills likely to be demonstrated by students who reached each level.

The assessment instrument does not change based on changes in curricula or in educational practices in order to measure trends of student achievement (Johnson, 1992; NCES, 1999b). Therefore, these assessments use subject-area frameworks and questions that remain relatively constant over time (Johnson, 1992; Silver & Kenney, 1993a). In other words, long-term trend assessments are used to estimate changes in performance from previous assessments using the same methodology and population definitions as in those assessments (Johnson, 1992).

However, several changes were made to the long-term trend assessment in 2004 to align it with best current assessment practices and with policies by NAGB. As a result, two assessments were given in 2004: 1) a modified assessment that contained many changes from previous assessments; and 2) a bridge assessment that was used to link the modified assessment to the 1999 assessment so the trend line could be continued. The modified assessment included the following changes:

- Removal of science and writing items,
- Inclusion of students with disabilities and English language learners,
- Replacement of items that used outdated contexts,
- Creation of a separate background questionnaire,
- Elimination of "I don't know" as a response option for multiple-choice items, and
- Use of assessment booklets that pertain to a single subject area (whereas in the past, a single assessment booklet may have contained reading and mathematics

In 2004, students were randomly assigned to take either the bridge assessment or the modified assessment. The bridge assessment replicated the instrument given in 1999 and used the same administration procedures to maintain trend lines across years. The modified assessment will provide the basis of comparison for all future assessments.

Design

Student Sample Design Because relatively small samples of students selected must be representative of the entire population, a complex sampling scheme, rather than simple random sampling, is used to collect NAEP data (Beaton & Zwick, 1992; Johnson, 1989, 1992). The national assessment has used a stratified multistage probability sampling design. For national and long-term trend NAEP assessments, a three-stage sample design is employed: (1) the sampling of students from (2) selected schools within (3) selected geographic areas, called *primary sampling units* (PSUs) consisting of counties of groups of counties.

In the first stage, PSUs are selected within each of four geographic areas (Northeast, Southeast, Central, and West) (Allen, Carson, & Zelenak, 1999; Johnson, 1989, 1992; Rust, 2004). The second stage of selection consists of elementary and secondary schools within each PSU. Private schools, schools with high-minority students, and SD/LEP students (since 1996) are deliberately sampled at a higher rate to assure a sufficient sample size for increased precision of estimates for these subpopulations, but not oversampled for the long-term trend assessments (Allen et al., 1999). In the third stage, students are selected within those schools chosen. Beginning in 2002, however, the national sample was obtained by aggregating the samples from each state to reduce state

burden (Mullis, 2003). Since 2003, state and national samples are drawn in the same way in odd-numbered years, while national samples are drawn using the three-stage method in even-numbered years. For samples to be representative of the population as a whole, the data from the students in the oversampled schools are weighted during analysis (Johnson, 1989, 1992).

For the state assessment, a sample of public schools and students in the fourth and eighth grades are selected to represent a participating state based on a two-stage sample design. Schools are stratified hierarchically by small- or large-district status, urbanization, school size class (measured by student enrollment), and percent minority (Allen, Donoghue, & Schoeps, 2001). The students within a school are randomly sampled from a list of students within the appropriate grade (four or eight).

Instrumentation Design NAEP produces estimates of performance distributions for the nation or state and for subpopulations (not for individual students) for each grade or age level. NAEP assessments include many cognitive items to measure a large and diverse body of student knowledge, and thus each student is presented a subset of the items from the entire pool. This design also increases statistical efficiency and reduces a burden placed on the student and the school (Beaton & Zwick, 1992; Burstein et al., 1995/1996; Johnson, 1989, 1992; Koretz & Deibert, 1995/1996; Mislevey, Johnson, & Muraki, 1992; Mullis, 2003).

NAEP uses a focused, balanced incomplete block spiraling (BIB-spiraling) scheme to assign test items to students (Beaton & Zwick, 1992; Johnson, 1989, 1992; Linn & Dunbar, 1992). BIB-spiraling design begins by dividing the items within a subject area

into units, called blocks, where each block is designed to take the same amount of time for completion. The blocks are assembled into a booklet that contains background questions and two or three blocks of subject area items (Beaton & Zwick, 1992; Johnson, 1992). “Focused” requires each student to answer questions from only one subject area. “BIB” ensures students receive different interlocking sections of the assessment forms, and “spiraling” involves interweaving the booklets in a systemic sequence to ensure approximately equal numbers of each booklet are used (Allen et al., 1999; NCES, 1999b). Consequently, each student in an assessment session receives a different booklet, thus markedly reducing cluster or order effects (Johnson, 1992; NCES, 1999b).

Four types of instruments are used in the assessment:

- Student assessment booklets, containing cognitive items and background questions (demographic and subject-specific)
- Teacher questionnaires
- School characteristics and policies questionnaires
- SD/LEP questionnaires (for national assessments from 1996, state assessments from 1998, long-term trend assessments from 2004)

NAEP Frameworks

A NAEP assessment is constructed based on a framework that is the blueprint that determines the content to be assessed and guides the development of the assessment instrument (Mullis, 1992). Since its inception in 1988, NAGB has had responsibility for developing the frameworks that underlie the assessments. Frameworks capture a range of subject-specific content and thinking skills needed by students to deal with complex

issues they encounter inside and outside their classrooms (Mullis, 1992; NAGB, 1995, 2002a). The content goals and objectives in the framework are intended to maintain a balance between what students are currently learning in schools across the country and the knowledge and skills that subject experts believe students should acquire (NAGB, 1995, 2002a; NCES, 1994).

NAEP frameworks take into consideration broad national input from a variety of stakeholders and experts. The framework for each subject area is determined through a *consensus process* by a planning committee (teachers, curriculum specialists, subject-matter specialists, and disciplinary researchers) and a steering committee (a broad group of education administrators, policymakers, parents, subject-area experts, and members of the general public) (NAGB, 2002a; Pellegrino et al., 1999). NAGB has made major efforts to develop thoroughly updated assessment frameworks for each assessment based on widespread, in-depth consensus processes in order to ensure they are appropriate for current educational requirements (Mullis, 1992; NAGB, 1995; NAGB 2002a; NCES, 1994). Most NAEP frameworks specify that the subject-area assessments be constructed around two or more dimensions. In science, for example, two major dimensions are "fields of science" and "ways of knowing and doing" supplemented by two underlying dimensions, "nature of science" and "themes." In mathematics, two primary dimensions, "content' and "mathematical abilities," are supplemented by a dimension designated as "mathematical power."

Developing a framework involves the following elements (NCES, 1999b):

- Widespread participation and reviews by educators and state education officials in the particular field of interest

- Reviews by steering committees whose members represent policymakers, practitioners, and the general public
- Involvement of subject supervisors from the education agencies of prospective participants
- Public hearings
- Reviews by scholars in that field, by NCES staff, and by a policy advisory panel

Scaling Procedures

Through scaling, the performance of students assessed in a subject area or subarea can be summarized on a single scale even when different students have been administered different items (Allen et al., 1999, 2001). The scaling process takes the information collected from all of the items administered and summarizes this information onto the subscales defined in the framework. The separate subscale scores are then combined in a weighted formula to reflect their relative importance as defined by the framework at each grade level within a subject area, which results in overall proficiency scores (Allen et al., 1999, 2001; Glaser et al., 1996).

In addition to NAEP's collection of incomplete data through the BIB design, the NAEP's design is further complicated by the stratification and clustering of schools, unequal probabilities used for selecting minority students, and the higher proportion of constructed-response items (Johnson, 1989, 1992; Glaser et al., 1996; Pellegrino et al., 1999). NAEP's complex design led to the use of an elaborate approach for developing NAEP's score scale and for estimating population characteristics in order to minimize biased estimates (Beaton & Zwick, 1992; Johnson, 1992; Mislevy et al., 1992).

Student item responses collected have been summarized for analysis and reporting using an item response theory (IRT) scaling method, which is a test analysis procedure that assumes a mathematical model for the probability that a given examinee will respond correctly to a given exercise (Beaton & Zwick, 1992; Johnson, 1989, 1992; Mislevy et al., 1992). IRT is used to develop a scale that spans age and grade levels and that defines proficiencies in terms of a latent variable without requiring that all students take the same items (Beaton & Zwick, 1992; Johnson, 1989, 1992; Mislevy et al., 1992). Instead of using methods that produce estimates of individual proficiency, NAEP employs marginal estimation procedures that yield consistent estimates of major population characteristics directly from item responses (Beaton & Zwick, 1992; Johnson, 1992).

NAEP has adopted “plausible values,” five random draws, taken from the distributions that are estimated in the marginal analyses and used these plausible values for estimating population characteristics (Beaton & Zwick, 1992; Johnson, 1989, 1992; Mislevy et al., 1992). In other words, plausible values are not test scores for individuals, but are offered only as intermediary computations to estimate population characteristics (Allen et al., 1999, 2001).

In order to combine the variety of item types onto a single scale, three separate scaling models were used beginning in 1992. These are: one for multiple-choice items (three-parameter logistic model), one for dichotomously-scored constructed-response items (two-parameter logistic model), and one for extended constructed-response items (partial credit model) that receive different scores depending on the quality of the response.

Performance Standards

The Goals 2000 legislation, the reauthorization of Title 1 of the Elementary and Secondary Education Act, and recent policies of NAGB all call for the establishment of explicit standards for what students should know and be able to do (Koretz & Deibert, 1995/1996). Current reforms tie performance standards to specific assessments and place great reliance on standards-based reporting (Burstein et al., 1995/1996; Koretz & Deibert, 1995/1996). Performance standards on NAEP, called achievement levels by NAGB, have their origin in the 1988 reauthorization of the National Assessment, which created NAGB and gave it the responsibility for identifying appropriate achievement goals for each age and grade level in each subject area to be assessed (Berends & Koretz, 1995/1996; Burstein et al., 1995/1996; Hambleton & Cadman, 1992; Koretz & Deibert, 1995/1996; Linn & Dunbar, 1992).

NAGB decided to develop the means for reporting NAEP results in terms of what students *should* know and be able to do. The decision means that NAEP results must describe what students should know in addition to describing what students know. In previous years, NAEP results had been reported in terms of percents correct or scale scores, which made it impossible to determine whether performance is good enough (Koretz & Deibert, 1995/1996). The use of achievement levels may contribute to moving NAEP away from being a low-key indicator toward becoming an instrument of reform (Linn & Dunbar, 1992).

The first use of performance standards-based reporting of NAEP results was with the 1990 mathematics assessment. NAGB developed policy definitions for three separate achievement levels: Basic (partial mastery), Proficient (competency over challenging

subject matter), and Advanced (superior performance beyond Proficient). The proficient level is intended to reflect the reasoning used in framing National Education Goal 3 (which calls for students to demonstrate competency over challenging subject matter). The policy definitions are generic, free of both subject-matter content and grade levels.

Starting with these policy definitions, panels of judges develop "operational definitions" of the levels and identify cut-scores for the levels based on ratings of the NAEP item pools for each content area and for each grade. In other words, the achievement levels setting process focuses on judgmentally estimating the skills and knowledge that define the lower bound of each policy definition in terms of content and skills from relevant NAEP frameworks. Such skill definitions are then mapped onto the NAEP reporting scale, using estimates of performance on actual NAEP items (Reckase, 1998). A panel of teachers, non-teacher educators, and non-educators establish those levels for each grade using a several-step method (Koretz & Deibert, 1995; Pellegrino et al., 1999):

1. Panels of judges are given simple "policy definitions" of what students should be able to do to be considered as having reached each level.
2. Judges are then asked to estimate what proportion of the students at the border line of each of three achievement levels would be able to answer specific items correctly (for each multiple-choice item panelists estimate the probability that a hypothetical student at the boundary of a given achievement level will get an item correct, while for constructed-response items they estimate mean item scores for students at each of these same three boundaries).

3. Item judgments are averaged across items and cutscores are reached that distinguish the levels of performance.
4. The average judgments on a final set of ratings (estimated *p* values) are then mapped onto the NAEP scale.
5. On the basis of this process and other information, judges refine and elaborate on their descriptions of performance at each level.
6. Items are then chosen (based on a variety of considerations, including appropriateness of content and actual patterns of performance) to exemplify the levels.

The achievement levels have been controversial regarding the process and the outcome of the undertaking (Baker & Linn, 1997; Koretz & Deibert, 1995/1996; NAGB 2000; Pellegrino et al., 1999). In general, critics agree that: 1) the judgment task posed to raters is too difficult and confusing; 2) there are internal inconsistencies in raters' judgments for different item types; 3) neither the descriptions of expected student competencies nor the exemplar items are consistent with actual student performance; 4) NAEP item pools are not adequate to reliably estimate performance at the advanced levels, and 5) appropriate validity evidence for the cutscores is lacking. Nonetheless, it has been widely perceived that the standards-based reporting is likely to enhance the usefulness of NAEP results for policy makers and educators.

Administration

Westat was responsible for sample design and selection, data collection, and field operations as a subcontractor to ETS when ETS was awarded the NAEP grant in 1983. Westat has continued to be responsible for these activities since then, at times being under subcontract to ETS and at other times being contracted directly with NCES. Westat is responsible for the following field administration duties: 1) selecting the sample of schools and students; 2) developing the administration procedures, manuals, and materials; 3) hiring and training staff to conduct NAEP assessments; and 4) conducting an extensive quality assurance program (NCES, 1999b).

For the state assessments, NAEP legislation required each participating state to handle data collection activities. Westat employed and trained state supervisors to work with the state-appointed coordinators who carried out the necessary organizational tasks. In addition to training local administrators, Westat ensured quality control across states by monitoring 25 % of the sessions in states that had previously participated in NAEP and 50 % of the sessions in states that participated for the first time (NCES, 1999b). However, beginning with the 2002 assessment, the NAEP state assessment program was administered in schools by Westat field operations staff, and state and national samples were combined.

National Computer Systems (NCS) (Pearson since 2003) produces the materials needed for NAEP assessments as a subcontractor to ETS. NAEP guarantees the anonymity of participants, and student or teacher names are never recorded on assessment booklets. The subcontractor handles all receipt control, data preparation and processing, scanning, scoring activities, and delivering the assessment data files to ETS

for analysis and reporting. Using an image-processing and scoring system introduced during the 1994 assessment, the subcontractor scans the multiple-choice selections, the handwritten student responses, and other data provided by students, teachers, and school administrators, eliminating paper handling during scoring.

Accommodations

The enactment of Goals 2000, the Improving America's Schools Act (IASA), and the individuals with Disabilities Education Act (IDEA) brought a huge change to state standardized assessment programs (Martin, 2004). Since the legislation required large-scale assessment program to include limited-English-proficient and special education students who had been excluded from testing until then, states began developing appropriate accommodations. NAEP was exempt from this legislation, but began to be pressured by advocacy groups and the Office of Civil Rights to be more inclusive (Martin, 2004).

Finally, NAEP began using split samples to explore the effects of being more inclusive in the 1996 national mathematics assessment. This enabled the program to accomplish three key goals: to maintain data trends to the past, to study the effects of providing assessment accommodations, and to begin new trend baselines in which accommodations were allowed (Glaser et al., 1996; Martin, 2004). Three experimental conditions were established: 1) exclusion criteria that applied in prior assessments with no accommodations; 2) a revised set of criteria with no accommodations offered to students with disabilities (SD) and those with limited English proficiency (LEP); and 3) the revised set of criteria with accommodations. The design for the national science

assessment was similar, except that the old exclusion guidelines were not used since there was no trend line to maintain (a new science framework was introduces in 1996).

In the 1998, 2000, and 2001 assessments, the sample was split equally between conditions 2 and 3 for students in which trend results were expected to date back to assessments in 1996, while all students were assigned to Condition 3 for assessments with no expectations for trend results. Beginning with the 1998 assessments, state samples were designed to be half samples (Conditions 2 and 3). Since the 2002 assessments, only inclusive samples have been used. Further, long-term trend assessments provided accommodations in 2004, thus assessing a greater proportion of students.

The decision to exclude any of these students is made by school personnel who complete the SD questionnaire or LEP questionnaire for the students selected for the sample. Criteria for including SD students and LEP students in NAEP assessments are as follows:

1. SD students should be included in the NAEP assessment unless:

- The IEP (Individualized Education Plan) team or equivalent group has determined that the student cannot participate in assessments such as NAEP, *or*
- The student's cognitive functioning is so severely impaired that he or she cannot participate, *or*
- The student's IEP requires that the student be tested with an accommodation that NAEP does not permit, and the student cannot demonstrate his or her knowledge of reading or mathematics without that accommodation.

2. LEP students should be included in the NAEP assessment unless:

- The student has received reading or mathematics instruction primarily in English for less than 3 school years including the current year, and
- The student cannot demonstrate his or her knowledge of reading or mathematics in English even with an accommodation permitted by NAEP.

Background Questions

In addition to measuring student performance for specific academic areas, NAEP collects background information from students, teachers, and school administrators to provide a context for reporting the NAEP findings. The background information that NAEP collects helps to give context and meaning to the cognitive results. Four general sources provide context for NAEP results: 1) student questionnaires; 2) teacher questionnaires; 3) school questionnaires; and 4) SD/LEP questionnaires.

Some background variables relate to basic demographics including gender, region, race/ethnicity, type of community, type of school, and parental education. Other background information focuses on five major educational policy areas: instructional content, instructional practices and experiences, teacher characteristics, school conditions and context, and conditions outside school that affect learning and instruction (Baker, 1995; Blank, Porter, & Smithson, 2001).

Background variables are identified during the consensus development process and that the committees examine current research and bring their knowledge and experience to bear on developing background questions (NCES, 1994). NAEP ensures that the background questions, in addition to being grounded in research, do not infringe on

respondents' privacy and that the answers can help inform the debate about educational reform (NCES, 1999b). Under NCLB, NAGB is responsible for selecting and approving all of NAEP's background questions (NAGB, 2003).

In sum, the background questions are used: (1) to define subgroups of the examinee population for reporting purposes; (2) to support research on the factors that affect NAEP scores; and (3) to improve the estimation of the students' proficiency distribution on the cognitive component of the NAEP (Reckase, 2002).

Alignment between Standards and NAEP tests

Alignment is used to characterize the match among multiple components of the educational system (Webb, 1997). This section will focus on research studies that investigate the extent to which NAEP is aligned with its framework and national standards. Two studies examined alignment of NAEP and its framework (Pearson and DeStefano, 1993; Silver and Kenney, 1993a), and a study conducted by Silver and Kenney (1993b) identified the degree of alignment of NAEP and the NCTM Standards.

Silver and Kenney (1993a) conducted a study designed to analyze the content and curricular validity of the 1992 NAEP framework and items used in the Trial State Assessment (TSA) in mathematics. The development of the framework was coincident with the drafting of the National Council of Teachers of Mathematics (NCTM) *Curriculum and Evaluation Standards for School Mathematics*, which focused heavily on problem solving, critical reasoning, and communication of mathematical ideas (Mullis, 1992).

The framework was a content-by-process matrix specifying five content areas: numbers and operations, measurement, geometry, data analysis, statistics, and probability, and algebra and functions. The three process or ability areas include conceptual understanding (CU), procedural knowledge (PK), and problem solving (PS). According to the authors, although the framework for the 1992 NAEP mathematics remained the same as that for the 1990 NAEP, some important changes were made at the item level. It used a new item type, called extended open-ended, in addition to multiple choice and open-ended, and its responses were scored using a focused holistic-scoring scheme. Manipulative items were also included.

The purpose of the study was to examine the adequacy of the framework and the grade 4 and grade 8 TSA items, with respect to the identified ability and content categories, the intended weighted distribution of items by category, and the actual distribution of items per cell within the framework matrix. The framework and items were examined by an expert panel of seven mathematics education professionals for grade 4, while the authors independently examined the mathematics framework and items for grade 8.

For grade 4, the researchers found that the intended versus actual distribution of items by content was approximately equal, while with respect to mathematical ability, there were fewer procedural knowledge (PK) items and more problem solving (PS) items on the NAEP assessment than were specified in the Mathematics Framework (PK: 30% intended vs. 20% actual; PS: 30% intended vs. 38% actual). Fewer PK items were found in all content areas, especially in categories of geometry (GY) and algebra and functions (AF).

The seven panelists working as pairs classified the NAEP items by ability and content. The pairs were given 4 of the 13 blocks of items, with one block classified by all pairs to provide a measure of agreement. With respect to the match between the actual NAEP and expert panel (EP) classifications for *content*, the EP members agreed highly with the NAEP classifications (agreement of 89 %). The authors argued that most of the disagreements involved the categories of numbers and operations (NO) versus algebra and functions (AF) and NO versus measurement (MT), resulting from ambiguity regarding definitions of each content area. The overall percent agreement between the NAEP and EP classifications regarding *ability* was 69 %. Agreement was fairly high for the PK-PK classification (78 %) and for the PS-PS classification (78 %), but the agreement for CU- CU (conceptual understanding) classification was only 58 %. The authors argue that the disagreements indicate the shortcomings of the matrix design, which required each item to be associated with only one framework cell, because the two aspects of mathematical proficiency (CU and PK) are interrelated rather than exclusive.

The EP members also examined the 1992 NAEP grade-4 TSA items, with respect to the four major themes emphasized in the NCTM standards: problem solving, reasoning, communication, and connections. They found that approximately 60 % of the items were related to one of the standards themes, whereas 40 % of the items did not reflect any. The strongest relationship was found in the categories of reasoning (25 %) and problem solving (23 %), while 6 % were classified as communication and 5 % as connections. According to the authors, the new items for the 1992 assessment were more likely related to important themes contained in the NATM standards than items administered as part of the 1990 assessment (“trend” block items).

In addition, the EP members analyzed grade 4 special item types: extended open-ended items, manipulative items, and calculator items. Of the five extended open-ended items, one was judged to be exemplary, one to be seriously flawed, and the other three to have some problematic aspects either in the wording, the scoring guide, or both. The panel found that the wording used was somewhat ambiguous and confusing, which might have created a problematic situation for fourth-grade students with weak reading comprehension skills. They also found that some scoring guides did not discriminate well between adjacent levels.

Manipulative items required students to use manipulatives in the form of cardboard geometric shapes as part of the solution process. The EP members found that most of the six manipulative items contained in the same block assessed important mathematical knowledge and skills, including spatial visualization and the use of concrete models to represent abstract ideas and to facilitate problem solving. Yet, they questioned whether 15 minutes for the six items provided students with sufficient time to familiarize themselves with the geometric shapes and then produce adequate solutions.

Of 37 calculator items organized in three blocks, 11 were classified as calculator-inactive, 20 as calculator-neutral, and 5 as calculator-active. The panel found that most of the calculator-active items could be characterized simply by their use of messy numbers and that some of the calculator-neutral items were much more interesting mathematically than the calculator-active items. In addition, the panel members believed that the presence of the calculator might have had a somewhat negative effect on the cognitive processes used by students.

For grade 8, the extensive analysis for the relationship between the 1990 framework and items had already been released, and thus the authors provided only an update obtained through their independent examination of the framework and items. Regarding the intended versus actual distribution of items by *content*, the percentages were almost equal, with the exception of the category of AF (20 % intended versus 15 % actual). The distribution by *ability* category showed that fewer CU (40 % intended vs. 36 % actual) and PK items (30 % intended vs. 25 % actual) and more PS items (30 % intended vs. 40 % actual) were on the test than in the Mathematics Framework, which is similar to the pattern found for the grade 4 assessment.

In conclusion, the authors argued that the actual percentages of NAEP-classified items in the five content categories were approximately equal to those recommended in the Mathematics Framework for both grades, while there were some discrepancies in the ability categories with fewer PK items and more PS items. According to the authors, these findings indicated the existence of ambiguity in the definitions of NAEP ability categories, which might have decreased the validity of ability-oriented interpretations made based on student performance in NAEP. When the actual NAEP classifications on ability and content were compared to the classifications assigned by the expert panel, there was a fairly high level of agreement. They added that from the perspective of content validity, the 1992 NAEP TSA for grades 4 and 8 represented an improvement over the 1990 NAEP TSA and that the 1992 NAEP was more closely aligned with the NCTM Standards than was the 1990 NAEP (60 % in 1992 vs. 50 % in 1990).

The content validity of the 1992 reading assessment was examined by Pearson and DeStefano (1993). The authors report that the 1992 Reading Framework included many

innovations: (1) a definition of *a good reader* based on a constructivist view of learning; (2) use of authentic texts; (3) classification of three reading situations; (4) assessment of cognitive aspects of reading; and (4) an increased emphasis on open-ended questions. The framework took the form of a 4 X 3 matrix of reading situations by reading aspects. The reading situations include reading for literary experience, reading for information, and reading for performance of a task, while the reading aspects include initial understanding, developing interpretation, personal reflection and response, and demonstrating a critical stance.

Twelve reading experts, including teachers, administrators, and science educators, judged the degree of the alignment of the reading framework and test items. Within the grade-level groups each member categorized all of the items for a passage according to the three-by-four reading–aspect-by-situation matrix, and then panelists as grade-level groups compared and discussed their categorizations to reach a consensus. The authors compared the expert panel's categorization with the classification prescribed by the MAEP framework. Finally, they compared the expert panel and official NAEP classifications (by ETS) by reading situation and aspect on an item-by-item basis.

The authors found that the item classification across reading *situations* by the experts closely approximated the distribution targeted in the framework. The differences in percentages of items classified in each reading situation ranged from 3 % to 8 %, with the largest differences at grades 8 and 12. The experts agreed with the official NAEP classifications for 81 % of the grade-4 items. The agreement was the highest regarding the situation of reading to be informed (87%), while the panel agreed less on items related to reading for literary experience (74%). The experts agreed with the NAEP

categories on 89 % of the grade-8 items, while the greatest disagreement was found at grade 12 with an agreement of 73 %. In particular, the experts believed that 24 items officially classified for grade 12 by NAEP as information items should be considered literary-experience items. For all three grades combined, the experts assigned the same classification as NAEP for 276 of the 341 items, or 81 %. The largest discrepancy involved reading for information.

The framework assigns the response time across reading aspects; initial understanding/developing interpretation 33 %, personal reflection and response 33 %, and demonstrating a critical stance 33 %. The panelists' classifications were substantially skewed toward initial understanding/developing interpretation (67%, 86%, and 59% for grades 4, 8, and 12, respectively). The authors found that critical-stance items were strikingly underrepresented at grade 4 and 8 levels (9 % and 3 %, respectively), while time for personal-response items were more than 20 % below the specified level at grade 12.

For grade 4, the expert panel agreed with official NAEP classifications with an agreement of 67 %, with the largest agreement of 97 % for the aspect of developing an interpretation. However, the panel felt that 22 other items should have been classified as that aspect as well, especially 13 of 26 the critical-stance aspect categorized by NAEP. For grade 8, the agreement level was only 44 %, with the agreement of 9 % regarding critical stance: only 4 of 43 items classified as demonstrating critical stance by NAEP were so categorized by the expert panel. For grade 12, the level of agreement was also only 44 % overall and was considerably low (31 %) for items classified as critical stance.

Only 15 of 49 items classified as this aspect by NAEP were classified as the same aspect by the experts.

The data for all three grades combined showed an agreement of only 50 %. The largest disagreement concerned items classified as critical stance, with an agreement of 24 %. The authors report that the panel members found it difficult to assign a single situation-by-aspect category to an item, which reflects a common criticism of the matrix design raised by the Silver and Kenney's (1993a) study.

The authors conclude that the expert panel categorization of the items across reading situation on the 1992 NAEP in reading closely approximated the distribution targeted in the Reading Framework, whereas significantly less assessment time was devoted to personal reflection and response and to demonstrating a critical stance than specified in the framework, with respect to the aspect category. For consistency of categorization, the authors concluded that high levels of agreement between expert panel and ETS were obtained on the categorization of reading situation, while agreement was much less achieved for reading aspect, especially with personal response and critical stance items. The authors argue that the disagreement was attributed to ambiguity in the definitions of these two categories and to inconsistencies between these definitions and the criteria used to score items.

Silver and Kenney (1993b) explored relationships between the grade 8 cognitive items used in the 1990 NAEP TSA for mathematics and four selected themes from the NCTM Standards: Mathematics as Problem Solving, Mathematics as Reasoning, Mathematics as Communication, and Mathematics as Connections. The 1990 NAEP assessment was the first NAEP for which state-level results were reported.

An Expert Panel (EP) of eight mathematics education professionals examined 137 items contained in seven blocks that were used in the 1990 NAEP TSA for grade 8 (about 74 % multiple-choice and 26 % open-response). Each EP pair was given two of the seven blocks, and two EP pairs received the 1990 NAEP “trend” block, which was the set of 23 items administered in both 1986 and 1990 to link performance across assessments, as the basis for calculating interrater agreement. The EP pairs were instructed to reach a consensus for each item. Further, the study examined the distribution of items judged to match or not match selected NCTM Standards by NAEP ability and content categories.

The inter-pair agreement was made on 20 items (87 %) of 23 items: 3 item were classified as Reasoning and 17 items fit none of the NCTM Standards themes. Overall, the authors found that about half the items were related to the NCTM Standards themes, with the strongest relationship to Problem Solving (14 %) and Reasoning (28 %). Only one item was classified as Connections, and 9 % of the items were classified in the category Communication. The authors argue that strikingly almost half of the items (48 %) did not reflect any of the four selected Standards themes.

For the NAEP item classifications and the NCTM Standards themes, the EP members classified over 25 % of the items in the category of Reasoning, most of which involved the content area of geometry. Over 80 % of the items in the category of PS (problem solving) were judged to be related to the NCTM Standards themes of Problem Solving or Reasoning, while over half of the items in the CU (conceptual understanding) category were classified as Reasoning or Communication. More than 80 % of PK (procedural knowledge) items were assigned the designation None. Most items classified in one of the four NCTM Standards themes involved GY (geometry) or DA (data analysis,

statistics, and probability), while most of NO (numbers and operations) items were designated as None.

The authors found that a majority of the items judged to fit none of the Standards themes appeared in the trend block and calculator blocks and that almost all items in the open-response block were classified as at least one of the NCTM Standards themes, which indicates that open-ended items are more likely to be in alignment with the cross-cutting themes of the Standards than are multiple-choice items. The authors also pointed to the shortcomings of the matrix design, arguing that assessments based on the matrix design tend not to present the domain of mathematics as an integrated body of knowledge involving a rich network of interconnected ideas and processes.

The Psychometric Quality of NAEP

Validity is a multifaceted concept that depends on the particular uses and interpretations of assessment results as well as on the instruments and administration conditions (Messick, 1989; Linn et al., 1996). In an effort to examine the validity of NAEP results, this section reviews three research studies addressing a range of validity questions relevant to various uses and interpretations of NAEP.

The validity of any assessment of student achievement depends on the quality of the raw data provided by students. Accordingly, high rates of non-response can distort proficiency estimates, and especially non-response rates to constructed-response items might be too high overall or differentially high for minority students (Koretz, Lewis, Skewes-Cox, & Burstein, 1993). Koretz et al. (1993) examined patterns of non-response

in all three age/grade groups (age 9/grade 4, age 13/gade 8, and age 17/grade 12) in the 1990 assessment of mathematics, which involved increasing numbers of open-format items. In their study, an item was considered to have a high omit rate if more than 10 % of the students in a grade-only sample omitted it, while a 15 % cutoff was used to define items as having high not-reached rates. They examined non-response rates for each item in the seven scaled blocks (one with constructed-response items only, one with multiple choice only, and five with both formats) and identified differences in non-response for population groups and gender. They also examined the relationship between non-response and several aspects of items (e.g., format, difficulty, content areas).

The authors found that overall not-reached rates were substantially less in the 1990 mathematics assessment (8%) than in the 1986 assessment (23%). They found that omit rates were strongly related to item format and that omit rates were modest at grades 4 and 8, but mostly high at grade 12 particularly on the constructed-response items. In contrast, according to the authors, not-reached rates showed a weak relationship to item format and this might be due to the structure of blocks since most of not-reached items at the end of each block were multiple-choice.

They found that in general, the constructed-response items with high omit rates were mostly more difficult than other constructed-response items and the average multiple-choice items and that items with high not-reached rates tended to be more difficult than others (*p*-value for those items is .27, compared to .56 for other items). Thus, they asserted that high omit rates seem to be a function of difficulty as well as format. In addition, the authors revealed that neither the content nor the process classification of items was strongly related to omit or not-reached rates.

The authors revealed that no items showed large gender differences in not-reached rates, but that a few showed sizable differences in omit rates. Female students had higher omit rates on the five Grade-12 items, but lower omit rates than males on the two Grade-4 items (five of the seven items were constructed-response). To examine whether these few gender differences were independent of mathematics proficiency, the authors conducted a Mantel-Haenszel (MH) analysis in which the outcome was omit rates rather than the conventional *p*-values. The MH analyses suggested that gender differences in omit rates could not be fully explained by differences in proficiency.

Regarding non-response for population groups, they found that Hispanic and African American students had higher omit and not-reached rates than whites in all three grades, that differences in not-reached rates were somewhat larger than those in omit rates, and that most of the items showing the differences were open-ended. They also found that the population-group differences in omit rate did not show striking or consistent relationships with content area. Mean differences were relatively large in data analysis, probability, and statistics in all grades, but the disparity between this content area and the others was modest. The authors conducted a Mantel-Haenszel (MH) analysis similar to that done with gender differences to determine the influence of ability differences. According to them, the results suggested that the black-white differences in omit rates can be accounted for by differences in proficiency, but not all.

In conclusion, the authors contended that overall, omit rates were modest in grades 4 and 8 but mostly high at grade 12, and that not-reached rates were greatly reduced from 1986 levels. They also argued that the higher omit rates on constructed-response items for African American and Hispanic than for White grade 12 students raises concerns,

particularly in light of the increasing reliance on open-ended items by NAEP. Yet, they suggested that differences in non-response for White and Black/Hispanic students can be partially explained by differences in mathematics proficiency. Finally, the authors recommended that omit and not-reached rates should be routinely monitored because substantial non-response has the potential for introducing error into the scaling of the test.

The achievement levels set by NAGB have been controversial since the first effort which focused on the 1990 mathematics assessment (Burstein et al., 1995/1996; Koretz & Deibert, 1995/1996; Linn & Dunbar, 1994). A study by Burstein et al. (1995/1996) examined the degree to which the descriptions of the achievement levels for the 1992 assessment in mathematics accurately represented what students at each level were able to do. Three approaches were used: (a) review of exemplar items; (b) classification of items based on descriptors; and (c) characterization of items that statistically differentiate student performance.

First, the authors reviewed the performance of students at each level on the exemplar items NAGB used to illustrate the levels. According to the authors, exemplar items were selected by NAGB on the basis of two criteria: (1) at least 50.1 % of students just exceeding the cut score for a given achievement level (borderline students) are expected to answer the item correctly; and (2) the item's content is matched with the description of the levels. Ten exemplar items were chosen for grade 4, eight items for grade 8, and eleven for grade 12. The premise was that a reasonably high proportion of students at a given level should be able to answer correctly the exemplars for that level and that exemplars of higher levels should be more difficult than exemplars of lower levels. Accordingly, the authors considered two thresholds for what constitutes a reasonable rate

of success on the exemplar items: a lenient criterion that $p > .50$, and a more stringent criterion that $p > .65$ (p value refers to the percentage of students at each level who correctly answered each item).

Second, they mapped each item in the entire item set onto NAGB's descriptions and explored the performance of students at each level on items that mapped to that level's descriptions. This was done by breaking each description into brief descriptors, each of which referred to a distinct attribute of items with being kept as close as possible to the NAGB description. A group of six mathematics educators was asked to map each item to all appropriate descriptors (the mapping was dichotomous).

Lastly, they determined empirically which items differentiated among the achievement levels and then explored their characteristics. This analysis reversed the logic of the previous one. The researchers defined differentiating items as those that met three conditions: (1) The p value for students at the target level must be at least .65; (2) The p value for students from the next lower level must be at least .30 less than the p value for students at the target level; and (3) at least 50 % of the students at the next lower level must get the item wrong.

The authors found that the exemplars selected for the 1992 mathematics assessment were poorly suited to illustrate the actual performance of students at each of the achievement levels. They report that using even the lenient criterion, two of ten grade-4 exemplars (p=.49 for Basic, p=.48 for Proficient) and two of eight 8th-grade items (p=.36 for Proficient, p= .42 for Advanced) failed to qualify as exemplars. They found that using a more stringent criterion of .65 disqualified additional seven exemplar items across the three grade levels, that is, approximately two thirds of the exemplar items

across the three grades failed to meet the criterion that 65 % or more of students at a given level answered the item correctly. Moreover, the authors argue that items exemplifying a given level were not consistently easier than items at higher levels.

The authors maintain that most items were successfully mapped to at least one descriptor by the judges but that NAEP's coverage of the attributes in NAGB's descriptions was very uneven. For example, in grade 4, the number of items mapped to each descriptor ranged from 1 to 87, 1 to 108 in grade 8, and 0 to 53 in grade 12. For grade 4 seven of the 18 descriptors were mapped to fewer than 9 items, 16 of 31 descriptors for grade 8 and 24 of 35 descriptors for grade 12. According to the authors, the number of items assigned to each achievement level also varied markedly and NAGB's descriptions for the different levels included phrases that were too similar to differentiate. For example, in grade 4 where the problem was most severe, 131 items were mapped to Proficient-level descriptors, but only 6 items were mapped to Basic-level descriptors and 13 items to Advanced.

They also found the NAGB descriptions were sometimes inconsistent with actual patterns of student performance. Using a median *p* value ≥ .65 as a criterion (for all descriptors mapped to at least 9 items by at least four out of six judges), fewer than half descriptors exhibited a pattern of student performance consistent with the achievement-level statements from which the descriptors were derived. For example, for 8th grade where the match was worse than in grade 4, five of the six Basic descriptors had median *p* values below .55 for Basic students, only four of the seven Proficient descriptors were appropriate for that level, and two of three Advanced descriptors seemed appropriate.

Grade 12 was found to have the worst match: only 4 of 13 descriptors matched their levels reasonably.

The authors report that in the third sub-study, as a first step, the achievement levels at which items differentiated were cross-tabulated with the levels at which they were mapped to descriptors. They found that the match was reasonably good for the Proficient level at grades 4 and 8, fair for the Basic level at grade 12, and poor in all other cases. Foe example, at grade 4, none of 28 items that differentiated at the Basic or Advanced levels were mapped to descriptors at the corresponding levels, and most were mapped to Proficient descriptors.

In conclusion, the authors argued that there existed serious inconsistencies between actual performance and both the descriptions and exemplars provided by the NAGB. They summarized: 1) some of the exemplar items were misleading; 2) the achievement-level descriptions overlapped considerably; 3) NAEP measured some of the attributes in the descriptions poorly and others not at all; 4) student performance on items corresponding to NAGB's descriptions was often unreasonably low; and 5) items that actually differentiated among achievement levels in terms of actual student performance did not correspond well to the achievement-level descriptions. They suggested that exemplars and descriptions of the achievement levels should be aligned with actual student performance on NAEP.

The performance of students at risk of low achievement has been of special interest for NAEP (Linn et al., 1996). Berends & Koretz (1995/1996) investigated the validity of the NAEP tests in terms of social context measures for diverse ethnic groups. NAEP reports have typically presented unadjusted differences among population groups without

attempting to adjust them for dissimilarities in social context (Berends & Koretz, 1995/1996; Koretz, 1992; Raudenbush, Fotiu, & Cheong, 1998). However, reporting only unadjusted differences among population groups may be misleading since student test score differences that statistically adjust for dissimilarities in social context are far smaller than the unadjusted differences (Berends & Koretz, 1995/1996; Raudenbush et al., 1998). In this study, the authors examined the adequacy of NAEP's measurement of social context for the specific purpose of reporting adjusted score differences among population groups from similar social contexts.

The authors focused on low-achieving students identified as those who score in the bottom quartile and decile of the achievement distribution and examined the strengths and weaknesses of NAEP for describing those at-risk students. The main body of analysis explored the practical impact of NAEP's choice of social-context constructs. This study was conducted by comparing the 1990 NAEP (grades 8 and 12) for reading and mathematics with the 1988 National Education Longitudinal Study (NELS) for grade 8 and 1980 High School and Beyond (HSB) for grade 12, both of which contain richer social context measures than does NAEP.

HSB and NELS were treated as benchmarks to explore the extent to which NAEP could replicate results obtained with those databases. The research questions were: (1) for the purpose of reporting adjusted score differences among population groups, how adequate are NAEP's social context measures?; (2) if there are problems involving the selection of social context constructs, how much practical impact do they have?; and (3) is there any practical impact of relying on student-reported information when adjusting the score differences?

To compare test scores in NAEP, HSB, and NLES, the authors used the percentage of change in the standardized group differences as the metric for evaluating the effects of controlling for social context. They used several sets of social-context variables in NAEP, HSB, and NLES, including measures of family background, language use, community and school characteristics, and curricular differentiation. Family background variables included parents' education, parents' occupation, and family income, and family composition included two-parent household and number of siblings. For language use, they created variables such as other than English spoken in home, other than English spoken generally, and frequency of speaking language other than English. The community characteristics variable was categorized into region of the country, locale, and size and type of community (STOC). The authors also analyzed information the NELS and HSB gathered from students' parents to examine consistency between student and parent responses on social context measures: 1) comparing student and parental reports of the same variables using NELS and HSB data; and 2) conducting parallel regression analyses.

To provide a baseline to compare the adjusted estimates, the unadjusted mean score differences between population groups were calculated. The authors estimated a series of models that cumulatively added an increasingly wide array of social context variables to estimate the changes in the mean score differences between Whites and both African Americans and Hispanics when controlling for a variety of social context measures. They explored the adequacy of the NAEP variable sets by comparing the regression analyses of all three databases, by estimating the reduction in the unadjusted difference at each stage.

First, they presented comparisons of regression analyses in NAEP, HSB, and NELS after controlling for a wide variety of social context variables. For grade 8, they found that *Black-White* unadjusted score differences in mathematics were .93 SD in NAEP and .77 SD in NELS. Adjusting these differences with the available measures of family background reduced them substantially. The NAEP family background model resulted in a 16 % reduction of the unadjusted difference, while the NELS model reduced it by 32 %. According the authors, NELS's information on parent's occupations and family income accounts for the greater reduction in NELS. With additional controls for school and community characteristics and curricular differentiation, the cumulative impact of controlling for all of these variables in NELS was a 56 % reduction (from .77 SD to .34 SD), while the corresponding reduction in NAEP was 44 % (from .93 to .52 SD).

They found that *Black-White* unadjusted score differences in grade 8 reading were smaller (.67 SD in NELS and .44 SD in NAEP). Yet, adjusting only for differences in family background produced a marked reduction in the gap, which was proportionately much larger in NELS than in NAEP. The cumulative reduction in NELS was 54 % (from .67 to .31 SD), whereas in NAEP it was 36 % (from .44 to .28 SD).

For grade 12, the *Black-White* unadjusted test score difference in mathematics was .83 SD in HSB and .86 SD in NAEP, while in reading .78 SD in HSB and .66 SD in NAEP. In mathematics, controlling only for the family background variable reduced the gap much more in HSB than in NAEP (32 % and 16 %, respectively). Adding controls for family composition and language use had small effects in both NAEP and HSB, whereas adding community and school characteristics had more of an effect in NAEP than in HSB. The total reduction, while accounting for all variables, was 43 % in both

NAEP and HSB. In reading, a similar pattern appeared. The cumulative reduction in the gap was 36 % in NAEP and 40 % in HSB.

For grade 8, they found that *Hispanic-White* unadjusted score differences in mathematics were .70 SD in NAEP and .60 SD in NELS, while in reading .53 SD in NAEP and .56 SD in NELS. After controlling for all variables, the gap in mathematics was .16 SD in NELS and .38 SD in NAEP, while in reading .12 SD in NELS and .31 SD in NAEP. According to the authors, these results indicated that several important social context measures are absent in NAEP and that in particular, NAEP's inadequate language variables hindered its ability to portray Hispanic-White differences independent of social context.

The grade 12 *Hispanic-White* contrast was the only case in which controlling for social context reduced the unadjusted test score differences more in NAEP than in HSB. The authors argued that it might be due in part to the dramatic growth and changed composition of the Hispanic population between 1980 (HSB) and 1990 (NAEP). The greatest reduction in the NAEP mathematics difference was from .60 to .20 SD in the school and community model and in HSB from .65 to .35 SD. The pattern in reading was similar, that is, the greatest reduction was also from .51 to .22 SD in NAEP in the school and community model and from .73 to .46 SD in HASB.

Second, the authors presented the degree of consistency between students and parents on family characteristics. Overall, they found that students and parents were generally consistent about relatively obvious characteristics such as population group membership, family composition, and language use, while less consistent regarding family background characteristics such as parents' education levels and occupations and family income. For

example, the general agreement about population group membership in NELS and HSB was over 90 %, but the consistency about family background measures was moderate to low and differed dramatically by grade level. For example, the authors found that the agreement was 54 % in grade 8 and 72 % in grade 12 regarding mother's education level and that the consistency about family income was worse, 31 % in grade 12. Both grade 8 and grade 12 students were also inconsistent with their parents in terms of parents' occupation: 45 % in grade 8 and 52 % in grade 12.

Lastly, the authors reported the impact of NAEP's reliance on student reports of family characteristics. They compared identical regressions based on information from parents and students, using NELS and HSB (NAEP lacks parent reports) and found that there were virtually no differences between regressions in both grades 8 and 12. For example, the unadjusted mathematics score differences between grade 8 African Americans and Whites were .77 SD in the student data and .82 SD in the parent data in NELS. While adjusting for social context measures, they found that the percentage reduction differed by no more than two percentage points between the parent and student data (and the regression lines were parallel).

In summary, the authors concluded that NAEP lacks a number of social context measures and that the quality of some of NAEP's measures is low because of its reliance on student self-reports. They suggested that additional measures including parental occupation, number of siblings, family income, and ability grouping in class, might improve adjusted estimates for achievement differences among population groups in NAEP. In addition, they indicated that NAEP's overestimates for performance difference

among student groups at the secondary school level are due primarily to its lack of some important social-context measures, rather than its reliance on student reports.

Utility of NAEP Data

Limited studies have been conducted to investigate how NAEP data are actually used at federal, state, or local levels. This section reviews four studies focusing on the use of NAEP data at the federal, state, or local levels and three research projects aimed at linking state test results to NAEP. In addition, a study examining policymakers' views of student assessment is reviewed since it has some implications for appropriate use of NAEP.

Sebring and Boruch (1983) conducted case studies of seven state education agencies and school districts within these states regarding use of NAEP through interviews and documents. Case studies were also developed for the National Council of Teachers of Mathematics (NCTM) and the National Council of Teachers of English (NCTE). For legislative uses, legislative hearings were reviewed and interviews were conducted at federal and state levels of government. In addition, they reviewed NAEP records covering the past 10 years.

The authors revealed three categories of use based on the purposes for which NAEP information was used: professional, policy, and research. According to the authors, *professional use* refers to employing NAEP data, methods, and materials to enhance programs and instruction. *Policy use* refers to utilizing NAEP data to inform decision-making in Congress, state legislatures, or state and federal agencies, while *research use*

refers to NAEP data use to advance measurement techniques or to examine relationships between achievement and student and school variables.

For professional use, the authors found that five states replicated the NAEP model to conduct parallel assessments at the state level and that two states adapted the NAEP model for their state tests. They found that in most states studied, NAEP reports were distributed to local schools and/or workshops were conducted to share findings and recommendations. For example, in Texas NAEP results were presented at public hearings on the development of the writing objectives for the Texas Assessment of Basic Skills (TABS), and in Minnesota the 8th-grade NAEP science results led the State Board of Education to commission a task force to study the middle school science curriculum. They also found that local schools used NAEP materials at times to develop curricula, write competency tests, inform teachers of national trends, and analyze curriculum weaknesses and strengths, but that the frequency of use was unknown. The NCTM and the NCTE were found to assist in interpreting and disseminating NAEP results. For example, NAEP findings served as a basis for NCTM's recommendations for mathematics curricula.

For policy uses, the authors found that the Congressional Research Service made extensive use of NAEP in response to Congressional committee requests for data on education. According to the authors, NAEP's own report of utilization showed many instances when information was provided to the House Subcommittee on Elementary, Secondary, and Vocational Education and the Senate Education Subcommittee also used NAEP data, but less frequently. For instance, NAEP data were presented during hearings on the consolidation of the elementary and secondary programs and during the hearings

on the Omnibus Budget Reconciliation Act that played an influential role in Title I being cut 12 percent rather than 25 percent. They found that federal agencies also used NAEP for policy purposes. For example, the Department of Labor conducted a study of out-of-school 17-year-olds using NAEP items. At the state level state legislatures in Maine and Texas depended on NAEP-inspired assessments when considering legislation mandating minimum competency testing, and the experience with NAEP was incorporated in the subsequent statute creating the TABS.

For research uses, the authors report that transfer of NAEP to the National Institute of Education (NIE) emphasized research use of raw data. As a result, they found that 170 public use data tapes had been distributed to researchers since 1979 for data analysis, which was a huge increase in research use compared to before. Yet, they found that NAEP data use to make decisions is much less common than use of data to persuade or confirm one's own beliefs, to understand issues, and to recognize problems.

In conclusion, the authors argued that NAEP products and services were used by diverse audiences and for multiple purposes. In particular, they suggested that NAEP data are most useful at the national level, including Congress, federal agencies, professional organizations, and researchers. The most use of NAEP at the national level makes sense since State NAEP was not developed yet at the time their study was conducted.

In addition, they argued that examination of the use of NAEP was extremely difficult because of the transitory nature of use and because reports of use were subject to bias. They pointed out that a major difficulty in investigating use was the ambiguity of the term itself, suggesting that the term be defined by NAEP in terms of audiences (local and state education agencies, federal agencies, and organizations), types of use (professional,

policy, and research), functional nature of use (decision-making, persuasion, enhancement of understanding, etc.), and the elements of NAEP used (NAEP items, frameworks, NAEP results, sample methods, etc.).

Bullock and DeStefano (1998) investigated the perceptions of state assessment directors regarding the usefulness and credibility of various components of the 1992 Trial State Assessment (TSA) in reading. For the first time in the history of NAEP in reading, results were reported at the state level, and thus the authors examined state-level administrators' perspectives on credibility and usefulness of the Reading Framework, the achievement-level descriptors, reporting, dissemination of NAEP results, and provisions of technical assistance. In addition, they explored how results from the TSA in reading were used by state-level administrators of assessment. Of 44 state directors of assessment whose states participated in the 1992 TSA in reading, 26 participated in the study through interviews.

The authors found that the presence of only one teacher on the NAEP Planning Committee troubled many of the state assessment directors, indicating that diverse stakeholders need to be involved in planning and development. They also found that generally participants liked the achievement-level descriptors provided with the results to aid in interpretation, considered them easy to understand, and felt they made NAEP results more useful. However, they revealed that the category "Below Basic" was perceived as vague since no Below Basic descriptor was provided in NAEP, although a large percentage of students were in that category in many states.

According to the authors, consistent with the NAEP's purpose, TSA results were being used primarily to inform policymakers at state and district levels and to provide

information to teachers, administrators, and the public, in terms of how the states were performing comparatively. They report that participants (62%) recommended simplifying the results to make them more useful, that is, they suggested that reports: 1) provide standard errors in a different manner on the charts since the current manner was confusing; 2) be sent to them in computerized form to be easily manipulated into their own customized reports; and 3) provide a quick index. In addition, some (46 %) suggested adding an SES indicator to results to help judge the validity of state comparisons and more graphical depiction in the results (42 %).

NCLB implies that NAEP might be used to confirm state test results in grades 4 and 8 reading and mathematics under the purview of the U.S. Department of Education as an element of the system of state-level accountability for student achievement results (NAGB, 2002b). To help plan for the use of NAEP in this new role, NAGB commissioned exploratory studies through an Ad Hoc Committee to investigate NAEP's usefulness in a confirming role under NCLB. The committee, assisted by a Planning Work Group (PWG) consisting of technical and policy experts, examined state test results in eight states, prepared arguments about performance in three of those states, and used relevant NAEP data to study NAEP's capacity to serve as a source of confirmatory evidence for state test results. In addition, the committee explored new ways of representing achievement gains and achievement gaps using achievement distribution charts and achievement distribution gap charts.

Holland, a member of the PWG, suggested displays of achievement data that help convey the size of changes in scores over time and changes in gaps between two subgroups over time. He introduced the cumulative distribution function (CDF) curve of

the scores for a given group of test takers. For each possible score, the CDF is the percent of students in the group with a test score less than or equal to that value. CDF curves, roughly S-shaped, start at 0 on the left and rises up to 100 on the right and display percentiles in a backward way. According to the author, shifts to the right indicate higher score distributions while shifts to left indicate lower score distributions, showing changes in location to the right are associated with achievement improvement. Likewise, the gap between two subgroups is the space between the two curves that displays gaps between the two distributions at all levels of achievement.

Holland reports that we can measure the gaps by measuring either the vertical distance between two CDFs or the horizontal distance between them. The former is used to compare percents above achievement level score cuts in NAEP (gap in percents), while the latter used to display the differences in percentiles against each possible percent on the horizontal axis (gap in percentiles). The former shows the gap by comparing percents at or above the three achievement level cuts for Basic, Proficient, and Advanced, while the latter shows gaps over the whole range of possible choices of the percent, p. According to the author, percentile differences are more robust to small changes in the score distributions than differences in percents, and these displays provide a context for interpreting a single number used to report a gap in scores.

Three state arguments were prepared regarding use of NAEP as confirmation by a three-member subgroup of the PWG. Yen conducted a thought experiment with data from state A, imaging 1993 is the base year for the implementation of the new ESEA legislation. State A data came from state test data in reading and mathematics since 1993 and NAEP data from 1992, 1994, and 1998 (grade 4 reading) and 1992, 1996, and 2000

(4th- and 8th-grade mathematics). The NAEP Basic level was chosen as the primary focus of attention since it was most similar to the State A proficiency level. This study reviewed trends in the percents of students reaching the Basic level and then compared these trends with the percents of students reaching the Proficient level for the state test. Because the state data are compared to targets under NCLB, the Primary Target was calculated for the NAEP data available (primary target is the percent of students needing to reach the proficient level so that all students are projected to meet the standard in 12 years, given linear growth from the base year).

According to the author, NAEP results were found to confirm results for the state A test for the state as a whole and for the major ethnic groups, in terms of general trends in improvements in performance over time, not AYP goals. In terms of overall state performance relative to AYP targets, NAEP results are farther away from the NAEP targets for both grades and both reading and mathematics than the state results are from the state targets. According to the author, this discrepancy results from the fact that the NAEP performance in the base years is at a lower percent reaching the standard than the state test, thus leading the target growth rate for NAEP to be greater than that for the state test. In terms of meeting targets for subgroups, there was substantial inconsistency between the state test and NAEP results.

Reckase selected State B to determine if State B results and the NAEP results are consistent with each other. According to the author, State B makes heavy use of performance assessments in its testing program, the Criterion Referenced Examination, reports results according to standards, uses a rolling average for reporting results, and uses a variety of subscores rather than a single total score. The state assessment results

are reported as a three-year rolling average percent above the standard (the Achieved the Standard and the Achieved the Standard with Honors), which seems to be the closest to the both the Basic and Proficient categories for NAEP.

The author found somewhat inconsistency between the two tests for grade 4 reading. For example, there was no change between 1994 and 1998 for either NAEP Proficient or Basic levels, while State B results showed slight growth after 1998. For reading gaps, the size in gaps between subgroups was roughly comparable for grade 4 and grade 8 Blacks and Hispanics, but the pattern of results for Asian American students seemed different. In mathematics, he found that the trends in performance was similar for grade 4, but not for grade 8. For example, state test results showed growth in grade 4 and mixed results for grade 8, while NAEP results showed growth over time for both grades. In terms of mathematics gaps, NAEP data showed increases in gaps from 1996 to 2000 in almost every case, while gaps in state tests generally did not change or got smaller.

The author concluded that the NAEP results were inconsistent with the state test results, but not in a regular pattern. Thus, he suggests that NAEP could be used to confirm state test results only in a very general way and that trends in the same direction or lack of directly contradictory information might be considered as confirmation.

Holland focused on Black students using State C data from 1993-94 to 1999-2000 and used 1994 as the base year to emulate the NCLB use of 2001-2002 as a base year. The author devised an AYP system of his own because of lack of a statement about what AYP meant from the perspective of State C. The author reports that it takes an average yearly increase of from 3.5 to 5.7 percentage points depending on the grade and subject to meet the 12-year target of 100 % of Black students meeting the proficient standard.

In reading, there was no change in grade 4 regarding the percent above NAEP Basic from 1994 to 1998, while there was improvement in state tests over the period of 1994 to 2000. In contrary, NAEP supported the conclusion of improvement in the performance in mathematics in state tests. In terms of gap closing, in grade 4 reading the trend in percents at or above NAEP Basic was contrary to a clear decrease over the 7-year period in the gap between White and Black students on the state assessment, whereas mathematics trends were in the same direction for both NAEP and state test results for both grades, showing a decrease. In addition, using the achievement gap distribution, the author found NAEP reading showed an increase in scores for both Black and White students, but only for the lower scoring part of the distribution for Black students, thus leading Black students above the NAEP Basic level to be unchanged.

Finally, based on a review of test results from eight states and three demonstration arguments designed to illustrate how states might report test results to meet the requirement of the NCLB Act, the committee presented principles for the use of NAEP as confirmatory evidence. First, NAEP can be used as evidence to confirm the general trend of state assessment results in grades 4 and 8 reading and mathematics. Second, confirmation should not be conducted on a point-by-point basis or interpreted as a strict validation of state test results. Third, limitations in using NAEP to confirm the general trend of state test results should be acknowledged explicitly. Fourth, test frameworks for NAEP should continue to be developed through the active participation of states; content coverage must be broad and inclusive; the focus should continue to be on measuring what students know and can do; and changes in NAEP frameworks should be made infrequently to maintain the goal of stability in measurement. Fifth, sampling procedures

for NAEP should ensure that results for major subgroups within a state are reliable and that the size of standard errors is minimized. Sixth, the release of NAEP data should include comprehensive means for presenting and analyzing state NAEP results. Lastly, the percent at or above basic, proficient, and advanced, and the percent below basic should be presented and considered in light of the full range of state standards.

Wilson & Blank (1999) approached NAEP and TIMSS results from the perspective of mathematics educators and education decision-makers at state and local levels to guide their efforts toward improvement of teaching and curriculum. The authors focused on analyzing content areas of strength and weakness in mathematics covered by both assessments. They also examined school and classroom factors that might be related to higher achievement, using NAEP data.

The authors argued that it was possible to analyze results from NAEP and TIMSS by content area since both assessment frameworks were similar. The major difference is that TIMSS framework was developed by an international panel of scholars and educators in mathematics and science. The NAEP framework for mathematics has five strands and two domains that cross the strands, while the TIMSS framework has eight content categories and five performance expectations categories. The authors examined some of content areas covered by both assessments.

According to the authors, long-term trend NAEP results indicate that the overall trend is of increased performance over time at all age levels, steadily but gradually, in terms of both scale scores and achievement levels (from 1973 through 1996). Yet, TIMSS results from the 1997 assessment show that, after fourth grade, US students lag far behind their counterparts around the world, which is a concern for mathematics educators. That is,

U.S. fourth-grade results were above the international average, but eight-grade results below the average and 12-graders almost at the bottom. Also, the change in cohort growth (between the fourth and eighth grade) between 1978-1982 and 1992-1996 showed less score gain in mathematics from grade 4 to grade 8 (Barton & Coley, 1998). The authors contended that these findings highlighted that the overall scores on NAEP continued to be low despite the continuous improvement over time.

Regarding performance on NAEP and TIMSS by content area, the authors found that in algebra, students at grades 4 and 8 improved their performance on NAEP from 1992 to 1996, but that on TIMSS, 4th-graders were above the international average, 8th-graders at the average, and 12th-graders below the international average. They report that algebra results from NAEP were higher than other content areas at all three grades, but that the overall knowledge level was relatively low. For example, only 4% of eight graders and 2 % of twelfth graders demonstrated the ability to identify and generalize complex patterns and solve real-world problems. In geometry, they found that fourth graders were strong but that eight and twelfth graders were weak. For example, on TIMSS US students scored above the international average in 4th grade, below the average in 8th-grade, and almost last at grade 12.

According to the authors, NAEP results show that U.S. students scored well on simple whole number computation and operations, but were weak in applying their knowledge to unfamiliar situations in number sense and estimation. On TIMSS, students at grade 4 were below the international average in number sense and 8th-graders at the average. Measurement was a particular weakness in U.S. mathematics at all grade levels. For example, U.S. students at both grades 4 and 8 scored below the international average,

and on NAEP students' scores in measurement was lower than the overall average at grades 8 and 12. The authors found that students had difficulties particularly with questions requiring unit conversions, calculations of volume and circumference, and estimation of measurement.

For item types on NAEP, the authors found that students scored significantly higher on NAEP multiple-choice items at all grade levels than open-ended items. The extended constructed response items were designed to assess higher levels of problem solving, reasoning, and mathematical communication. According to the authors, these findings indicate that overall, students in mathematics classes need more opportunities to work on non-routine problems, to use higher-order thinking skills, and to communicate their mathematical ideas.

In addition, to identify factors that might affect achievement, the authors analyzed differences in NAEP mathematics results by state context, variation on performance between students performing at high and low levels, type of community, and curriculum. They found that state context is related to student achievement in mathematics. For example, high-performing states have higher expenditures per pupil, fewer children living in poverty, and more adults with high school diplomas. Additionally, the authors found significant variation in performance between high-performing and low-performing students and that variation increased from grade 4 to grade 8. They also revealed variation in performance by school location and curriculum. For example, the authors found that students taking algebra performed better than those taking regular mathematics in the 1996 NAEP assessment.

The authors analyzed student performance in NAEP and TIMSS in terms of three variables in opportunity to learn mathematics: teacher preparation, implemented mathematics curriculum, and teaching practices. They found that low-performing states on NAEP had lower percentages of mathematics teachers with a major or minor in their field and that the amount of professional development received by teachers did not have a consistent relationship to student achievement, indicating mathematics educators need to focus on the quality of professional development. They also found that mathematics curriculum in the U.S. was very broad compared to other countries, covering too many topics at each grade with too little depth and that traditional practices were still widespread. For example, in most lessons, teachers showed students how to solve problems and that students did individual seat-work at grades 4 and 8.

The authors conclude that a more meaningful picture of mathematics education in the U.S. can be obtained by studying the meaning behind average scores and summary statistics reported about NAEP and TIMSS. They suggest that it is essential to make some changes to what mathematics is taught, how it is taught, and how teachers are supported in order to provide all students with a quality mathematics education.

The findings of this study have some implications for NAEP data use by states. For example, state education personnel might examine content areas within a subject that students show weaknesses in specific areas in need of attention and instructional improvements, by analyzing released questions and student responses. Further, this kind of information could be taken into consideration in planning K-12 programs and policies aimed to promote student performance at state levels.

McDonnell (1994) notes that there have been the sharpest disagreements between testing experts and policy makers on the policy uses of assessment. She notes that the use of test results as high-stakes purposes has been of particular concern to testing experts, despite the trend change from a sole reliance on multiple-choice tests toward performance-based assessments focusing on critical thinking skills, the application of knowledge, and the integration of knowledge across subjects.

According to the author, testing experts claim that the same test cannot be used for different purposes. For example, high-stakes tests used for accountability purposes cannot be validly used to provide information on the status of the educational system. However, policy makers see new forms of assessments as an instrument for both accountability and other purposes. Thus, the author analyzed the gap between policymakers' enthusiasm for the policy uses of student assessment and expert caution about its potential misuses. The study focused on policymakers' differing expectations of what assessment policy can accomplish and how they view the feasibility of assessment-based reforms, based on interviews with 34 national and state policymakers including White House staff, congressional staff, state legislators, governors' education aides, and interest group representatives.

Firstly, the author found seven types of purposes policymakers expect assessments to serve: 1) describing the status of the education system; 2) aiding in instructional decisions on individual students; 3) bringing great curricular coherence to the system; 4) motivating students to perform better and parents to demand higher performance; 5) acting as a lever to change instructional content and strategies; 6) holding schools and educators

accountable; and 7) certifying individual students as having attained specified levels of mastery.

Yet, the author argues that opinions differ about which of the seven purposes is the most important or what combination of those is most appropriate. For example, state policymakers viewed accountability as a major purpose of assessment, while national-level policymakers implied a weaker form of accountability (disseminating information about the status of the education system). In between these, according to the author, two views are those who believe that assessments can motivate students to perform better, bring greater curricular coherence to the nation's education system, and act as a lever to change instructional content and strategies.

Secondly, the author found that policymakers and testing experts viewed the feasibility of assessment-based reforms differently as well. She argued that testing experts had the notion that technical questions about generalizability of performance-based assessments are sufficiently problematic to warrant caution in moving to widespread implementation of new strategies. In contrast, the author found that policymakers did not see these problems as a barrier to broad-based implementation of new forms of assessments, although they did not consider technical constraints unimportant.

There are also other constraints that set limits on the implementation of assessment as a policy instrument, such as test costs and political constraints. According to the author, most policymakers understood that alternative assessments cost more than multiple-choice tests, but viewed them as the least expensive strategies for reforming schools. In

addition, the author found that there was generally broad support for the concept of national standards and new forms of assessments among policymakers.

In conclusion, the author maintains that policy uses of assessment remain unresolved between policymakers and testing experts. She argues that policymakers need to develop more realistic expectations about what assessments can accomplish, acknowledging even the best assessments are imprecise measurement tools with real limits on their generalizability and appropriate use. The author adds that testing experts "should be explicit about how much of their criticism stems from principled opposition to high-stakes tests and how much is specific to particular types of assessments and their policy uses" (p. 43) and then need to outline the conditions under which the problems they have identified with new forms of assessment can be mitigated.

A growing number of states have been interested in linking statewide standardized test results to NAEP. The word "linking" is a generic term that includes a variety of approaches to make results of one assessment comparable to those of another (Linn, 1993). Although prior studies indicate that the linking of assessments with differences in content, item format, and motivational levels is challenging, the benefits might outweigh the concerns (Mullis, 2003).

Both Mislevy (1992) and Linn (1993) describe three levels of linkage that are potentially useful for expressing the results of statewide tests on the NAEP scale: equating, calibration, and prediction. *Equating*, the most demanding type, can be done only when two distinct assessments measure the same construct with equal degree of reliability (forms are interchangeable). *Calibration*, usually based on the models of item response theory (IRT), is used to provide comparable scores on tests that measure the

same thing (Linn, 1993; Mislevy, 1992), possibly with different degrees of precision (e.g., short and long forms of a test). *Projection* or prediction uses an empirical relation between scores on tests that do not measure exactly the same thing to predict the distribution of one test from that of scores on another test.

Linn and Kiplinger (1995) conducted a linkage study to provide a better understanding of the degree to which existing statewide assessments may be linked to NAEP despite violations of basic underlying assumptions that the assessments measure the same thing with equal precision, by using equipercentile equating procedures. The equipercentile equating defines observed scores at matching percentiles to be equivalent (Ercikan, 1997).

From two statewide assessments in Grade 8 mathematics and NAEP TSA data obtained for both 1990 and 1992, they obtained an equating function that converts the statewide results in 1990 to the 1990 NAEP-TSA results and then use the data collected in 1992 to evaluate the accuracy of that conversion when used 2 years later. Data from statewide tests in Grade 8 mathematics in 1990 and 1992 were obtained from four states that participated in both the 1990 and 1992 TSAs. Two states used the SAT (Stanford Achievement Test; forms E, K, L), one state used the Iowa Tests of Basic Skills (ITBS), and one used the CAT (California Achievement Tests).

The equating functions for the 1990 SAT Total Mathematics and the NAEP Overall Proficiency scores (corresponding to total group percentiles of 95, 90, 75, 50, 25, 10, and 5) for the state total and for male and female students were established. According to the authors, if all conditions required for equating are satisfied, the equating functions should be invariant across subpopulations except for sampling error. However, the authors found

that for state 1, a given score on the SAT would be converted to a somewhat higher score on the NAEP if the equating function for male students were used rather than the equating function for female students. The difference between the two equating functions tends to be larger at the low end of the distribution than at the high end. The equating functions for state 2 were similar to those for state 1, but the difference between the male and female equating functions almost disappears for SAT scores of 91 or higher.

In addition, the authors were concerned about the differences in content between the two tests and thus performed separate equipercentile equatings using the SAT Total Mathematics scores and the NAEP "Numbers and Operations" scores (since the majority of the items on the standardized tests belong to Numbers and Operations). However, the equating functions were similar.

When the 1992 estimated NAEP scores obtained using the 1990 equating functions were compared to the 1992 actual NAEP scores for state1, generally the differences between estimated and obtained scores were reasonably small. The differences were greater than twice the standard error only at the low end of the distribution (5^{th} and 10^{th} percentiles). For state 2, a new form of the SAT (form L) used in 1992 was equated to the form used in 1990 and then mapped into the NAEP scale. According to the authors, estimated and observed performance on NAEP was similar for the low half of the distribution, whereas the observed performance was higher than the estimated performance for the top half (differences exceeded two standard errors). Additionally, equipercentile equating underestimated the 1992 NAEP scores for state 3, particularly at percentiles of 75, 90, and 95, while equipercentile equating overestimated the 1992 NAEP scores for state 4, particularly above the median and at the 5^{th} percentile.

In summary, the authors argue that the difference between equating functions for male and female students was larger than expected based on sampling error alone for some parts of the distributions (for states 1 and 2). The differences in the region between the 5^{th} and 95^{th} percentiles are as large as 11 points for state 1 and 8 points for state 2. The authors maintain that when 1990 equating functions were used with 1992 statewide test data to estimate the 1992 NAEP results, differences between estimated and observed NAEP scores were larger than expected in one or both tails of the distribution in all four states. Thus, the authors conclude that the linking might be considered adequate for purposes of estimating average achievement on the NAEP scale, but not for estimating performance at the lower or upper ends of the distribution.

A study by Ercikan (1997) also used equipercentile equating methods to link 1990 state-level test results (California Achievement Tests, Form E [CAT/E], and Comprehensive Tests of Basic Skills, Fourth Edition [CTBS/4]) in four states for eighth grade mathematics to the 1990 NAEP mathematics scale. In particular, the study focused on an examination of whether the population property holds for the linking function, when the function obtained for an individual state is compared to the functions obtained for other states either individually or combined.

For content match, the author reports that CAT/E was designed according to objectives fairly similar to those of CAT/5 and CTBS/4 designed based on the same content blueprint. According to the author, the CAT/5 items corresponded to 26, CTBS/4 items to 27, and CAT/E items to 17 out of 46 of the NAEP objectives. When standardized mean differences for each state pair were computed using pooled within-state standard deviation for the statewide and NAEP tests, the author found that the standardized

differences based on the statewide tests were much smaller than those based on NAEP, indicating the content coverage between the two sets of tests were different.

The author states that the total mathematics scale scores from CAT/E or CTBS/4 were converted to CAT/5 scale scores and then converted to the Normal Curve Equivalent (NCE) scale of the CAT/5 to provide comparability of results across states and to reduce effects due to differential statewide testing dates. The first set of analyses linked each state separately, which involved conversion of scale scores from statewide tests to the CAT/5 scale, then to CAT/5 NCEs, and finally linking of NCE distributions to NAEP scale score distributions. The author comments that the second set combined data from three states that had all used CTBS/4 tests, excluding state 3 that had used the CAT/E test. The third set combined data from four states. The author assumed that if the conditions for equating were fully satisfied, the linking function would be identical for each analysis set and each state except for sampling error.

The author found that when the NAEP scores for NCE scores of 10, 25, 50, 75, and 90 from each of the analyses were computed, there were large differences in the linked NAEP scores for each state separately based on analysis set 1. For example, an NCE score of 90 on the CAT/5 predicted NAEP scores ranging from a low of 305 in one state to a high of 325 in another state. According to the author, these differences were up to 17 times the estimated standard errors, whereas the differences were very small in the two linking analyses (Analysis Sets 2 and 3). She argues that when the linking function from each analysis was represented by CAT/5 NCE versus NAEP score plots, the plots also showed that the relations between NCEs and NAEP scale scores varied from state to state.

In addition, the author found that when she examined the difference between the actual and linked NAEP median scores using the linking function based on Analysis Set 3, the difference at the median was 9 NAEP score points for state 1, 10 points for state 2, 5 points for state 3, and 5 points for state 4. The sets of linkings were most similar at the middle of the scale for all four states.

From NAEP's score cutpoints corresponding to the achievement levels (basic, proficient, and advanced) for grade 8 and their corresponding CAT/5 NCEs (using the data from all four states pooled), the author found that the actual median scores for states 1, 3, 4 belong to the Basic level, and that the score for state 2 is Below the Basic level. She added that the linked NAEP median scores for all the states belong to the Basic level. When the actual percentages of students in each of the achievement levels for each state was compared with those obtained by Analysis Set 3, the author found that the differences between the actual and the linked percentages ranged from 1 to 14.

In conclusion, the author argued that the equating functions were different across the states and between the individual states' results and the combined results. She speculated that the differences in the linkage results may have been caused by differences in test administration dates, differences in motivation, and differences in content between the two tests. The author suggests that the link between state tests and NAEP does not provide precise information and that information from a linking study such as this one should be limited to rough estimates of percentages of students in each of the NAEP achievement levels.

Williams et al. (1998) employed projection methodology to link a North Carolina End-of-Grade (EOG) mathematics test for grade 8 to the NAEP scale. Linear regression

models were used to develop projection equations to predict state NAEP results in the future. The authors argue that there is considerable overlap in their content frameworks.

A total of 2824 students were tested, and two ethnic classifications were created for the projection analyses: BHN (Black, Hispanic, and Native American) and WA (White, Asian, and Other). The test administered in February 1994 contained 78 items, including a short version of the 1994 EOG mathematics test (40 multiple-choice items) and two blocks of released 1992 NAEP mathematics items (29 multiple-choice and 9 open-ended). Two forms of the test were administered: Form A contained the NAEP test first followed by the EOG test, while Form B contained the EOG test first.

First, the NAEP proficiency distribution for the sample is estimated based on the sum of a NAEP posterior distribution for each student obtained using examinees' responses, the population distribution, and the IRT item parameters from the calibration of the NAEP items on the single scale. For this study, the NAEP items were calibrated on one scale instead of the usual five mathematics subscales. Second, NAEP proficiency distributions were created for each EOG score by ethnic classification, but gender was excluded since it did not explain reliable variation in the conditional distributions in initial analyses.

A polygonal representation of each examinee's posterior proficiency distribution was used in estimating group distributions, instead of plausible values methodology. Third, the means of the NAEP distributions were predicted from the EOG score and ethnic classification, using weighted least squares regression analysis, in which the means were weighted by the number of students in each score by ethnicity category. Lastly, the

standard errors for the regression coefficients (means and standard deviations) were estimated using a bootstrap technique.

To examine the effects of test administration at different times of year on linkage results, a total of 2313 students from the NC-NAEP linkage sample were matched with their actual May EOG scores to project the February NAEP results from the regular May administration of the EOG test. Parameter estimates were again estimated using weighted least squares, and standard errors for the parameter estimates were computed using the bootstrap procedure. The authors found that the linkage results are very similar to the projection equations obtained by linking results from the February EOG to the February NAEP data. In addition, when the 1994 NAEP TSA results for North Carolina obtained directly from the linkage sample were compared with the 1994 projection from statewide administration of the MAY EOG test, the authors found that the distributions corresponded closely.

When comparing projected and observed 1996 NAEP results for North Carolina 8th-graders in mathematics using the 1994 projection based on the May EOG scores, the authors found that the means differ by less than twice the estimated standard error of their difference, but that the medians differ by more than twice the estimated standard error of their difference. Thus, the authors suggest that a linkage may begin to become obsolete in as little as two years.

In conclusion, the authors contended that overall, the NC-NAEP linkage permits comparisons to national data and to national standards, but that the linkage does not remain stable over time. The authors argue that the fact that the relation between scores

on the two tests is not the same can be accounted for in the linkage by conditioning on any variable (e.g., ethnic classification in this study) over which the relation varies.

Methodological Approaches Guiding This Study

Using a combination of data types (triangulation) enhances the validity of findings as the strengths of one approach can compensate for the weaknesses of another (Miller, 1997; Patton, 2002). Data for this study were collected primarily through interviews and documents. In this section, to seek guidance on how to analyze these multiple data sources altogether, two studies are reviewed that utilized multiple sources of information such as observations, interviews, and documents. This review focuses on the methodological approaches to analyzing these data.

Anspach's (1987) study of prognostic meetings in two hospital-based neonatal units focuses on the micropolitical significance of institutional texts within decision-making settings. The study examined prognostic conflict between physicians and nurses in making life-and-death decisions in intensive care units for infants. Data were collected in the course of 16 months of field research in intensive care nurseries of two hospitals that differed in their size, relative prestige, referral patterns, and the demographic composition of their clientele.

Data sources included (a) observations and informant interviews regarding the organizational context of decision-making; (b) observations of 22 cases that represented life-and-death decisions; and (c) formal interviews concerning prognostic reasoning and specific cases. The author attended weekly social service rounds and ethics rounds,

mortality review conferences, and decision-making conferences. These participant observations were supplemented by interviews with 58 staff members and 16 patients.

According to the author, there were three sources of information used in making predictions in medical settings studied: technical, perceptual, and interactive cues. *Technological cues* refer to any information obtained through diagnostic technology, while *perceptual cues* refer to information collected via direct perception of a patient including palpation, percussion, and observation. *Interactive cues* refer to information gathered through the social interaction between patient and practitioner including eye contact and aspects of infants' facial expression.

The author found that physicians and nurses differ substantially in their views of infants' prognoses because of their experiences in the intensive care nursery. For example, although all groups relied on a combination of technological and perceptual cues, a much larger number of nurses than physicians considered interactive cues in decision-making. These nurses sustained continuous contact with infants, whereas physicians had limited contact with infants and were technologically focused. Hence, the author argued that it was necessary to go beyond gross tabulations and to examine the interview and observation data to understand the nature of the prognostic process.

The author emphasized that the different daily work experiences of the physicians and nurses might provide access to different types of knowledge to be used in making predictions, thus resulting in conflicts concerning how to reach life-and-death decisions. The author reports that one part of prognostic meetings involved considering various medical texts and technological cues that both physicians and nurses treated as relevant,

but that they disagreed on the nurses' direct experiences with the infants. The author maintained that interviews confirmed evidence from theses observation data.

The author asserted that a major issue of conflict between these two groups involved their orientations to theses different sources of knowledge. She indicated that each group of practitioners might approach the life-and-death decisions based on a partial knowledge base, selectively filtered through the culture and social structure of circumstances. The author found that, not surprisingly, these conflicts were resolved by treating technological cues as more reliable than interactive cues. However, the author argued that these conflicts were unlikely reported in the infants' case records. Moreover, the author reports that many of the interactive cues noted by the nurses were rarely systematically entered into patients' charts. Clearly, institutional texts constructed to explain past decisions are very likely to overlook the openness and complexity of the decision-making process (Miller, 1997).

Conversation analysis (CA) is a method for investigating the structure and process of talk and interaction between humans, and CA studies use transcripts based on video and/or audio recordings made from naturally occurring social interactions (Perakyla, 2005). Marlaire and Maynard's (1990) study involved paying close attention to features of social interaction that are often glossed over by even very careful observers of institutional settings, focusing on the social organization of standardized testing as an interactional activity.

According to the authors, the testing practices under study were intended to measure children's cognitive skills by asking them to respond to diverse questions asked by clinicians. One type of question asked that the children recognize and elaborate on word-

association patterns, such as “bread is to eat as milk is to …?’ The clinicians recorded if children’s responses were appropriate, thus producing institutional texts that might be used later in assessing their levels of cognitive development.

The authors carried out this study in a special education section of a developmental disabilities clinic at a Midwestern university, which played a key role in the diagnostic processes of the clinic. Clinicians administered a test battery including different kinds of examinations that were chosen to sample particular domains of mental capability. The authors videotaped three clinicians who were individually paired with 10 children ranging in age from 3 to 8 years. The authors focused exclusively on the examination process and its social organization, with the assumption that actors produce social structures interactionally and locally.

According to the authors, the test-item sequences contained three parts: a testing prompt, a reply, and an acknowledgement. Firstly, one obvious source of variation in testing prompts is the nature of a particular subtest and the directions given. Some subtests are of the “point-and-say” kind, while others are of the “hear-and-say” type. Some questions required the child to “repeat after me,” while others were “fill-in-the-blank” questions requiring the child to complete an analogy. In addition, clinicians frequently initiated repair on a reply and thus modified the prompt over the course of several turns. In this sense, the authors found that testing prompts were formulated in the interaction between clinician and child, rather than being given as simple stimulus items.

Secondly, according to the authors, replies came in three varieties: (1) unmitigated replies were those that children produced straightforwardly and without hesitation; (2) absent replies were those in which the child declined to produce a reply; and (3) tentative

replies were those in which children provided a partial reply eliciting a repair initiation from the clinician, which was the focus of the analysis. The authors found that replies were very sensitive to their interactional environment, like testing prompts that elicited them. Thirdly, the authors stated that acknowledgement also displayed significant variability in relation to different interactional contingencies. They found systematic relationships between acknowledgement and the nature of replies. For example, acknowledgement was skipped when the child's attention and answers were proper and correct. However, when the child's hearing or seeing a prompt was not a problem and when the child's understanding of the test idiom was appropriate, an incorrect answer was treated as reflecting the child's intellectual ability.

Overall, the authors found that the tests involved more than questions and answers since clinicians responded to children's answers in a variety of verbal and non-verbal ways. For instance, the clinicians responded to children's correct answers by saying 'good' and to their incorrect answers with 'okay' or by re-asking the question. Sometimes the clinicians responded to children's answers by asking for specifying their answers further. In conclusion, the authors argued that test-givers were inevitably implicated in testing situations and relationships on an item-by-item basis, and therefore test sores should be viewed as collaborative productions. They noted that these aspects of interactions were not discussed in documents that reported the outcomes of the testing process.

Discussion and Conclusions

The review of the history of and several aspects of the NAEP program was intended to provide readers with background information on the NAEP program so that they can better understand the case under study. The focus of this literature review was on issues relevant to alignment, NAEP validity, and the use of NAEP data that were directly associated with this study.

Alignment is critical to current efforts of systemic and standards-based education reforms. The review of this literature indicates that NAEP is reasonably aligned with its frameworks and national standards (Pearson & DeStefano, 1993; Silver & Kenney, 1993a, 1993b). The studies reviewed did not directly establish the rigor of the NAEP frameworks, but pointed out that the matrix design (e.g., the content-by-process matrix in mathematics) tends not to present the domains of mathematics and reading as an integrated body of knowledge involving a network of interconnected ideas and processes. In fact, the most recent NAEP evaluation suggested that in general, NAEP frameworks in reading, mathematics, and science were reasonably well balanced with respect to current disciplinary reform efforts and common classroom practices, but that the frameworks still did not adequately reflect contemporary research and theory from cognitive science and the subject-area disciplines about how students understand and learn (Pellegrino et al., 1999). At this point, some questions are raised. How do state education personnel perceive NAEP frameworks? Do states refer to NAEP frameworks when developing their state standards? Little research has been conducted on these issues.

An understanding of the psychometric characteristics of NAEP was deemed to help state education staff draw valid inferences from NAEP results. This review focused on providing an overall understanding of the validity of specific interpretations of NAEP results and helping identify interpretations and uses that are suspect or areas where the assessment needs to be strengthened. The literature suggests that NAEP lacks many social context measures that might relate to achievement (Berends & Koretz, 1995/1996), which indicates that data analysts would face limited family variables collected with NAEP that are strong predictors of performance. Moreover, the literature indicates that those variables collected by NAEP, especially data reported by 4th-grade students, might make their quality problematic (Berends & Koretz, 1995/1996). Yet, there exists the potential burden of collection of family factors such as family income through surveys of parents. Accordingly, previous NAEP evaluations suggested the integration of NAEP results with data on education inputs by non-NAEP sources (Glaser et al., 1997; Pellegrino et al., 1999). "How policy-relevant do state education personnel consider these disaggregated data provided by NAEP?" "How do state education personnel perceive the integration of NAEP results with indicators collected outside of NAEP?"

Burstein et al. (1995/1996) found that there were inconsistencies between actual student performance and both descriptions and exemplars of NAEP achievement levels. The findings reflect the major criticisms that have been leveled against the standards by various evaluations. Yet, very limited research has been conducted on examining the perceptions of NAEP users' perspectives on the usefulness of the NAEP achievement levels. An investigation of how state education personnel view the technical aspects

surrounding the achievement levels is beyond the scope of this study, but the current study explored their perceptions of the use of the achievement levels.

Research studies on the actual uses of NAEP directly related to the present study. Sebring and Boruch's study (1983) explored the use of NAEP at the federal, state, and local levels. The study found that NAEP products and services were used for multiple purposes by various audiences and that the use of NAEP data to make decisions was less common than to understand relevant issues. A study by Bullock and DeStefano (1998) examined the usefulness of the 1992 TSA in reading and the actual use of the state NAEP results through interviews with state assessment directors. The study revealed that the results had been used primarily to inform policymakers at state and local levels and to provide information to teachers, administrators, and the public in terms of how the states were performing comparatively.

There was a long time gap between those two studies and moreover, there has been little research on actual uses of NAEP since then. There have been several changes in NAEP in terms of design and reporting since the two studies were conducted. For example, congressional authorization of state NAEP was made in 1994 and thus the term "trial" was dropped beginning with the 1996 state NAEP assessment. In addition, more constructed-response items and performance-based item types were incorporated into NAEP tests. Further, state NAEP is not necessarily voluntary under NCLB and the position for a NAEP state coordinator has been created in each state for better communication between NAEP and the SEA in conjunction with the NCLB mandate for states. In addition, NAEP reports and resources have been posted on the NAEP web site since 1996, thus leading to innovations in dissemination of NAEP. These changes in

NAEP evoked the following questions. “How do state education personnel perceive the usefulness of NAEP?” “Are NAEP data currently used by state education staff more than before?” “How are NAEP data used in the NCLB accountability system at the state level?”

Taken together, this literature review led to a big question of how NAEP data are currently used at the state level in the NCLB accountability system in terms of informing education and/or policy decision-making processes. It was assumed that NAEP results and resources might be utilized more than ever before within this context. This study was conducted to investigate the state-level use of NAEP.

CHAPTER III

DESIGN AND METHOD

NAEP data provide a good source of information from a state perspective in terms of how students in the state are performing against NAEP criteria compared to the nation and other states. States might assess progress in the achievement of their students in a broader framework through this large-scale survey with a design that uses representative samples of students to make statistical estimates at the national and state levels. However, it has been widely speculated that the use of NAEP data are limited, and the literature review of Chapter II suggests that if NAEP data could be made more accessible and useful, their utility would be improved. Since mid-1990s, NAEP information and resources have been provided on the NAEP website and further the NCLB forces states to pay greater attention to state NAEP results. In this context, it was assumed that the use of NAEP is recently increased at the state level and this study was intended to empirically verify the assumption.

Although the ways in which NAEP might be used independently to monitor state achievement trends are not specified in NCLB, there might be potential uses of the results from the required biennial State NAEP. What do state education personnel think about this issue? How do they perceive the use of NAEP? How do they use NAEP data to inform their educational decisions? This study first examined the potential utility of NAEP data posted the NAEP website from a state perspective. Then the study investigated the perceptions on the usefulness of NAEP of state education personnel and how NAEP data are actually used in a state context in relation to their potential utility.

Qualitative methods of data collection and analysis were utilized for this case study. The following three research questions were addressed for the present study:

1. What is the nature of the NCES website in terms of NAEP as sources of data to inform state educational decisions?
2. What are the state education personnel's perceptions of the use of NAEP in making informed educational decisions?
3. How are NAEP data used in supporting the state in responding to current issues in education?

Method

In this investigation, a case study approach was employed to provide thick description and detail of a phenomenon of interest. "A case study is both a process of inquiry about the case and the product of that inquiry" (Stake, 2005, p. 444). Stake identified three types of case study: 1) intrinsic case study; 2) instrumental case study; 3) and collective (or multiple) case study. *Intrinsic* case study is conducted to better understand a particular case, which itself is of interest (Stake, 2005). *Instrumenta*l case study is undertaken primarily to provide insight into an issue, and a particular case being examined advances an understanding of the issue. *Collective* (or multiple) case study is instrumental study extended to several cases where many cases may be examined altogether to investigate a phenomenon (Stake, 2005).

This study is very close to an instrumental case study since a case (US State education agency) was examined to advance an understanding of the use of NAEP data

by states across the country. According to Stake (2005), in instrumental case study the case itself is of secondary interest and plays a supportive role in facilitating an understanding of something else. All of the depth and detail looked at regarding the case helps researchers pursue an external interest. Stake (2005) argues that the case might be seen as typical of other cases or not and that the case is chosen to advance an understanding of the external interest that is larger in scope.

The researcher's experiential and contextual descriptions and interpretations are expected to assist readers in the construction of their knowledge about the case being studied. The case study helped identify how NAEP data are currently used to inform educational decisions and program planning at the state level. The case study approach allowed the provision of thick descriptions and detail of the state education personnel's perspectives on the usefulness of NAEP and their NAEP data use in a particular state context, thus helping the researcher pursue a better understanding of NAEP data use by states across the country in general.

There were two phases of the investigation: a) analysis of the NAEP website, and b) a case study of how NAEP data are currently used in a state context. Data collection and analysis were carried out through the two phases as depicted in Figure 1. The job descriptions of all the participants and types of documents used in the analysis are presented in Figure 1. In the first phase, the NAEP website was analyzed through thorough content analyses of the site and analysis of NAEP data provided in the system focusing on potential utility of NAEP information for state education personnel in making state educational decisions. Additional information was obtained through

interviews with NCES staff: one is responsible for the site management and database and one for supporting NAEP state coordinators.

In the second phase, a case study was conducted to investigate how state education personnel in US State perceived the usefulness of NAEP and used NAEP data in making educational decisions. This case study was performed using a combination of interviews, documentation, and the SEA's website analysis. The context of the study is presented in Figure 2. The case chosen was the US State's education agency. The context of the study in Figure 2 was primarily used to identify tasks to be carried out, documents to be analyzed, and interviews with the state education personnel to be conducted. Figure 2 illustrates the state context within which NAEP data are used by the state education personnel, including the US State's history of state assessment programs, the state's performance standards, its accountability system, and its assessment purposes. This study was an instrumental case study, and thus the researcher was deeply interested in the case under study but intended to probe into an understanding of the use of NAEP across the states.

Phase I. Analysis of the NAEP website

⇨ Navigation of the website
- Access
- Intention
- Content analysis
- Interpretation of features
- Potential utility

⇨ Interviews with NAEP staff
- The NAEP webmaster
- Project officer for state NAEP coordinators

Phase II. Case Study

⇨ Interviews with SEA staff
- Reading specialist
- Mathematics specialist
- Science specialist
- Chief policy officer
- NAEP state coordinator
- Education specialist
- Education program specialist

⇨ Document analysis
- Reports on student achievement
- Reports on state testing programs
- Newsletters
- Press release
- Minutes of State Board of Education meetings
- Periodic publications
- Special NAEP reports
- Memoranda
- Gap analysis reports

⇨ US SEA's website analysis
- Content
- Analysis
- Dissemination of NAEP information

Figure 1. Study Design

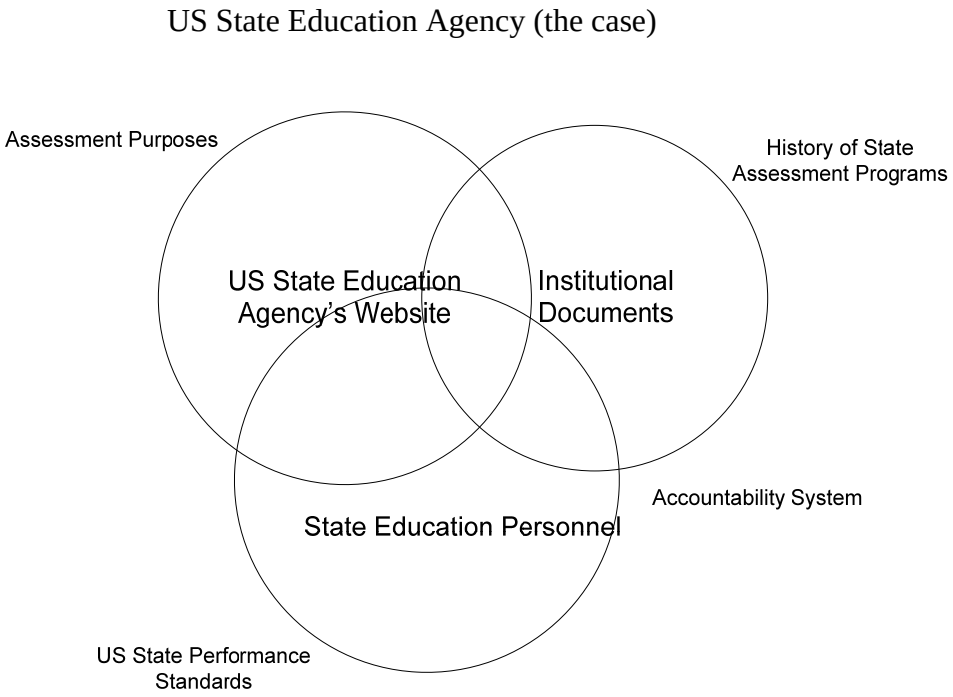


Figure 2. Context of the Study

Participants

The participants for the present investigation consisted of seven volunteers from a SEA and two volunteers from NCES. The state education personnel, who should be knowledgeable about the NAEP program, were recommended as potential interviewees by a NAEP state coordinator. The two volunteers from NCES were recommended to be interviewed through consultation with NCES. Confidentiality was maintained through the uses of identifying codes instead of participants' names and of a pseudonym for the state to ensure protection of the privacy of the participants and the identification of the organization where they worked. Participants' confidentiality will also be protected in publications based on this study by providing the opportunity for the participants to

review a manuscript in order to ensure that the manuscript accurately reflects what they said or intended to say.

State Education Personnel

The state NAEP coordinator in US State was asked to indicate state education staff members who were familiar with NAEP and had experience using NAEP data. She recommended six people including the state director of assessment, a chief policy officer, the mathematics specialist, and reading specialist. Due to the requirement to participate in biennial state NAEP in reading and mathematics at Grades 4 and 8, the researcher ensured that a mathematics specialist and a reading specialist were included.

Eight staff members were asked for interviews and six agreed to participate in the study. Later, after consultation with the NAEP state coordinator, a science specialist was contacted and participated in the study. Thus, a total of seven staff members participated in the present study. Informed Consent Forms (Appendix A) were explained and signed before the face-to-face interviews started. In the case of phone interviews, the consent forms were mailed before or immediately after the interviews and signed forms were then collected. The interviewees were given codes ranging from A to G to maintain confidentiality. These codes and their job descriptions are presented in Table 1. The information provided in the table was obtained during interviews.

NCES Staff

Two staff members at NCES were recommended as potential interviewees by NCES. After initial contact, a project officer for NAEP state coordinators agreed to participate in

an interview. The NAEP webmaster was contacted several times and finally agreed to participate indirectly by answering the researcher's questions via e-mail. Informed Consent Forms are presented in Appendix B. Interviews with NCES staff were intended to gain a deeper understanding of the role of NAEP state coordinators and the NAEP website design and management.

Table 1
Participant Codes and Job Descriptions

Code	Title	Job Description	Experience in current position	Experience w/ NAEP website
A	Reading assessment specialist	Supervision of construction of state reading assessment	10 years	None
B	Mathematics curriculum specialist	Supervision of development of state mathematics standards	3.5 year	Sometimes
C	NAEP state coordinator	-Overseeing administration of NAEP tests -NAEP data analysis & reporting -Liaison between state & NCES	3 months	Frequently
D	Education specialist	Provision of leadership & regional support for "State Reading First" initiative	5.8 years	Once
E	Chief policy officer	Management of policy development & implementation, & analysis of new policy needs	3.5 years	Many times
F	Science curriculum specialist	Supervision of development of state science standards	2 years	Sometimes
G	Education program specialist	- English language proficiency assessment - Real assessment for real success	4 months	Very often

Note. Participants are indicated by codes ranging from A to G.

Data Collection

Qualitative research methods were used to collect data for this study, including (a) interviews, (b) written documents, and (c) Web Sites. Using a combination of different data sources increases validity of findings and reduces the likelihood of misinterpretation (Hodder, 2000; Patton, 2002; Stake, 2005). The question of NAEP utility has two components: perceived utility and actual evidence of usefulness (Jaeger, 2003). Hence, research data on the utility of NAEP were collected through interviews and institutional documents. Data for the present study were collected over a period of seven months (from October 2005 to May 2006).

The NAEP Website

NAEP has expanded the dissemination of NAEP information and resources via its website and committed significant efforts to developing the website as a means of improving access to and use of NAEP products. This website is a user-responsive site, not merely one mimicking print publications. It provides much more data including web-based data tools than printed publications. For example, the NAEP website provides a variety of NAEP information that might help inform state-level educational decisions.

An analysis of the NAEP website was focused on the information being provided, as well as on the value of the information to various stakeholders, especially personnel in the SEA. In particular, attention was given to how appropriately the site was designed to answer questions that might be asked by its state-level users. This analysis was performed in advance of interviews with state education personnel in order to be better prepared to

construct interview questions and then probe participants' perspectives on the use of NAEP during interviews.

The structure of the website was first examined via navigation, referring to the site map on the site. Major sections were then explored, focusing on information that might help states make educational decisions. These sections were navigated carefully to find out what is available and not available in terms of assisting states in improving teaching and learning in the state. For example, the NAEP Data Explorer (NDE) was examined for its structure and features. A variety of its features were then explored using NAEP data (e.g., US State results) embedded in the NDE system to create statistical tables and graphics.

In fact, the NDE provides NAEP information on decades of NAEP results and background factors that may be related to achievement. The NDE was explored seeking answers to questions that might be asked by state education staff. Examples of such questions were: "What are the overall performance trends at each tested grade level over the last decade in US State?" "Have achievement gaps among race/ethnic groups narrowed in US State in the last decade?" " How has US State performed in comparison to the nation and other states?" Finally, their potential utility at the state level was examined in detail.

Interviews

Interviews are active interactions between the researcher and respondents that lead to negotiated, contextually and politically bound results (Fontana & Frey, 2005). Fontana and Frey (2005) argue that the meaning of the qualitative interview "is accomplished at

the intersection of the interviewer and the respondent" (p. 717). Thus, it was kept in mind that an interviewer is responsible for providing a framework within which interviewees can respond accurately and honestly to questions asked during an interview (Patton, 2002).

The interviews were conducted either in person or by phone. In this study, the interviews lasted for 40 through 90 minutes and all of the interviews were audiotaped. With state education personnel, open-ended interviews based on semi-structured protocol were conducted to elicit and probe their perceptions and perspectives on the usefulness of NAEP and their use of NAEP data. The focus was on how the participants are currently using NAEP data and resources as well as how they view NAEP issues in terms of the usefulness of NAEP. The following is a description of the semi-structured interview protocols for state education personnel and NCES staff, respectively.

State Education Personnel Interview questions were developed to answer research questions two and three: "What are the state education personnel's perceptions of the use of NAEP in making informed educational decisions?" "How are NAEP data used in supporting the state in responding to current issues in education? The construction of the interview protocol was based on the literature review on utility features of NAEP data and the analysis of the NAEP website. In this study, "NAEP data" included general NAEP information, NAEP materials, and methodology as well as the results of NAEP assessments.

The protocol had seven sections that ranged from background information to participants' perceptions on assessment itself and NAEP to NAEP data use. These seven

sections were based on the research questions, in general and on issues addressed by the literature review and by previous NAEP evaluation studies, in particular. The seven sections included: background information, assessment purpose, NAEP data use, standards-based reporting, the NAEP website, information dissemination, and demands for information. Questions in the section six were asked only of the NAEP state coordinator since the dissemination of NAEP information within the state was part of the coordinator's job. These questions were all open-ended (Appendix C), and the general questions asked were as follows:

- How do you obtain NAEP information?
- How have you used NAEP data?
- How policy-relevant do you find NAEP's findings?
- How useful do you find state-to-state comparisons?
- It might be possible that under NCLB, NAEP could be used to confirm general trends in state performance on the statewide test. What would be you think about this matter?
- How do you feel about NAEP performance standards?
- What questions would you expect to be answered through navigation of the website from a state perspective?
- What is the protocol in your state for communicating results?
- Have you made informal requests for information through the NAEP site?

Below are example questions for each section:

1. Section1: Background information

- Please tell me about your primary responsibilities.

2. Section 2: Assessment purpose
    - What purposes do you expect assessments to serve?
3. Section 3: NAEP data use
    - How have you used NAEP data? Please specify.
    - How useful do you find state-to-state comparisons?
    - If NAEP might be used to audit state measures of yearly educational progress under NCLB, how would you react?
4. Section 4: Standards-based reporting
    - How easy to understand do you find the achievement-level descriptions provided with NAEP results to aid in interpretation?
5. Section 5: The NAEP website
    - What NAEP product(s) provided on the site have you used?
    - What questions do you expect to be answered through navigation of the website from the state perspective?
6. Section 6: Information dissemination
    - How does NCES assist in producing your state's own reports?
7. Section 7: Demands for information
    - Have you made informal requests for information through the NAEP site? If yes, how?

Once the protocol had been designed, pilot interviews were conducted with a staff member from the SEA and a director of research and evaluation center for classroom

teaching and learning at Northwest Regional Educational Laboratory. They provided recommendations for improvement, and the protocol was revised accordingly. Further, the interview questions were very slightly refined for clarification as the interviews went on.

When time permitted, several questions (arranged in the end of the section "The NAEP website") were asked to probe their perceptions on some of the issues addressed for improvement by previous NAEP evaluation studies (Glaser et al., 1997; Pellegrino et al., 1999). These questions included topics such as: reporting data on the various aspects of achievement, integration of NAEP results with education inputs from non-NAEP sources, inclusion of students with disabilities and English-language learners, and making NAEP results more visible and informative. Usually some of those questions were asked depending on time availability because pilot interviews suggested that 50 minutes was not sufficient to ask all questions in the protocol and that these topics might be unfamiliar to participants.

Either face-to-face or phone interviews were conducted by the researcher during March, April, and May 2006. Participants were sent copies of the interview questions in advance to clarify questions especially in the case of phone interviews and were given reminder e-mails two or three days before each interview. All of the interviews were taped in a recorder with permission of participants and transcribed. One participant was interviewed twice with his permission due to a recording failure. After the interviews, follow-up questions were asked when necessary.

NCES Staff It was essential to interview NCES staff in order to acquire an in-depth understanding of how NCES aids states in using NAEP data as well as of the site content and management. A request has been made through an e-mail form set up on the NAEP site, and then two people were indicated as potential interviewees. One is in charge of supporting NAEP state coordinators and the coordinators in the NAEP State Service Center, while the other in charge of managing the contractors who do the website development and maintenance. The former was interviewed via phone, while the other answered the researcher's questions by e-mail.

The interview questions were constructed based on the literature review, the analysis of the NAEP website, and interview guides suggested by Hert, Eschenfelder, McClure, Rubin, Taffet, Abend, and Pimentel (1999). Hert et al.'s study was conducted to evaluate several of the U.S. Department of Education websites to inform the Department's on-going efforts to improve these websites. One protocol had seven sections: background information, website purpose and audience, site content and management, log & transaction analysis, NAEP data use, user feedback & requests, and evaluation processes (Appendix D). The other one had four sections: background information, NAEP state coordinators, NAEP data use, and information dissemination. These interview questions were all open-ended (Appendix E).

The major questions asked were as follows:

Protocol B

- What is the purpose of the website?
- What is the process for creating content and getting it on the website?

- How does NCES identify information needs and information preferences of the site's users?
- How does NCES actively solicit the input of states to better support them in terms of NAEP data use?
- How does NCES respond to ongoing user demands for information?
- How does NCES assess the usefulness of the site from the perspective of users?

Protocol C

- Why has the position for NAEP state coordinators been established?
- What is the job of a NAEP state coordinator?
- How does NCES provide ongoing support for NAEP State Coordinators?
- How does NCES actively solicit the input of NAEP state coordinators to better support states in terms of NAEP data use for policy and/or education decisions?
- In what ways does NCES involve state departments of education in facilitating accessibility and dissemination of NAEP results within their states to make NAEP data more accessible to intended audiences?
- How does NCES support states in generating their own reports of NAEP findings customized to their own situations and needs?

Below are example questions for each section of the protocol B:

1. Section1: Background information
    - Please tell me about your primary responsibilities.
2. Section 2: Website purpose and audience

- What is the purpose(s) of the NAEP website?
- What are the audiences intended for the site?

3. Section 3: Site content and management
    - What is the process for creating content?
    - What is the process for getting it on the website?
    - How is the site reviewed?
4. Section 4: Log & transaction analysis
    - How does NCES regularly analyze data in the web server to better understand the ways in which users use the website?
5. Section 5: NAEP data use
    - How does NCES identify information needs and information preferences of the site users?
    - How does NCES actively solicit the input of states to better support them in terms of NAEP data use?
6. Section 6: User feedback & requests
    - What mechanisms are in place to receive feedback or requests from users of the site?
    - How do good ideas feed into the content provision process?
7. Section 7: Evaluation processes
    - What processes are in place to evaluate the success of the website on a regular basis?

Below are example questions for each section of the protocol C:

1. Section1: Background information
    - Please tell me about your primary responsibilities.
2. Section 2: NAEP State Coordinators
    - What is the purpose(s) of the NAEP State Service Center?
    - How does NCES aid NAEP state coordinators in serving as the liaison between the State Education Agency and NAEP?
3. Section 3: NAEP Data Use
    - What kinds of expectations does NCES perceive states have about NAEP?
    - How does NCES identify NAEP information needs and information preferences of the states?
    - Has NCES conducted research focusing on the usefulness of NAEP data provided on the NAEP website from a perspective of states?
4. Section 4: Information Dissemination
    - How does NCES support states in generating their own reports of NAEP findings customized to their own situations and needs?

The initial interview questions were reviewed by science education faculty in a university, and their suggestions for improvement were implemented. A phone interview with one person was conducted by the researcher in early March, and the other answered interview questions by e-mail in mid-March. Participants were sent copies of the interview questions in advance to clarify questions, and the interviewee was given reminder e-mail one day before the phone interview. The interview was taped with permission of the participant.

Written Documents

Records and documents provide a rich source of information about organizations and programs including historical insight (Hodder, 2000; Patton, 2002). The information provided by texts may differ from and may not be available in spoken form. Institutional documents provide the cultures and operations of contemporary institutions (Miller, 1997) and might provide other specific details to corroborate evidence from other sources (Yin, 2003). In this study, evidence of how NAEP data and resources had been used at the state level was also sought through content analyses of written documents and records. Through consultation with a NAEP state coordinator, it was found that a wide array of reports (such as state reports on student achievement, reports on state testing programs, memoranda, proposals, minutes of meetings, press release, state assessments and education regulations, etc.) were posted on the SEA's website.

Documents reviewed included: memoranda (1999-2006), agendas, policy publications, press releases (1997-2006), test administration manuals (2005-2006), minutes of State Board of Education meetings (October 2005-April 2006), State Board white papers, newsletters, NAEP-state reading assessment linkage study: Report to US State, gap analysis reports, and periodic publications by the SEA such as Assessment Updates (5/12/05-4/24/06), Superintendent's Pipeline (2004-05 ~2005-06), Academic Content Standards, Annual Statewide Report Cards (2000-2005), and State Standards Newspaper (2005-2006).

Some of these documents were indicated during interviews. When documents indicated by interviewees were not provided on the website, the researcher asked them for these documents. For example, the mathematics specialist stated that her group had

presented to the State Board of Education evidence that the state mathematics standards revised in 2002 were aligned with NAEP frameworks. From follow-up questions, it was found that a match gap analysis had been done between the state standards and NAEP frameworks but that the document was not available on the site. Thus, the side-by-side analysis of the standards and NAEP frameworks was obtained through the specialist.

In summary, documents included rich information on historical insight into institutions and their programs. In addition, documents and reports sometimes provided the information that might not be obtained through interviews. For example, some participants of this study were unaware that some uses (e.g., use of NAEP data in meetings, a match gap analysis, linking study) constituted use or they did not simply remember some of their uses. Therefore, the focus was primarily given to finding out whether evidence from interview data was consistent with that from document analysis and whether documents contained information that might not be available from interviews.

State Education Agency's NAEP website

Yin (2003) argues that physical artifacts might be an important component in the overall case being studied, when relevant. Most of state education department websites include a NAEP section where NAEP data and resources are introduced to parents, educators, and the public in the state. In this sense, an SEA's website is helpful in developing a more precise understanding of how NAEP data and resources are currently disseminated and utilized in a state context.

The SEA in US State also has posted NAEP information on its website. The NAEP website consisted of News Announcements, Pages (FAQ, NAEP frameworks, NAEP glossary, NAEP news, NAEP resources for more information, NAEP results, and NAEP sample questions), Miscellaneous internal Links, External Links, and Contacts. The focus was on examining the site content in terms of assisting potential NAEP users in the state in better understanding NAEP and providing useful tools for interpretation and understanding of the data.

Timeline for Data Collection

The primary sources of data were collected through interviews, documents, and websites. The overall data collection was conducted in four stages, but those stages were not exactly linear. The first stage was to analyze the NAEP website to find out what is available in terms of NAEP data use at the state level in preparation for interviews with state education personnel in US State. The second stage was to conduct interviews with NCES staff. The intent of the interviews was to gain a better understanding of the operation of the NAEP site, as well as the communication between the SEA and NCES in terms of promoting the understanding of NAEP and NAEP data use. The third stage was interviewing with state education personnel, which was intended to probe their perceptions on the use of NAEP and actual uses of NAEP. The last stage included analyses of documents and records and the state's NAEP website to examine the actual evidence of utility. The focus of this website was on how the site was designed to assist districts and parents in understanding and using NAEP resources. Finally, all of these data sources were used to triangulate the research data.

Data Analysis

Case study analysis was used to address the research questions of interest due to the nature of data collection methods employed. According to Stake (2005), case studies concentrate on experiential knowledge of the case and close attention to the influence of its social, political, and other contexts. Thus, this case study analysis was conducted to produce thick and detailed description of participants' experiences. Further, this case study focused on describing and interpreting participant's activities, contexts, and organizational processes. In order to conduct a high-quality analysis, attempts were made to (a) attend to all the evidence available; (b) present the evidence separate from any interpretation; (c) show adequate concern for exploring alternative interpretations; and (d) address the most significant aspect of this study, as noted by Yin (2003).

Interviews

The analysis of the interview data was an ongoing and iterative process. For data from state education personnel, the continual process of comparing episodes within and across sections in interview protocols began during the data collection process and continued after the data were collected. A summary of each participant's views was generated for comparison. Then summaries of responses were prepared using interview transcriptions for each section to compare responses across interviewees.

All participants' responses to interview questions were analyzed in the same manner. First, audiotapes were transcribed for analysis. Coding and analysis were done whenever responses were given. In most cases, a set of tentative codes was not

established at the outset. Instead, important aspects from responses were picked out as provisional codes that could be revised later. Responses were then categorized according to themes or patterns that emerged during the analysis process, which was centered mostly on the major issues asked. For example, these responses were categorized into 15 themes: assessment purpose, NAEP information source, NAEP data use, policy relevance of data, state-to-state comparisons, relationship between NAEP & OSAT, NAEP under NCLB, improving NAEP utility, achievement levels, consistency of achievement-level descriptions, NAEP site use, expectations for NAEP site, reporting of higher-order thinking skills, integration of NAEP results with other sources, NAEP inclusion, and use of web features. All interviewee's responses to each category were summarized and compared.

Several rounds of analysis were conducted to check against confirmatory or otherwise contradictory evidence in the data sources, making modifications accordingly. Responses to different questions asked were sometimes found to be overlapped. Thus, when reporting results, the categories were reorganized. For example, for reporting of their perceptions of the usefulness of NAEP, five categories were generated: 1) policy relevance of NAEP data; 2) state-to-state comparisons; 3) NAEP under NCLB; 4) NAEP performance levels; and 5) improvement of NAEP (integration of NAEP results, higher-order thinking skills, NAEP inclusion, and use of web's features).

Written Documents

Document analysis provides a behind-the-scenes look at a program or an event that may not be directly observable and may not be found through interviews (Patton, 2002). Since most documents and records were produced for a specific purpose and a specific audience (Yin, 2003), these conditions need to be appreciated in analyzing texts. In this study, actual evidence of NAEP data use at the state level was sought through content analyses of documents and records. Document analysis was focused on analyzing how texts were constructed and used in the SEA in order to better understand the actual uses of NAEP information within institutional contexts.

In particular, focus was on whether there was a match between interview data and documents. For example, a mathematics specialist stated that the last revised state mathematics standards were based primarily on NAEP frameworks and evidence for that was presented to the State Board of Education. Thus, a report verifying her argument was obtained and examined. A chief policy officer also argued that the SEA did a gap analysis between what the state's existing standards contained and what was in NAEP frameworks and that when something was found missing in the state standards, they incorporated that information from relevant NAEP frameworks into the state standards and had them approved by the State Board of Education. A document entitled "Academic Content Standards" confirmed his comment. He also asserted that all of NAEP information is placed in an annual booklet entitled "Annual Statewide Report Cards" published by the SEA. These reports were reviewed as well.

In addition, some information, which was not available from interviews, was gained through analysis of documents. For instance, document analysis revealed that the SEA

had conducted a study linking the state's own 1998 reading test scores to NAEP. The study was performed to assess the correlation between NAEP and the statewide assessment results in reading. A draft of a report on that linkage study was reviewed. Another example was that the SEA utilized NAEP data to address the relationships between achievement and family factors that might affect student performance. This use was not identified through interviews.

Websites

Increasingly, federal agencies depend on web-based technologies to provide a wide range of information resources and services (Hert et al., 1999). Likewise, since its inception in1996, the NAEP website has grown and expanded in conjunction with the ongoing development of the web technologies. On-line information is part and parcel of the social world since it is developed by people who hold certain values, assumptions, goals, or power relationships (Eschenfelder & Miller, 2005). This descriptive analysis of the website was conducted by thoroughly navigating the whole website at http://nces.ed.gov/nationsreportcard and by analyzing NAEP data in the system whenever needed. This analysis focused not only on information being provided but also on the value of the information for state-level users. In addition, the focus of the website analysis was given to how the site encourages feedback and how it is designed to respond to user feedback, as indicated by Eschenfelder and Miller (2005). The major focus of the analysis was as follows:

- Examine how the content of the website is structured

- Examine what features of the content provided are useful from the state perspective and interpret those features
- Examine how the site is taking advantage of the Web's interactive and multimedia features in reporting student performance
- Examine the ease with which users could utilize web-based tools such as the NAEP Questions Tool, State Profiles, and the NAEP Data Explorer
- Check for contact information and the presence of instructions for providing feedback

The structure and function of the NAEP Data Explore (NDE) was first explored through using the tool repeatedly. One of the top initiatives in US State was closing the achievement gap. Therefore, the trend in the gap change was examined with the NDE. The first step with the NDE was to select criteria, including grades, subjects, jurisdictions, assessment years, and variables (student, family, and school factors that may relate to achievement). To identify trends in scale score differences in mathematics between major subgroups at grades 4 and 8, the variable "gaps and changes in gaps" was selected (data analysis was carried out by choosing only one specific grade at a time). The results were that any change in score gaps between either White and Black students or White and Hispanic students in US State was not statistically significant for both grades 4 and 8.

A NAEP section in the SEA's website was also analyzed to examine how NAEP information is currently organized online to disseminate NAEP information, facilitate an understanding of NAEP, and to promote NAEP data use across the state. The state NAEP

coordinator has posted NAEP reports, NAEP information, and relevant NAEP resources on the website. The content analysis of the website was intended to gain a more comprehensive understanding of the distribution and actual uses of NAEP data in the state context. When information addressed from interviews was not identified on the website, questions were asked of relevant interviewees. In addition, when less state reports of NAEP findings customized to the state's own situations and needs than expected were found to be available on the website, additional state NAEP reports generated by the state NAEP coordinator were requested and reviewed (e.g., NAEP-language and literacy issues: Factors related to student performance).

The Researcher

NAEP is a project that periodically surveys the knowledge, skills, and attitudes of young Americans. Its basic mission is to provide information on performance useful to educational decision-makers and practitioners. However, its assessment results have been of limited use, and the perspectives on the usefulness of NAEP by its users have not been fully investigated. In particular, the NCLB legislation has brought greater attention to NAEP more than ever. Accordingly, this study was aimed to investigate the use of NAEP in a state context. The researcher's academic background, teaching experiences, and interest in educational assessment appear to offer a context within which she analyzed and interpreted the data collected for this study. What follows is a brief description of her background and her perceptions of issues related to the present study.

The researcher holds a bachelor's degree in Biology. After graduation, she began her career as a biology teacher in high school in Seoul, Korea, with the hope that she would pursue a graduate degree in molecular biology later. While teaching at high school, the researcher became interested in scientific inquiry that represents the key conceptual basis for "scientific literacy" for all citizens. Korean high school science curricula at that time were primarily focused on scientific facts and concepts and cookbook style laboratory work. Students were learning science passively and spending most of their time in preparing for a college entrance exam. Thus, the researcher decided to pursue a PhD in Science Education in the USA in order to obtain a deeper perspective on current issues surrounding science education.

During course work in her PhD program, the researcher's attention was given to large-scale assessments. NAEP measures the progress of education periodically based on national and state probability samples, using innovative design, sampling, measurement, and reporting. What struck her was the fact that NAEP could enable educational researchers to detect and report significant shifts in student performance associated with major policy decisions (NCES, 2003). There has been a line of research studies on technical topics associated with this issue.

The researcher believes that the wealth of information that NAEP provides could help reform American schools toward better learning and teaching in multiple ways. However, previous research indicates that the rich NAEP data are under utilized. Moreover, there have been rare systematic investigations of how a variety of NAEP audiences are reached, what kinds of information are desired, and how NAEP data are actually used by

them. This lack of research on the use of NAEP led the researcher to conduct the present study in the hope that this study would make meaningful contributions in the field.

A national assessment similar to NAEP has been administered in Korea since 1998. The assessment was designed to measure the educational achievements of elementary, middle, and high school students and their progress over time. It purpose is to serve as a barometer of student learning. However, it appears evident that Korea's national assessment needs to be improved in several aspects to be further used to inform educational decisions being made by educators and policy makers across the country. Moreover, an independent website entirely dedicated to providing the achievement data and relevant resources including web-based data tools such as the NDE has not been created yet. After completing a PhD program, she plans to make a valuable contribution to advancing the national assessment program in Korea on the basis of her research experience with the current project as well as the quality education she has received through graduate work.

CHAPTER IV

RESULTS

This chapter presents the results of data analysis and is organized in five sections. The first section, *State Context*, presents a description of the US State context focusing on the state's assessment system and performance standards. The second section, *Nature of the NAEP Website*, presents an analysis of the website with some examples of their applications in terms of the potential utility of NAEP data to inform state-level educational decisions, which answers the first research question. The third section, State *Education Personnel's Perceptions of the Usefulness of NAEP*, examines state education personnel's perceptions of the usefulness of major features of NAEP, which answers the second research question. The fourth section, *NAEP Data Use*, explores how NAEP data have been used at the state context, which answers the third research question. The fifth section, *Use of the NAEP website*, examines the perceptions on NAEP products on the website of state education staff and their expectations for the website from the state perspective, which answers the second and third research questions. The last section, *Summary*, presents a summary of the findings of this study.

State Context

State assessment programs share some common purposes and methods, but they are also different. States differ in the emphases they place on test results, the use of the scores, the status of curricular reform, the educational policy climate, and the kinds of

attention test results receive from the press (National Research Council [NRC], 2000). These factors play an important role in setting the context and the environment within which testing occurs.

US State was selected for this study to investigate how personnel in the SEA perceive the use of NAEP and how they actually use NAEP data to make educational or policy decisions. US State has established a standards-based system with a focus on determining whether students have met the standards set for the Certificate of Initial Mastery (CIM) at grade 10. The state has also committed considerable efforts to improving student achievement through assessments aligned to the standards and used criterion-referenced examinations as a major component of its state assessment system (US State Department of Education [USDE], 2005).

US State has participated in state NAEP since its implementation, and its student performance on NAEP has been in the middle of the achievement distribution among states. This state is one of the states that changed state policy due to NAEP performance (NRC, 2000), but has not used NAEP data substantially to inform its educational policy. NAEP frameworks have influenced the development and/or revision of the state's standards, and the state adopted standards-based reporting following the NAEP models. However, the US State's standards are not highly aligned with NAEP frameworks, and its performance standards are different from NAEP achievement levels in several aspects (e.g., number of levels, standard-setting method).

Taken together, US State appears to stand in the middle among states in terms of NAEP data use and following NAEP models, and hence this case study is expected to help provide some insight into NAEP data use in the larger group of interest. This section

presents the context for testing and the testing program itself in US State to help readers to better understand the perceptions and use of NAEP by state education personnel illustrated in later sections.

US State Assessment Program

The 1995 amendments to the US State Education Act for the 21st Century led the state assessment program to a standards-based system, focusing on determining whether students have met the standards set for the Certificate of Initial Mastery (CIM) at around grade 10 (USDE, 2005). In addition, the revised assessment program includes content and performance standards for benchmarks leading to the CIM at grades 3, 5, and 8. Further, the NCLB Act led to the development of the state's grade-level foundations (Grades K-2) and grade-level standards (Grades 3-8 and CIM) in English/language arts and mathematics (USDE, 2005).

US State has had a criterion-referenced assessment program since 1991 and reported results according to standards. The US State Education Act for the 21st Century created three forms of assessment: 1) knowledge and skill tests; 2) state performance assessments; and 3) classroom work samples. The knowledge and skill tests are multiple-choice assessments, while performance assessments are administered only in writing and mathematics problem solving (mathematics problem solving currently suspended). Classroom work samples are completed in areas of mathematics, science, speaking, writing, and social sciences.

Overall, the US State assessment is different from NAEP tests in terms of test formats and stakes attached to the test results. For example, on NAEP approximately 50

percent of the questions are multiple choice and 50 percent are constructed response, while the state test is 100 % multiple choice. The state assessment is high stakes for schools and school districts, while NAEP is a low-stakes test. Further, the state's assessment is considered to be in higher alignment with its content standards than relevant NAEP frameworks.

US State Performance Standards

The state performance standards define how well students must perform on state assessments and classroom assessments leading to the CIM (USDE, 2005). The state has established performance standards for grades 3, 4, 5, 6, 7, 8, and 10 in the subject areas of reading/literature, writing, mathematics, science, and social sciences. The labels for the performance standards are: exceeds, meets, nearly, low, and very low.

The US State's performance standards are not similar to NAEP achievement levels in several aspects. First, US State has five achievement level categories (exceeds, meets, nearly, low, and very low), while NAEP has four (Advanced, Proficient, Basic, and Below Basic). Secondly, the names of achievement levels are different as shown above. Thirdly, the method that US State used to set cut-scores that distinguish the levels of performance is different from that used by NAEP. Lastly, US State has not developed achievement-level descriptions for its performance levels, while NAEP provides descriptors for each of the three achievement levels (Advanced, Proficient, and Basic). According to the state NAEP coordinator, the state will revise its performance standards during the winter 2006-2007 and relevant achievement-level descriptors will be created at that time.

Nature of the NAEP Website

The overall purpose of the NAEP website (http://nces.ed.gov/nationsreportcard/) is deemed to make NAEP information more accessible to its user, presenting much more information than printed reports, but information on what purpose(s) it serves is not provided on the website. According to the NAEP webmaster, its purpose is to provide National Assessment results for K-12 educational progress. The screenshot of the NAEP homepage is presented in Figure 3.

Figure 3. Homepage of the NAEP Website (captured on 12/18/06)

This section presents how the site is actually designed to serve its purposes. The website was primarily structured around six categories and thus the analysis of the website is described focusing on these categories. An analysis of each category of the site is presented with an indication of how it could be used by state education personnel. The

potential utility of NAEP results and resources is then addressed in terms of informing educational decisions at the state-level along with an analysis of NAEP data for US State when appropriate, which answers the first research question. Followed are examples of an analysis of NAEP data with the NAEP Data Explorer (NDE).

Overall, this section is organized in three subsections: the site structure, the site management, and potential utility of the site. The site structure and potential utility of the website are described based on a content analysis of the website, while the site management is presented on the basis of content analyses of the website and information provided by a NAEP person in charge of the site management.

The Site Structure

NAEP has more than 30 links on its homepage. The webpage can be broken into three major areas: left-hand navigation bar, horizontal navigation bar, and recent study results links. The horizontal navigation bar at the top has links to NAEP questions, NAEP Data Explorer, state profiles, and publications and appears on each page. On the basis of its content, the structure of the website was divided into six categories for this study: NAEP Subjects, About NAEP, Special Tools, NAEP Publications, Research e-Center, and Other Resources (Figure 1). These main categories (and their elements) are interconnected through links to the homepage and/or direct links to each other, and every component under each category has its own page. This section presents a brief description of the categories including how they work.

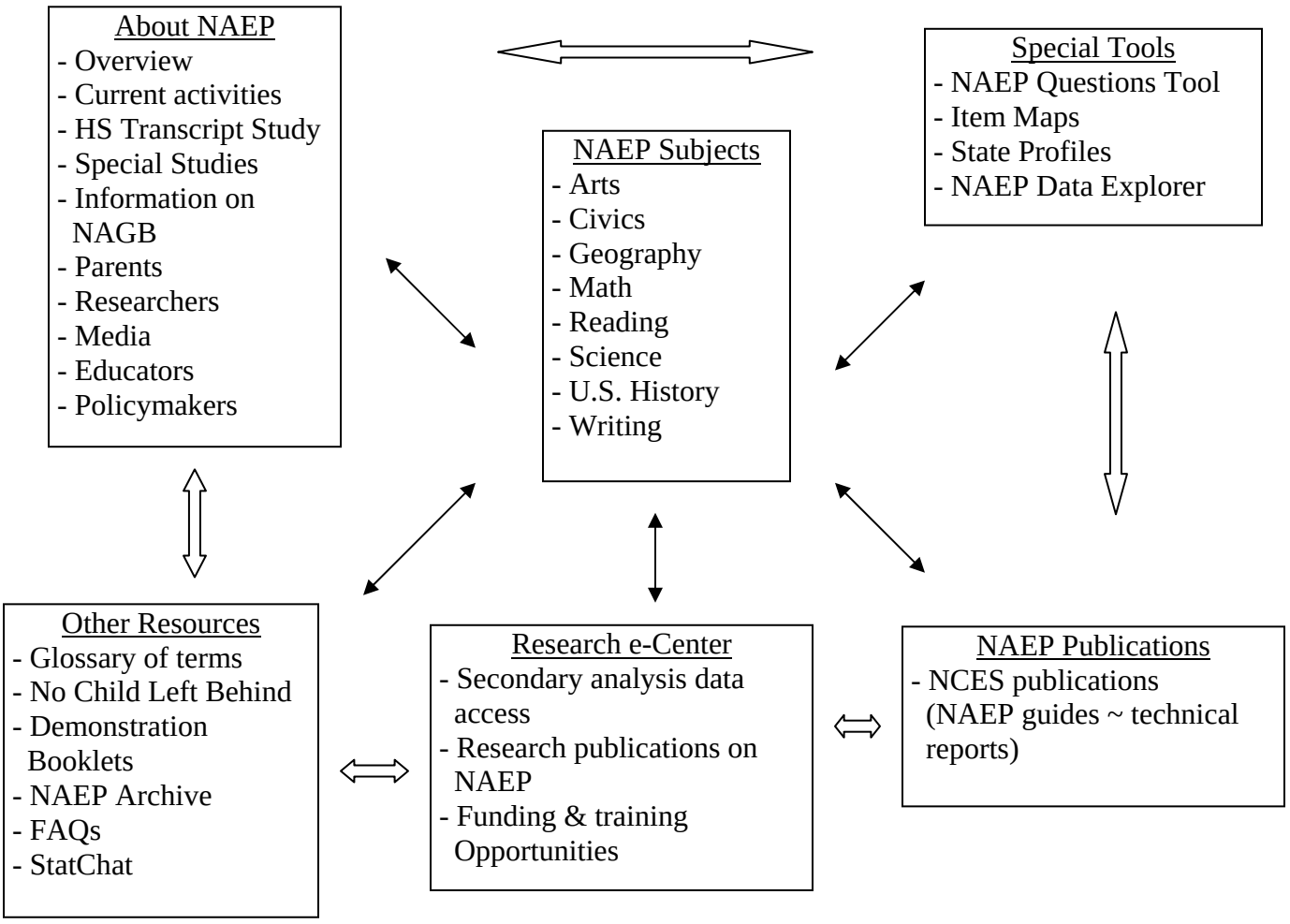


Figure 4. The structure of the Site

About NAEP This category provides an overview of NAEP, current NAEP activities, the long-term trend assessment, the high school transcript study, selected schools, and special studies. In addition, it provides NAEP information particularly targeted at a variety of audiences, such as parents, policymakers, the media, educators, and researchers. This category also provides updates, schedules, and information on NAGB, NAEP partners, and inclusion of special-needs students.

NAEP users including state education personnel can use this category to better understand the major features of NAEP including the structure and development of

NAEP, how the data are collected, scored, and analyzed, and how the results are reported. In particular, state education staff might explore the policymakers page dedicated to help inform the decisions they make.

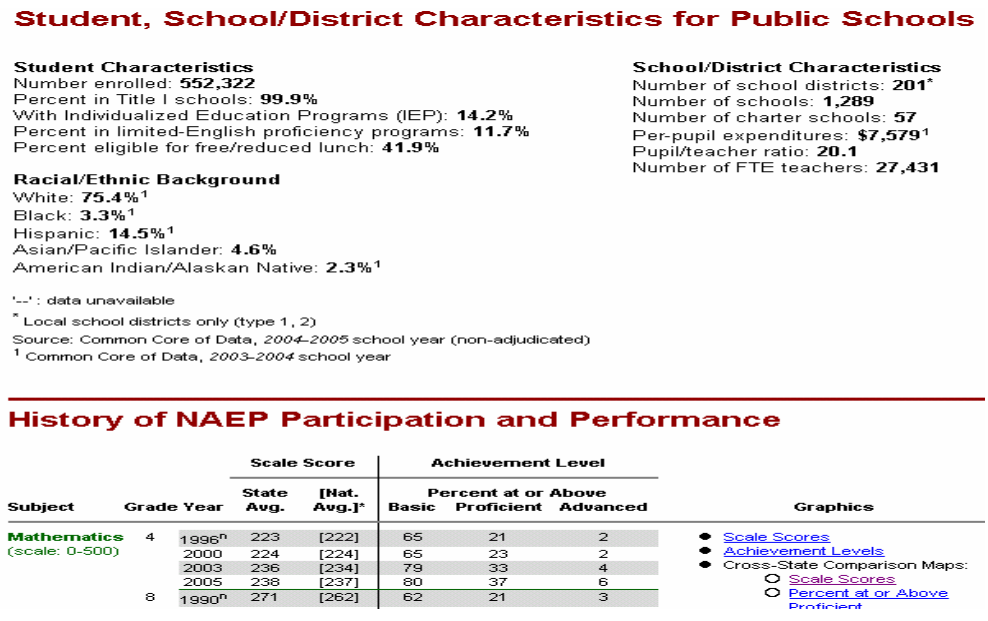
NAEP subjects This category provides information on eleven subjects, including reading, mathematics, science, writing, U.S. history, civics, geography, the arts, economics, world history, and foreign language. Within each subject page, users can find the most recent results and explore specifics through those results. For example, the mathematics page contains links to an executive summary of the results, the 2005 Mathematics Report Card, publications on NAEP mathematics, results from the 2004 long-term trend assessment in mathematics, 2005 state results and 2005 TUDA results, information on mathematics frameworks, test development, sampling, test administration, scaling, achievement-level descriptions, interpretation of NAEP results, and the NAEP Data Explorer (NDE).

NAEP users including state education staff could use information on this category to see the latest NAEP results for each of subject areas tested and relevant information on a specific subject including frameworks, achievement levels, item maps, sampling design, inclusion rates of students with disabilities and limited English proficiency, and NAEP scales. They can also explore sample questions linked to examine NAEP questions, student responses, and scoring guides that are released to the public. In addition, using the NAEP Data Explorer (NDE) linked in this category, state assessment staff might explore NAEP data for US State in depth and detail. Further discussion on NDE, frameworks, achievement levels, item maps, and released NAEP items is followed.

Special Tools This section includes the State Profiles, the Questions Tool, Item Maps, and the NAEP Data Explorer (NDE). These are web-based tools.

State Profiles

The state profiles present key data about each state's student and school population and its NAEP testing history and results (http://nces.ed.gov/nationsreportcard/states/, observed on10/26/2005). They compile data from Common Core Data and provide easy-to-use features for graphical presentation of results for each state. The profiles also present cross-state comparison maps by scale scores and achievement levels where each state's performance can be graphically displayed compared to other states, and the focal state can be changed using the NDE. The screenshots of the US State's profiles page and the cross-state comparison map by scale scores are presented in Figures 5 and 6, respectively.



Student, School/District Characteristics for Public Schools

Student Characteristics
Number enrolled: **552,322**
Percent in Title I schools: **99.9%**
With Individualized Education Programs (IEP): **14.2%**
Percent in limited-English proficiency programs: **11.7%**
Percent eligible for free/reduced lunch: **41.9%**

Racial/Ethnic Background
White: **75.4%**[1]
Black: **3.3%**[1]
Hispanic: **14.5%**[1]
Asian/Pacific Islander: **4.6%**
American Indian/Alaskan Native: **2.3%**[1]

School/District Characteristics
Number of school districts: **201***
Number of schools: **1,289**
Number of charter schools: **57**
Per-pupil expenditures: **$7,579**[1]
Pupil/teacher ratio: **20.1**
Number of FTE teachers: **27,431**

'--' : data unavailable
* Local school districts only (type 1, 2)
Source: Common Core of Data, *2004–2005* school year (non-adjudicated)
[1] Common Core of Data, *2003–2004* school year

History of NAEP Participation and Performance

Subject	Grade	Year	Scale Score: State Avg.	Scale Score: [Nat. Avg.]*	Achievement Level: Percent at or Above Basic	Proficient	Advanced
Mathematics (scale: 0-500)	4	1996[n]	223	[222]	65	21	2
		2000	224	[224]	65	23	2
		2003	236	[234]	79	33	4
		2005	238	[237]	80	37	6
	8	1990[n]	271	[262]	62	21	3

Graphics
- Scale Scores
- Achievement Levels
- Cross-State Comparison Maps:
  - Scale Scores
  - Percent at or Above Proficient

Figure 5. US State Profiles (captured on 12/18/06)

The problem using this tool is that the interactive cross-state comparison maps do not function in Mac as well as in PC. For example, when the cursor moves to other states, dynamic messages appear automatically that compare them with the focal jurisdiction. In Mac, these messages comparing the performance of the focal state statistically with another state do not appear. Moreover, users can make any jurisdiction become the focal state by simply clicking it on the map, but Mac computers do not support this function properly.

State education personnel can use this tool to find out the history of a state's participation in NAEP and its overall performance. Performance data since 1990 for each of participating states are summarized in tables in terms of scale scores and achievement-level results and these table data are automatically converted to graphics. They can also use the interactive US maps to compare their state's performance with other states and the nation graphically.

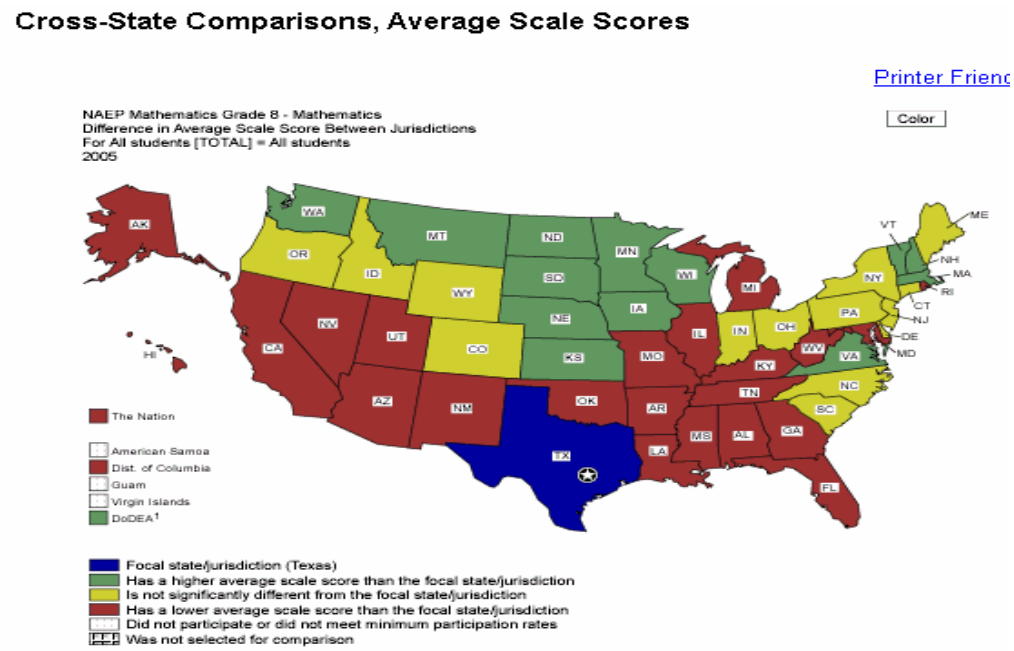


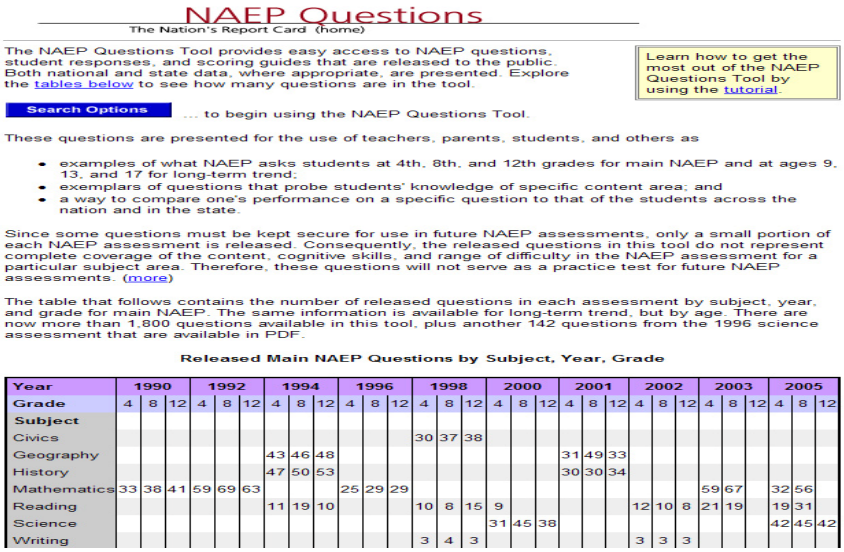
Figure 6. Cross-State Comparison Map (captured on 12/18/06)

Questions Tool

This tool is a web-based system that provides easy access to released NAEP questions, student responses, and scoring guides. National, state and long- term trend data are presented (http://nces.ed.gov/nationsreportcard/itmrls/, observed on10/31/2005). Questions can be searched within a subject by several factors, such as grade, content classification, question type, difficulty, and some subject-specific variables (e.g., mathematical complexity and mathematical ability in mathematics). For each question, *performance data*, *content classification*, *scoring guide/key*, *student responses*, and *more data* are provided:

1. *Performance data* provides a graphic indicating the percentage of students answering the item correctly, or the percentage of students at each score level for open-ended questions with more than two score levels.
2. *Content classification* indicates the content and ability categories the item represents and provides a description of these categories.
3. *Scoring guide/key* indicates the correct response option for multiple-choice questions. For open-ended questions, scoring rubric is provided.
4. *Student responses* shows examples of actual student responses to the essay questions for different score levels.
5. *More data* indicates the percentage of students selecting each response option for multiple-choice questions or the percentage at each score level for open-ended questions. It also provides cross-state data, and score percentages are disaggregated by demographic characteristics.

The screenshots of its front page and student responses, performance data, and cross-state data to a constructed-response question on grade 8 mathematics are presented in Figures 7, 8, 9 and 10, respectively.



NAEP Questions

The Nation's Report Card (home)

The NAEP Questions Tool provides easy access to NAEP questions, student responses, and scoring guides that are released to the public. Both national and state data, where appropriate, are presented. Explore the tables below to see how many questions are in the tool.

Learn how to get the most out of the NAEP Questions Tool by using the tutorial.

Search Options ... to begin using the NAEP Questions Tool.

These questions are presented for the use of teachers, parents, students, and others as

- examples of what NAEP asks students at 4th, 8th, and 12th grades for main NAEP and at ages 9, 13, and 17 for long-term trend;
- exemplars of questions that probe students' knowledge of specific content area; and
- a way to compare one's performance on a specific question to that of the students across the nation and in the state.

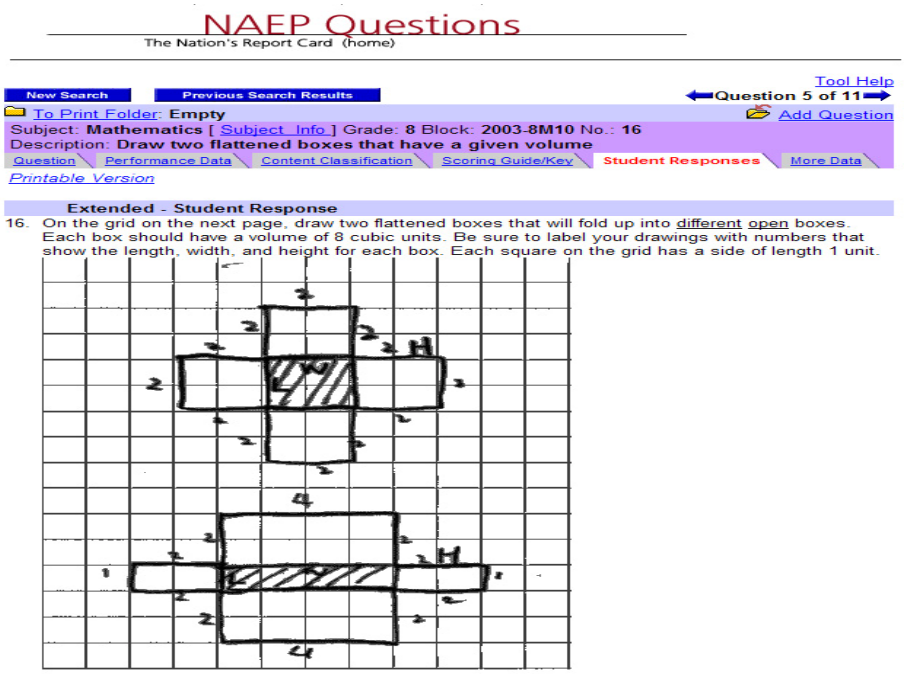
Since some questions must be kept secure for use in future NAEP assessments, only a small portion of each NAEP assessment is released. Consequently, the released questions in this tool do not represent complete coverage of the content, cognitive skills, and range of difficulty in the NAEP assessment for a particular subject area. Therefore, these questions will not serve as a practice test for future NAEP assessments. (more)

The table that follows contains the number of released questions in each assessment by subject, year, and grade for main NAEP. The same information is available for long-term trend, but by age. There are now more than 1,800 questions available in this tool, plus another 142 questions from the 1996 science assessment that are available in PDF.

Released Main NAEP Questions by Subject, Year, Grade

Year	1990			1992			1994			1996			1998			2000			2001			2002			2003			2005		
Grade	4	8	12	4	8	12	4	8	12	4	8	12	4	8	12	4	8	12	4	8	12	4	8	12	4	8	12	4	8	12
Subject																														
Civics													30	37	38															
Geography							43	46	48										31	49	33									
History							47	50	53										30	30	34									
Mathematics	33	38	41	59	69	63				25	29	29													59	67		32	56	
Reading							11	19	10				10	8	15	9						12	10	8	21	19		19	31	
Science																31	45	38										42	45	42
Writing													3	4	3							3	3	3						

Figure 7. Front Page of Questions Tool (captured on 12/18/06)



NAEP Questions

The Nation's Report Card (home)

New Search | Previous Search Results | Tool Help | Question 5 of 11

To Print Folder: Empty | Add Question

Subject: **Mathematics** [Subject Info] Grade: **8** Block: **2003-8M10** No.: **16**

Description: **Draw two flattened boxes that have a given volume**

Question | Performance Data | Content Classification | Scoring Guide/Key | Student Responses | More Data

Printable Version

Extended - Student Response

16. On the grid on the next page, draw two flattened boxes that will fold up into different open boxes. Each box should have a volume of 8 cubic units. Be sure to label your drawings with numbers that show the length, width, and height for each box. Each square on the grid has a side of length 1 unit.

Figure 8. Student Responses (captured on 12/18/06)

US State does not release its state assessment items on an annual basis since the items are used over and over again. Instead, the state agency provides sample questions on its web site, but does not have any web-based tool such as the NAEP Questions Tool available on the site. Hence, this tool could be used at the state and local levels for several purposes. For example, curriculum specialists at the SEA might use this question tool to identify content area weaknesses and strengths of students in the state and further compare them with other states and the nation. In the same sense, this tool can be used for diagnostic purposes. State education staff may share this information with teachers across the state so that teachers can improve their instruction. In addition, state assessment specialists might refer to NAEP items in developing items for their own state assessment.

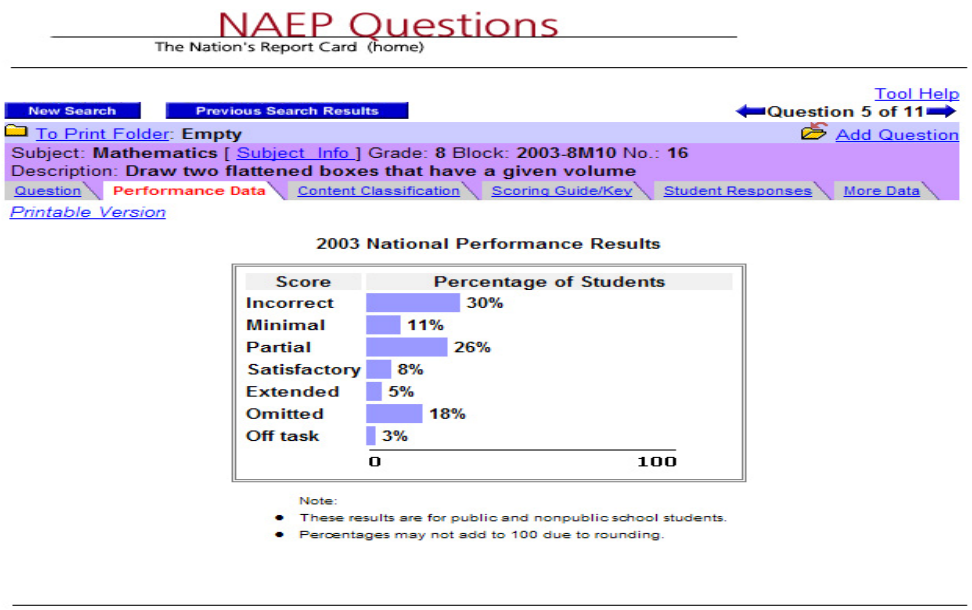


Figure 9. Performance Data (captured on 12/18/06)

Jurisdiction	Incorrect				Minimal				Partial				Satisfactory			
	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.
Alabama	255	(3.2)	32	(2.0)	‡	(‡)	13	(1.7)	273	(3.1)	27	(2.0)	‡	(‡)	4	(0.8)
Alaska	271	(3.4)	29	(2.2)	‡	(‡)	6	(1.2)	288	(3.3)	24	(1.6)	‡	(‡)	8	(1.3)
American Samoa	‡	(‡)	21	(4.4)	‡	(‡)	#	(***)	‡	(‡)	14	(4.3)	‡	(‡)	#	(***)
Arizona	271	(3.4)	29	(1.7)	‡	(‡)	7	(1.0)	280	(2.7)	27	(1.9)	‡	(‡)	9	(1.3)
Arkansas	257	(3.6)	35	(2.2)	247	(3.9)	13	(1.4)	277	(3.3)	26	(2.3)	‡	(‡)	4	(0.9)
Atlanta	243	(4.4)	28	(2.6)	‡	(‡)	22	(2.3)	‡	(‡)	14	(2.7)	‡	(‡)	2	(0.8)
Boston	‡	(‡)	27	(2.9)	‡	(‡)	10	(2.0)	‡	(‡)	18	(3.1)	‡	(‡)	5	(1.4)
California	263	(3.8)	30	(2.3)	249	(4.3)	9	(1.3)	272	(4.2)	24	(1.9)	304	(5.4)	9	(1.4)
Charlotte	262	(4.0)	32	(3.4)	‡	(‡)	10	(1.9)	‡	(‡)	22	(3.2)	‡	(‡)	9	(1.8)
Chicago	244	(3.1)	35	(2.8)	‡	(‡)	15	(2.1)	261	(4.0)	23	(2.5)	‡	(‡)	6	(1.5)
Cleveland	245	(3.9)	36	(3.9)	‡	(‡)	11	(2.0)	‡	(‡)	21	(2.6)	‡	(‡)	7	(1.6)
Colorado	275	(2.6)	29	(2.1)	‡	(‡)	9	(1.4)	291	(3.0)	28	(2.2)	‡	(‡)	12	(1.6)
Connecticut	277	(3.1)	27	(2.0)	‡	(‡)	11	(1.4)	291	(3.2)	20	(1.8)	‡	(‡)	10	(1.4)
Delaware	273	(3.3)	30	(2.2)	‡	(‡)	10	(1.5)	283	(3.4)	24	(1.8)	‡	(‡)	10	(1.3)
District of Columbia	243	(3.2)	33	(2.7)	232	(4.4)	19	(2.2)	‡	(‡)	17	(2.0)	‡	(‡)	3	(0.9)
District of Columbia (as district)	243	(3.2)	33	(2.7)	232	(4.4)	19	(2.2)	‡	(‡)	17	(2.0)	‡	(‡)	3	(0.9)
DoDEA	278	(3.0)	25	(1.9)	‡	(‡)	10	(1.4)	285	(2.4)	28	(2.2)	‡	(‡)	7	(1.0)
Florida	266	(3.4)	29	(1.9)	‡	(‡)	13	(1.6)	280	(3.6)	22	(2.4)	‡	(‡)	7	(1.4)
Georgia	259	(2.8)	30	(1.8)	251	(4.4)	15	(1.6)	281	(2.8)	24	(1.8)	‡	(‡)	6	(1.2)
Hawaii	265	(2.7)	31	(1.7)	‡	(‡)	7	(1.3)	276	(3.1)	25	(1.8)	‡	(‡)	5	(1.1)
Houston	259	(3.5)	36	(2.8)	‡	(‡)	13	(1.9)	270	(3.5)	23	(2.2)	‡	(‡)	6	(1.6)
Idaho	278	(2.4)	30	(2.2)	‡	(‡)	9	(1.4)	285	(2.8)	26	(2.2)	‡	(‡)	7	(1.3)
Illinois	264	(2.5)	34	(1.9)	254	(4.0)	11	(1.4)	280	(3.2)	24	(1.5)	‡	(‡)	8	(1.1)
Indiana	269	(3.1)	26	(2.0)	268	(4.0)	12	(1.6)	286	(1.8)	29	(2.2)	‡	(‡)	8	(1.1)
Iowa	276	(2.6)	31	(1.7)	273	(4.9)	11	(1.1)	288	(2.1)	28	(2.1)	‡	(‡)	7	(1.2)

Figure 10. Cross-State Data (captured on 12/18/06)

Item Maps

An item map is a visual representation combining scale scores, item descriptions, and achievement-level cutpoints in a way that highlights the kinds of questions a student can likely solve at a given level (Silver, Alacaci, & Stylianou, 2000). That is, the item map illustrates the knowledge and skills demonstrated by students performing at different scale scores on NAEP, and the three achievement-level marks are directly hyper-linked to relevant descriptions. The map location for each question represents the probability that, at any given score point, 65 % of the students (for a constructed-response question) and 74 % of the students (for a multiple-choice question) answered that question correctly (http://nces.ed.gov/nationsreportcard/itemmaps, observed on 11/2/05). Items released since the administration of the 2003 mathematics, 2005 reading, and 2005 science assessment are directly linked to the NAEP Questions Tool so that users can explore scoring guides, student responses, and performance data as well. The screenshot of the item map of the 2005 mathematics assessment for grade 8 is presented in Figure 11.

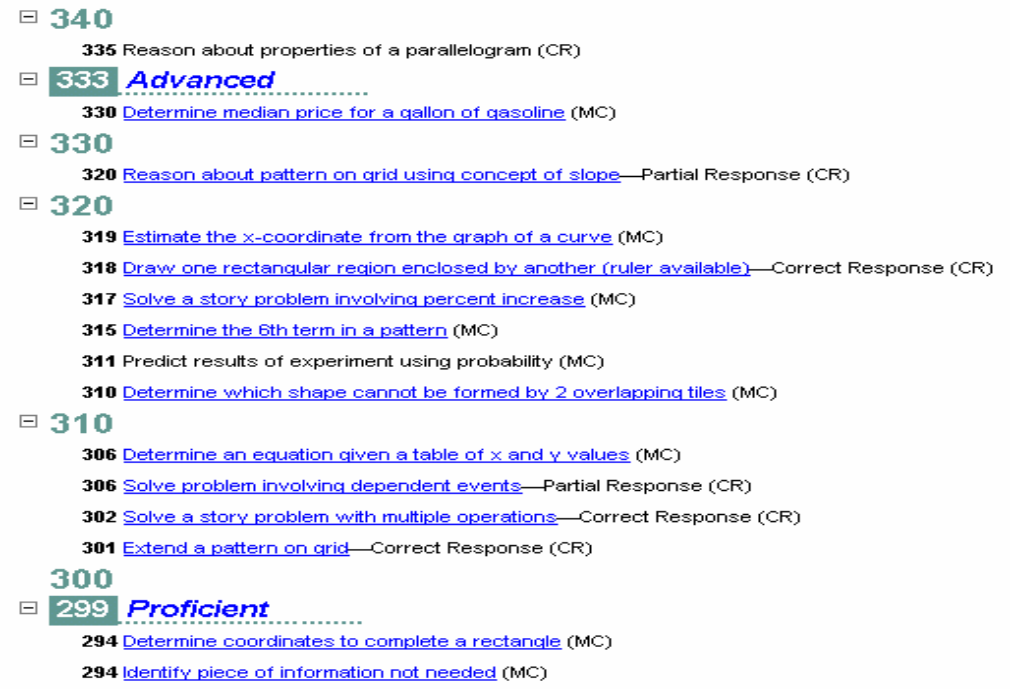


Figure 11. 2005 Mathematics Grade 8 Item Maps (captured on 12/18/06)

This kind of an item map is not provided for the US State's assessment. Accordingly, the state agency might use this tool to better understand NAEP scales and NAEP results. For example, state education personnel could use this tool to easily examine the specific knowledge and skills demonstrated by students who have performed at different scale scores on the NAEP scale. In this way, they might interpret the arbitrary NAEP scale scores in a more meaningful way. In addition, state assessment specialists could compare concepts students scoring at different levels on the scale have a grasp of (e.g. high-achieving vs. low-achieving students). By analyzing sample questions from different points on the map within each content strand along with performance data for each question (for multiple-choice) or student responses (for constructed-response), they might examine the type of material mastered within this content strand by students with varying degrees of mathematics proficiency.

However, to help users better understand subscales of each subject, the item mapping could be expanded to include an item map for each of the subject content areas for each grade. For example, item maps for grade 4 mathematics might include one for numbers and operations, one for algebra and functions, one for measurement, etc.

NAEP Data Explorer

The NAEP Data Explorer (NDE) is a web-based data analysis tool (http://nces.ed.gov/nationsreportcard/nde/, observed on 11/2/2006). The NDE allows for the disaggregation of data and comparison of state and national NAEP subject-area results. The tool is designed to search, display, and customize cross-tabulated variable tables using all major national and state assessments conducted for NAEP since 1990. The cross-tabulated tables can be displayed graphically. Users can also conduct statistical significance tests on the data in tables with the NDE. For example, users can view whether achievement gaps have changed significantly between years for subgroups of data with NDE. The screenshot of its front page is presented in Figure 12.

Figure 12. The Front Page of NDE (captured on 12/18/06)

To use the NDE, one should start an analysis with a question, such as: "How did the percentage of Hispanic grade 8 students in US State who performed at or above Proficient on the 2005 mathematics assessment compare to that of Hispanic grade 8 students at or above Proficient in the nation as a whole?" To answer this question, the NDE will calculate two statistics for the nation and for US State: the percent of Hispanic students at or above Proficient and its standard error. Only data generated by Hispanic students who scored at or above Proficient are used for the national and state calculations. The NDE then uses these statistics to run a pair-wide *t*-test to check the statistical significance of the performance difference between the two groups. The screenshots of these procedures (i.e., criteria selection, results table, and statistical significance test) are presented in Figures 13, 14, and 15, respectively.

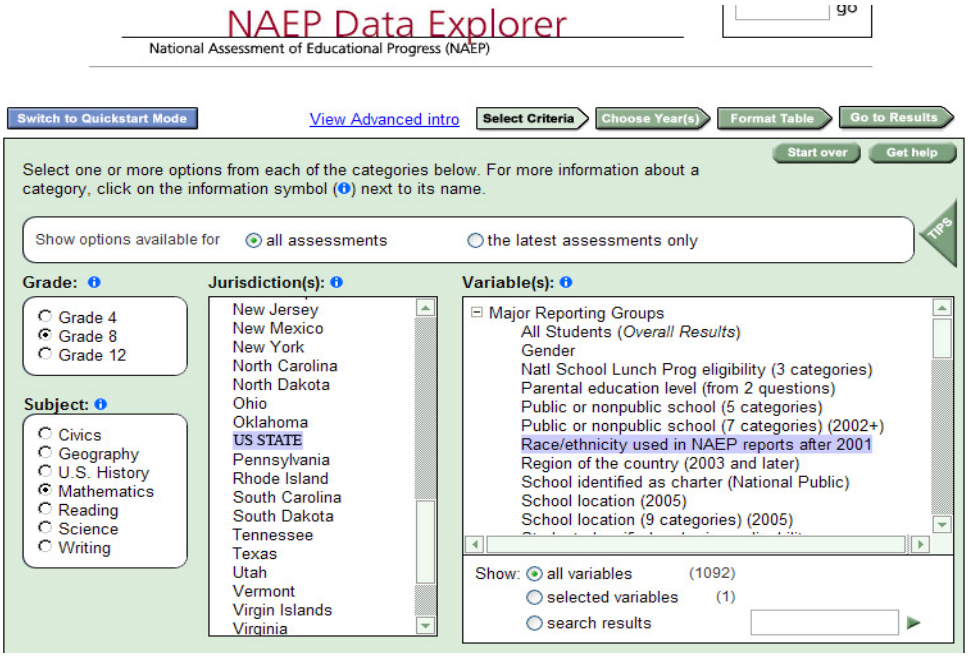


Figure 13. Variable Selections (captured on 12/18/06)

Race/ethnicity used in NAEP reports after 2001	Year	Jurisdictions	Below Basic	Standard Error	At or above Basic	Standard Error	At or above Proficient	Standard Error	At Advanced	Standard Error
White	2005	National Public	21	(0.2)	79	(0.2)	37	(0.3)	7	(0.1)
		US STATE	23	(1.2)	77	(1.2)	38	(1.6)	8	(0.9)
Black	2005	National Public	59	(0.6)	41	(0.6)	8	(0.3)	1	(0.1)
		US STATE	50	(8.1)	50	(8.1)	9	(3.4)	1	(***)
Hispanic	2005	National Public	50	(0.6)	50	(0.6)	13	(0.4)	1	(0.1)
		US STATE	56	(2.7)	44	(2.7)	10	(2.0)	1	(0.9)
Asian Amer/Pacif Isl	2005	National Public	19	(0.8)	81	(0.8)	46	(1.2)	16	(1.0)
		US STATE	18	(4.5)	82	(4.5)	50	(5.8)	19	(4.8)
American Indian	2005	National Public	45	(1.8)	55	(1.8)	14	(1.0)	2	(0.4)
		US STATE	37	(5.8)	63	(5.8)	23	(5.9)	6	(3.4)
Unclassified	2005	National Public	31	(2.7)	69	(2.7)	29	(2.3)	7	(1.2)
		US STATE	‡	(‡)	‡	(‡)	‡	(‡)	‡	(‡)

Figure 14. Results Table (captured on 12/18/06)

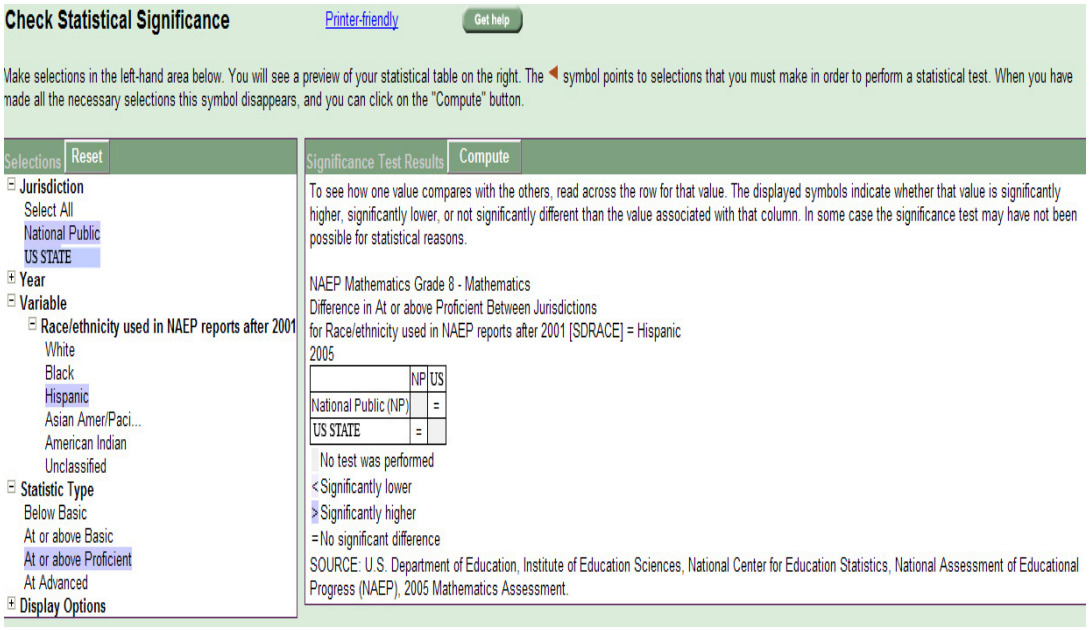


Figure 15. Statistical Significance Test (captured on 12/18/06)

In using the NDE, there are thousands of variables in the NAEP database to select from in the Advanced mode, which are collected from student, teacher, and school questionnaires. Users can select multiple variables at a time and then choose years,

format tables, and/or view results. The NDE is somewhat complicated to use. That is why NAEP state coordinators are offered training on how to use the NDE. A tutorial, which is a slide presentation, is provided for the first-time user, but it does not seem helpful enough. Maybe, it would be better for NCES to develop a more detailed tutorial or a simulation tutorial and then to seek feedback from users. Also, the tutorial needs to be placed at a location to be easily found on the NDE page.

State assessment staff can obtain most of NAEP data needed using the NDE. For example, they can compare their state's overall performance and progress with other states and the nation. They can also find out how selected subgroups of students in their state perform over time and further determine whether achievement gaps have narrowed over time. In addition, the NDE can be used to seek possible correlations of achievement with background factors that might affect student performance. However, users should keep in mind that much of the background information is collected by self-report and that NAEP data are cross-sectional in nature. Learning is cumulative, but NAEP design does not allow for the before-and-after testing since NAEP assesses different samples of students on each testing occasion. Also, the NDE can be useful in states generating their own state NAEP reports because of its ability to produce diverse tables and graphics based on NAEP data. Examples of analyses of NAEP data for US State conducted in the NDE system will be presented later in this section.

NAEP Publications The publications section provides NAEP publications ranging from NAEP guides to technical/methodological reports. These publications are listed by topic, including arts, civics, data products, geography, high school transcript studies,

long-term trends, mathematics, reading, science, technical/methodological, US history, working papers, and writing (http://nces.ed.gov/pubsearch/getpubcats.asp?sid=031, observed on 11/11/2006). Technical and methodological reports provide details on the instrument development, sample design, data collection, and data analysis procedures in specific subject areas tested. These reports are targeted at psychometricians and researchers.

State education personnel might search the publications section to view the latest research studies on issues of interest. These research findings may help inform making educational decisions on curriculum, instruction, programs, or professional development at the state level.

Research e-Center The research e-center provides secondary analysis data access, research publications on NAEP, and funding and training opportunities (http://nces.ed.gov/nationsreportcard/researchcenter/, observed on 11/11/2006). NAEP data for analysis include public-use data via the NAEP website and restricted-use data requiring a license from NCES. Public-use data are available in the NAEP Data Explorer, while Micro-level NAEP data in raw format are available for the purpose of secondary analysis with a license. For researchers and those who seek in-depth NAEP information, it provides links to various papers and reports on special NAEP research topics. These research papers include NCES publications and research publications on NAEP not published by NCES.

This section can be used by the state to acquire restricted-use data licenses for secondary analysis of NAEP data, for example, to link scores from the state's assessment

to NAEP. State assessment staff might participate in NAEP training seminars to improve their understanding of NAEP, especially technical aspects of NAEP and expand their understanding of issues related to NAEP further with research papers provided.

Other Resources Other resources include glossary of terms, NCLB, NAEP assessment schedules, NAEP calendar of events, demonstration booklets, StatChat archives, and Contact Us. The demonstration booklets contain many of the features of the actual test booklets. The NCLB page describes important aspects of the NCLB act related to NAEP. StatChat is a real-time online chat about NAEP findings with NCES staff such as Associate Commissioner, which occurs on the day NAEP results are released.

State education staff can gain an understanding of what features of the NCLB act are relevant to NAEP through the NCLB page. One of the important aspects of NCLB is that states must participate in state NAEP in reading and mathematics at grades 4 and 8 every two years. They might attend the StatChat to ask questions about NAEP results that have been just released. Through the online conversation, state assessment staff might also share their concerns about NAEP from the state perspective with NCES Associate Commissioner. They can ask more specific questions or share suggestions through Contact Us.

The Site Management

The content development and management of the website are performed through cooperation of several NAEP partners. The primary responsibility of the NAEP webmaster is to manage the contractors who do the website development and

maintenance. This section presents a description of the site management conducted in cooperation with the current NAEP contractors.

NAEP Partners Since 1983, NCES has conducted the assessment through a series of contracts, grants, and cooperative agreements with ETS and other contractors. The major activities of the current assessments (2003-2006) have been conducted under the cooperative agreements by a team of contractors. These activities are as follows (http://nces.ed.gov/nationsreportcard/contracts/procurements.asp, observed on 12/8/2006): 1) design, analysis, and reporting; 2) item development; 3) materials preparation, distribution, and scoring; 3) sampling and data collection; and 4) web operations and maintenance.

Site Content & Management The content is developed by NAEP Assessment Division Staff and various contractors, according to the NAEP webmaster. For example, the process of creating content for NAEP results is: 1) assessment results are analyzed and then presented in both text and graphical representation; and 2) this is provided in hardcopy and web pages. According to the NAEP person, the content development is managed through an automated content management system. He added that Web page review is performed by contractors, Assessment Division staff, Center-wide staff, and external reviewers. After then, a QC (Quality Control) process is completed. The pages are then staged on a development server where a technical review is performed and then placed on the production server for public access.

According to the webmaster, the site content is monitored on a daily basis by contractor staff and NAEP Webmaster. The review of comments received through

Contact Us is also made daily. NCES identifies information needs and preferences of the site users through questions from Contact Us, State Assessment Coordinators, contractor staff, online polls, and NAEP staff requests. In addition, he responded that NCES attempts to review the usefulness of the site through usability studies/groups and trend analysis of logs semi-annually in a formal manner, which is a really on-going effort. Yet, any product of these efforts is not provided on the website.

User Feedback & Requests The website receives feedback or requests from the site users through an automated Public Communications Tracking System that interfaces with Contact Us, according to the NAEP webmaster. He remarked that NCES monitors the user feedback daily and attempts to respond within 24 hours. A request received is forwarded to appropriate staff and monitored, and then response is sent via e-mail when completed. He added that user feedback is maintained in an enhancement database for possible future modifications of the content provision process. According to the project officer for NAEP state coordinators, NCES also conducts informal polls through the NAEP website to seek feedback on the site. Yet, none of the state education staff interviewed commented on their use of this system through the NAEP website. More will be discussed in a later section.

Potential Utility of Information on the NAEP website

An analysis of the NAEP website from a state perspective revealed that NAEP data provided on the site could be utilized in several ways to inform state policy decisions. This section presents a summary of potential utility of NAEP information placed on the NAEP website. This usefulness was pursued on the basis of the content analysis of the

site and NAEP data analysis with the NDE. Illustrated below is how the information available on the site can be used in making educational and/or policy decisions at the state level along with examples of an analysis of relevant NAEP data when appropriate. NAEP results and resources can be used in a state context as follows:

- Development or revision of state standards
- Examination of trends in student performance
- Examination of whether achievement gaps is narrowing
- State-to-state comparisons
- Comparison of state test results and NAEP results
- Development of state performance standards
- Identification of content strengths and weaknesses
- Examination of relationships between achievement and background factors

Development or Revision of State Standards NAEP frameworks reflect a careful national consensus on what 4th-, 8th-, and 12th-grade students should know and be able to do. The frameworks are designed to reflect a balance among the emphases suggested by current instructional efforts, curriculum reform, contemporary research, and desirable levels of achievement (Pellegrino et al., 1999). The frameworks tend to be inclusive since they consist of content that is considered to be significant by all states. The frameworks are also forward looking since they attempt to balance what is being taught among the states with expert judgment about what should be taught (Wise, 2003). In short, NAEP frameworks represent the best thinking of leading experts in the field from around the country and thus could be used as a reliable independent check on state standards.

In fact, many states have developed or revised their state curriculum standards on the basis of NAEP frameworks. A high degree of correspondence between state standards and the NAEP frameworks provides several advantages. First, a close alignment between the two may add credibility to the state standards, as noted previously. In addition, the tight alignment might enable comparisons of student performance on NAEP and state assessments. When the two assessments provide consistent information on achievement, NAEP results can be viewed as external validation of state test results. Yet, when the two results are discrepant, an in-depth analysis is necessary. Lastly, students are expected to perform well on NAEP, which is important since NAEP is the only measure by which interstate comparisons are made against high-standards.

Trends in Student Performance NAEP results can be used to probe the overall performance trends at each tested grade level for a specific subject over time. Short-term trend data provided by NAEP might help policymakers and educators find out how their students are making progress over time. As an example, mathematics data for US State and the nation (public schools only) were analyzed using the NDE for grade 4 in 1996, 2000, 2003, and 2005 and for grade 8 in 1990, 1996, 2000, 2003, and 2005 in terms of scale scores, achievement-level results, and percentiles. Statistical tests were conducted with the NDE to identify whether the performance difference shown between 2005 and each of other assessment years was statistically significant.

The US State's students have shown steady improvements since 1990 in terms of average scale scores and achievement-level results. In terms of scale scores, average scores were significantly higher in 2005 at grade 4 than in 1996 and 2000, but the actual

gains since then were not statistically significant. Grade 8 scores in 2005 were significantly higher than those in 1990 and 1996, but the gains since then were not significant. In comparison with national results, 8th-graders performed better than national counterparts, while the performance of 4th-grade students was not significantly different from that of national peers. In general, achievement-level results reflected the pattern of the trends in the average scale scores. For example, the percentages of students at grade 4 performing both at or above *Basic* and at or above *Proficient* significantly increased over time except for those between 2003 and 2005, while those of grade 8 students at or above *Basic* and at or above *Proficient* were significantly greater in 2005 than in 1990 and 1996.

In addition, an analysis of scale scores by percentile enabled identifying whether trends in average scale scores were reflected in all levels of performance (10th, 25th, 50th, 75th, and 90th) (Figures 3 & 4). The analysis shows that the improvement made at grade 4 as a whole was contributed by students at all the percentiles, while overall grade 8 gains were made by middle-scoring and higher-scoring students. Changes in gaps between the 75th and 25th percentile showed that 4th-graders' gaps in 2005 narrowed from 1996 and 2000 in statistically meaningful ways, while 8th grade gaps did not narrow over time. These findings indicate that the gap between higher-achieving students and lower-achievers does not decrease and is likely to widen as students move from grade 4 to grade 8.

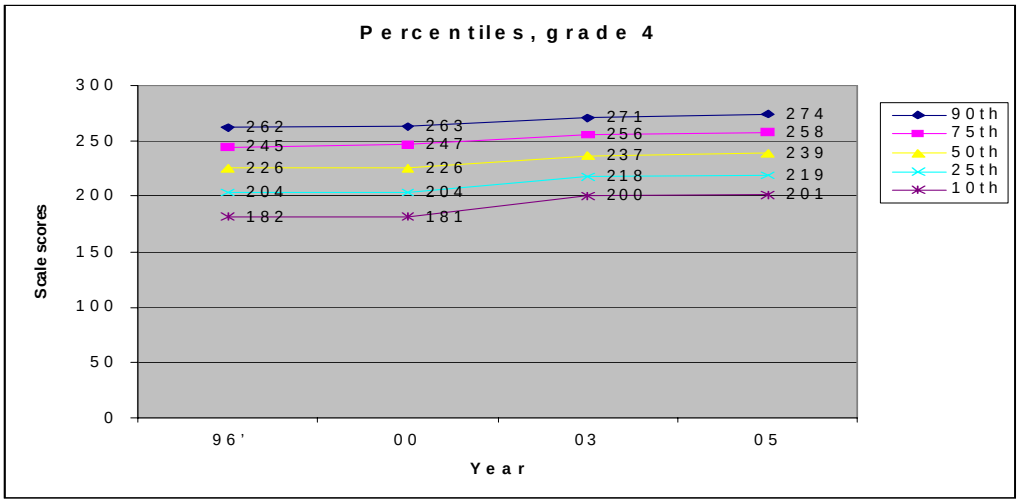


Figure 16. US State mathematics scale score percentiles, grade 4: 1996-2005

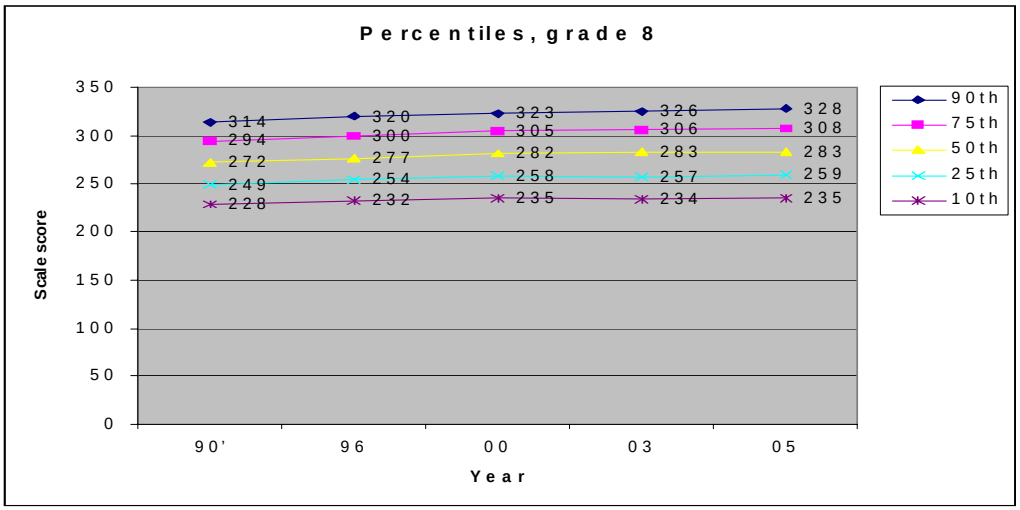


Figure 17. US State mathematics scale score percentiles, grade 8: 1990-2005

Closing the Achievement Gap The intent of the NCLB is to hold public schools accountable for closing the achievement gaps between different groups of students. NAEP is asked to contribute to this end by providing an accurate measure of the current levels of students' achievement and their progress over time (NAGB, 2003). In this context, NAEP results can be used to understand how achievement gaps among different

demographic groups have changed over time. Social class characteristics such as income and race/ethnicity in the US have actually influenced student achievement (Rothsteln, 2006). Accordingly, states have focused on reform efforts to narrow the achievement gap between lower-class and middle-class children. This section presents an analysis of the patterns of achievement gaps by subgroup (e.g., SES, race/ethnicity) conducted using NAEP mathematics data for grades 4 and 8 in US State.

SES

Socioeconomic status (SES) has been found to be a strong predictor of student achievement. NAEP collects information on indicators of students' SES including the highest education level completed by students' parents and eligibility for the free/reduced-price lunch program. Here, achievement gaps between students by SES were explored by comparing results in terms of eligibility for the free/reduced-price lunch program. Three levels of eligibility were examined: eligible, not eligible, and information not available.

At both grades 4 and 8 in US State, average scale scores for students who are not eligible for the free/reduced price lunch program have been significantly higher than those for students eligible for the program since NAEP collected these data in 1996 (Figures 5 and 6). At grade 4, the performance of students from both groups improved over time. Regarding changes in gaps between the two groups, the 2005 gap narrowed from gaps in 1996 and 2000, but the apparent change between 2005 and 2003 was not statistically significant. At grade 8, students for both groups did not make statistically

significant progress since 2000, and the gap between the two did not narrow over time. These changes in gaps reflect the pattern of the trends by percentile described previously.

The results indicate that there is a clear association at both grades 4 and 8 between SES and average scale scores on the mathematics assessment. In addition, score gaps between the two groups seemed to narrow over time at grade 4, but did not decrease at grade 8. These findings might provide state policymakers with indications about whether programs aimed at low-income students (e.g., early childhood preparation, after-school programs, or summer programs) are working or needed. For example, if programs targeted at low SES students have been operated in the state, it is possible that such programs are effective for grade 4 students, but not for 8th-graders.

In fact, document analysis revealed that the US State's assessment program began to provide performance data on students eligible for the lunch program in 2005. The 2005 US State's annual report card states that 48 percent of these students met the state mathematics standards at grade 8 in 2005, while 64 % of students not eligible for the program reached the standards. No trends in the performance of these disadvantaged students can be identified from the state assessment data, while NAEP data seem to be useful for the SEA to obtain some insight into how low SES students in the state are making progress over time.

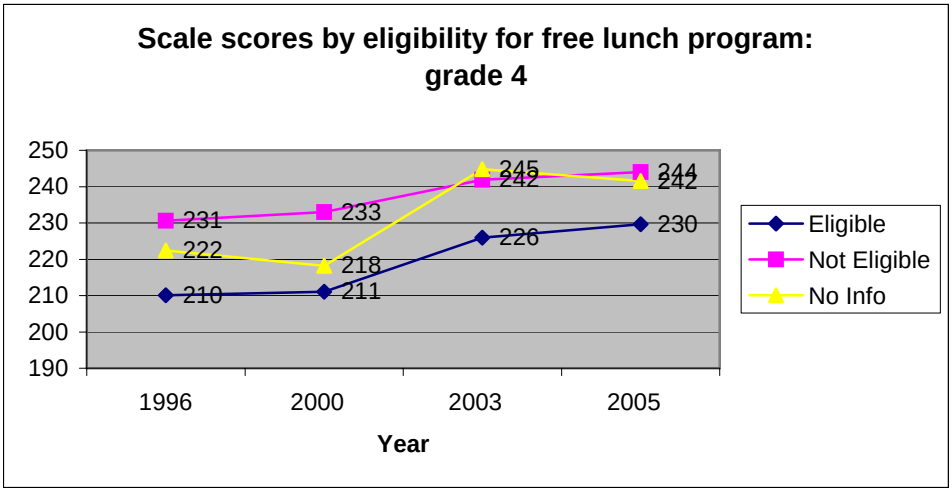


Figure 18. NAEP mathematics scores by National Lunch Program Eligibility: Grade 4

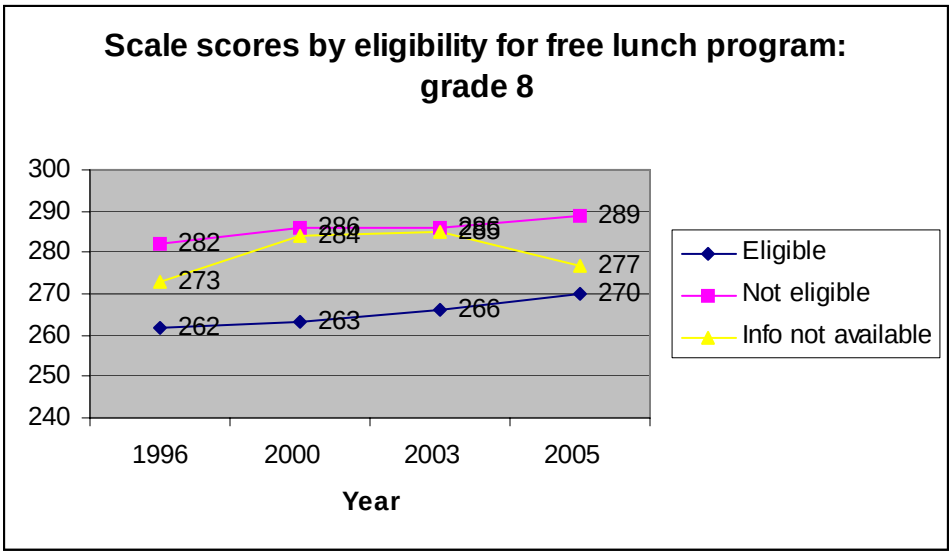


Figure 19. NAEP mathematics scores by National Lunch Program Eligibility: Grade 8

Race/ethnicity

In general, the size of the scale score gaps between the racial/ethnic subgroups has been found to be much larger than that among other subgroups. In this section, analysis was focused on major racial/ethnic groups: White, Hispanic, and Black. An analysis of

the 2005 mathematics results for the US State's students by race/ethnicity is presented in terms of scale scores and achievement-level results, as well as change in the gap over time by scale scores.

At grade 4, White and Hispanic made improvement in terms of scale scores and achievement-level results since 1996 (White: through 2003, Hispanic: through 2000), but no improvement for Black students. At grade 8, White students made progress in earlier assessment years, but no progress for Hispanic and Black students (Figures 7 & 8). The analysis shows that substantial performance differences exist among race/ethnicity subgroups at both grades 4 and 8. In addition, any change in score gaps between either White and Black students or White and Hispanic students was not statistically significant for both grades 4 and 8. These results indicate that performance gaps between White and minority students in US State have not decreased over time. Morover, improvement made over time as a whole does not seem to involve all ethnic/racial students making improvements. For example, meaningful 4th-grade gains were made mostly by White and Hiapanic students (plus Asian), while 8th-grade progress was made by White students (plus Asian).

NAEP and the US State data seemed to indicate somewhat inconsistent results regarding the change in achievement-level gaps for White-Black and White-Hispanic students in grade 8 mathematics. NAEP data indicated that the White-Hispanic gaps widened from 1990 through 2005 in almost every case and that the White-Black gaps widened as well. However, the state data indicated that the gaps between White & Black and White & Hispanic students decreased very gradually over time (the state assessment did not test grade 4 students until 2005). The discrepancy might be due to the fact that the

state's performance standards are different from NAEP's in terms of standard-setting methods.

By examining NAEP data for all ethnicities in addition to the state data, state education personnel can obtain a more comprehensive picture of how their minority students are doing in relation to their White peers and of whether or not the achievement gaps are decreasing over time. Also, this kind of information could provide some insight into whether or not the state's initiative of "closing the achievement gap"is making a difference. Further, the SEA might compare the change in achievement gaps over time in US State to that in like-states such as State A, using state NAEP data.

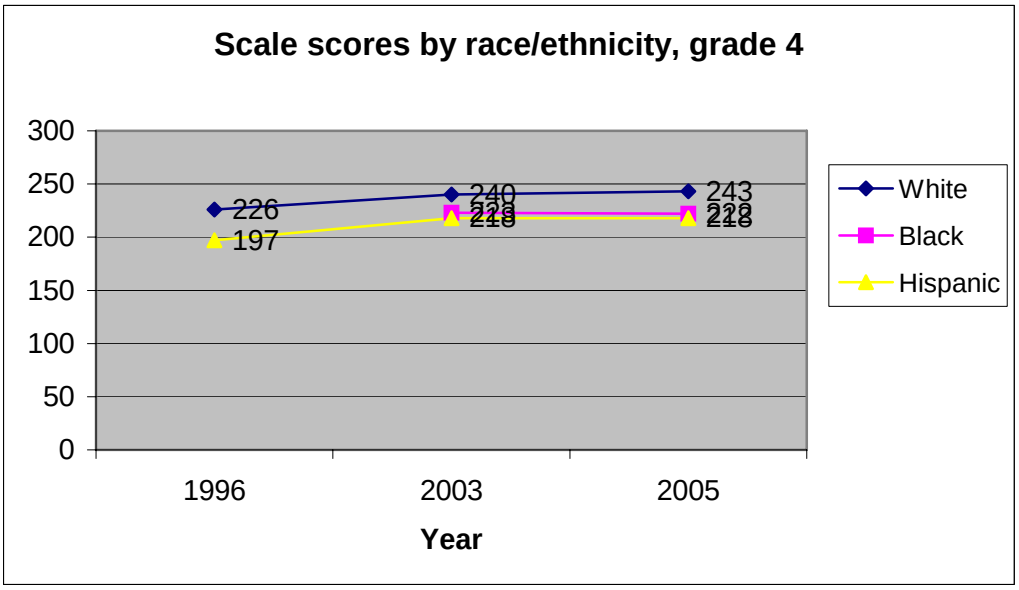


Figure 20. Average NAEP mathematics scores by race/ethnicity: Grade 4

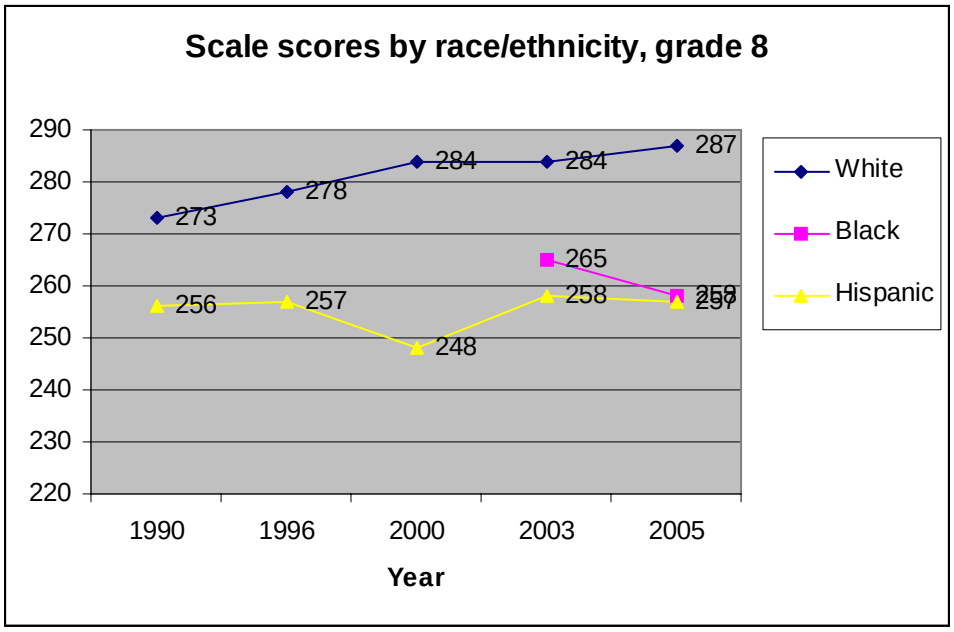


Figure 21. Average NAEP mathematics scores by race/ethnicity: Grade 8

State-to-State Comparisons NAEP results might be used to see how students are performing over time compared to other states. Demographic characteristics are a major factor that affects student achievement, but overall averages provided by NAEP do not take into account the different demographics of states' student populations in reporting. Thus, state comparisons need to productively focus on state differences in policy-relevant correlates of proficiency, rather than simply on state differences in mean proficiency (Raudenbush et al., 1999).

In this sense, it seems more meaningful to compare states with similar demographic characteristics. Comparing states with similar demographics may be useful in identifying whether there might exist such factors as state policies and characteristics that might affect differences in student achievement between states. US State is similar to its peer state, State A, in terms of demographics and money spent on education, according to the

state education data profiles provided on the NAEP website. Thus, this section illustrates how the US State's students at grades 4 and 8 are performing in mathematics in terms of scale scores compared to State A's counterparts.

Overall, there was no significant difference in performance between the two states over the past decade (since 1996) at both grades 4 and 8, except that 4th-graders in State A performed better in 2005 than their peers in US State. Comparison of grade 8 mathematics scale scores is presented in Figure 9. These results might suggest that the US State's key educational policy variables including per-pupil expenditure, teacher characteristics, and resource utilization are very similar to those of its peer state. Further analysis using a sophisticated research model is needed to test these hypotheses.

In addition, it must be noted that even states with different demographics can be compared by controlling for social and economic factors. In order to relate policy-relevant indicators to student outcomes, strong statistical methods must be employed using raw data since there are still significant barriers to analyzing substantially complicated NAEP data and obtaining reliable results policymakers need. One of the US State's primary needs is to support instructional program improvement efforts, and therefore it could be useful to examine policies and instructional practices that may guide improvement in achievement in other states with major improvements.

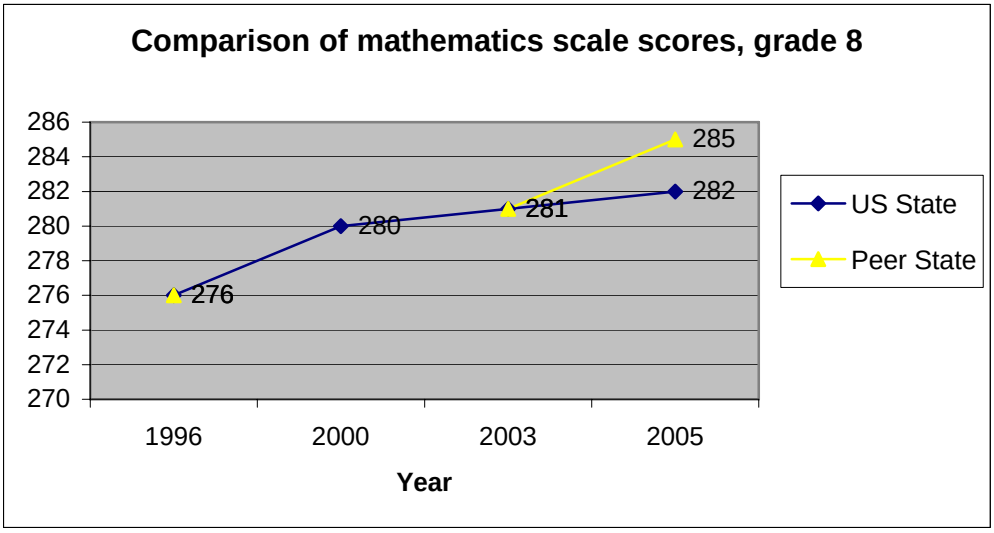


Figure 22. NAEP mathematics scale score comparisons at grade 8

Comparison of State Test Results and NAEP Results Do NAEP and the US State assessments provide consistent information on the performance of students in the state? NAGB studies (2002b) suggested that NAEP could be used as evidence to confirm the general trend of state assessment results in grades 4 and 8 reading and mathematics in terms of percentages of students meeting standards, but that the interpretation should be made with great caution because of the potential differences between the two tests.

This section attempts to examine if NAEP results are generally in the same direction as the US State's assessment results in terms of meeting standards. The NAEP proficient level is identified by NAGB as the level that all students should reach, while US State considers students performing at the "meets" or "exceeds" level to have met the standards. Yet, there is no clear correspondence between the NAEP Proficient level and the state's Meets level. In this study, both the NAEP's Basic and Proficient categories were used for an analysis of overall performance in 8th-grade mathematics, while only the NAEP Proficient level was used for an analysis of trends in achievement gaps. These

analyses were preliminary and adapted from those conducted in Linn et al.'s study (2002).

For overall performance, NAEP and state test results indicated the similar pattern, although the degree of performance gains was different (Figure 10). In addition, this analysis suggests that the US State's *Meets* standard is closer to the NAEP *Basic* level in terms of meeting standards than the Proficient level. As noted previously, however, there was some inconsistency for changes in achievement gaps between white and minority students. It is not clear whether this inconsistency for ethnic subgroups indicates a direct contradiction between the two test trends.

An examination of the sources of the inconsistencies of the two test results is beyond the scope of this study. However, it seems reasonable to speculate that these discrepancies might result from the differences between the two tests in their performance standards. In addition, their definitions of achievement level categories might have affected resultant estimation of the percentage of students at each level, but could not be compared because no descriptions had been established for US State. In addition, NAEP's small samples might not allow the confirmation concerning performance gaps.

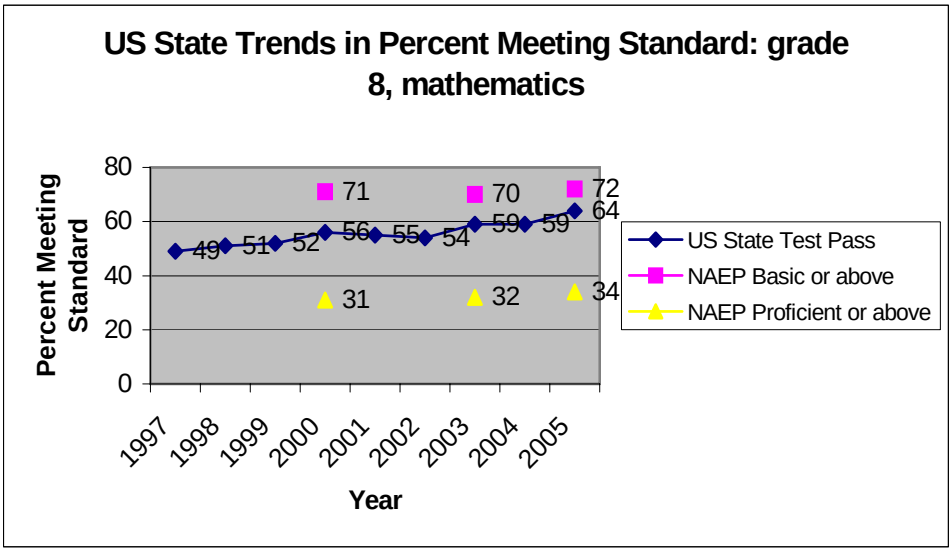


Figure 23. US State's trends in 8th- grade mathematics performance on state tests and on NAEP

Performance Standards There are two major approaches to providing meaning to test scores: norm-referenced and criterion-referenced interpretations. A norm-referenced interpretation is obtained by comparing student scores to a well-defined comparison group (e.g., a nationally representative sample), while a criterion-referenced interpretation is obtained by comparing student scores to a well-defined standard (Phillips et. al., 1993). Because criterion-referenced interpretations are grounded in specific content objectives, they are more instructionally relevant. These interpretations provide more meaning to the arbitrary summary scale scores and make assessment results more understandable to the general public, educators, and policymakers.

NAGB has addressed the level of mastery of each subject that constitutes proficiency since 1990, in addition to the content frameworks. Three levels are defined by minimum or cutoff scores: Basic, Proficient, and Advanced. With these achievement levels, student performance is reported in terms of percentages at or above a given level. Increases in the

percentage of students who are at or above the proficient level are considered to be more meaningful for policy makers and the general public than an increase in the average score on the arbitrary scale.

These NAEP achievement levels might provide a useful benchmark for state efforts to define their own performance standards. A study by NAGB (2000) indicates that 25 states have developed achievement level descriptions similar to those used by NAEP. If the two scales are aligned, the percentages of students at different levels on statewide tests may be compared with corresponding percentages from state NAEP. In addition, NAEP performance standards could have a potential impact on curriculum development and instruction. For example, descriptions of student knowledge and skills at each of the NAEP achievement levels might help curriculum developers at state and local levels focus their curriculum and instruction on areas judged to be the most critical to proficient performance.

Sample Assessment Questions NAEP has been innovative in the design of test items. NAEP items are developed in a variety of formats including performance assessments and constructed response questions. These NAEP questions reflect current thinking on how to assess accurately the knowledge and skills described in the NAEP frameworks. Fortunately, NCES has provided some portion of each assessment in the NAEP Questions Tool on the NAEP website. A list of questions is available for each subject and grade level with the Questions Tool, along with scoring guides and student responses.

These released items might be useful in guiding item development for state assessment measures. For example, the use of performance assessments and constructed

response questions in NAEP could lead to the inclusion of similarly formatted questions in the state assessments. Additionally, state education personnel might examine what misconceptions students hold within a specific subject by exploring student responses especially to extended constructed-response items and further use such information for decision-making on curricular or instructional policy.

These NAEP items could be also useful to classroom teachers. Teachers may conduct formative assessments using these items. For instance, the information provided for each question can be used for teachers to score responses, assess the types of questions that their students can or cannot answer well, identify student misconceptions, and compare classroom results to national outcomes. It might also have the potential to provide some basic guidance to teachers about areas of strengths and weaknesses at the state and national levels. Additionally, the information might be used in designing their instruction and classroom assessments to cover content areas described in the NAEP frameworks.

Relationships between Achievement and Background Factors NAEP has tracked changes in contextual and instructional factors related to student achievement and in the distribution of important educational resources, by collecting background information from students, teachers, and schools. The information collected includes instructional content, instructional practices and experiences, teacher characteristics, school conditions and context, and conditions beyond school. Contextual factors correlated with performance can be valuable and point to possible explanations for poor or promising performance. However, it is important to note that NAEP results cannot be interpreted in

a causal sense because NAEP data are cross-sectional in nature and because a number of factors including unmeasured variables might affect student achievement simultaneously.

This section focuses on the relationship between the emphasis placed on different content areas and performance for 8th-grade students in US State in the 1996 NAEP mathematics assessment. Five content areas assessed were: 1) algebra & functions; 2) data analysis, statistics, & probability; 3) geometry; 4) measurement; and 5) numbers & operations. The data were based on teachers' written self-reports using a four-category response scale: "a lot", "some," "a little," and "none" emphasis. Care must be taken in interpreting the findings since teachers might have interpreted the same category differently. For example, one teacher's reading of "a lot" might be another teacher's "some."

As predicted, the analysis indicates that emphasis on algebra was related to the highest student performance. Emphasis on data analysis, statistics, and probability had a positive impact on achievement as well. Performance on geometry showed mixed results, while there was no indication of patterns found in both the measurement and numbers & operations. These results indicate that high emphasis on such content areas as algebra & functions and data analysis, statistics, & probability in US State might affect student achievement in a positive manner. In fact, current mathematics reform efforts stress that more focus needs to be given to these content areas (algebra and probability & statistics) away from emphasizing only the number & operations strand.

The possible relationships identified in this analysis might help guide the state's efforts toward improvement of teaching and curriculum. In other words, emphasis placed on algebra and data analysis & statistics in curriculum could lead to the improvement of

student performance in mathematics. In particular, this kind of information may be helpful for US State to think about policy implications since its own assessment program collects relatively limited background information on students and school practices. Moreover, the state collects background and contextual information only from students and has not developed a web-based tool such as NDE for the state assessment. Clearly, these data NAEP collects provide the opportunity for state education personnel in the state to explore potential associations between achievement and background factors.

Summary

An analysis of the NAEP website revealed that NAEP data and resources can be utilized at the state level for: 1) examination of trends in student performance; 2) examination of the achievement gap change over time; 3) state-to-state comparisons; 4) comparison of state test results and NAEP results; and 5) examination of possible relationships between background factors and achievement. In addition, it was found that NAEP performance standards could be used to help develop state performance-level descriptors and that released NAEP items and student responses can be used to develop items for state assessment measures and to examine areas of strengths and weaknesses.

Perspectives of state education personnel in US State on the usefulness of NAEP are presented in the next section. Their actual use of NAEP data and their use of the NAEP website are then illustrated in the following sections.

State Education Personnel's Perceptions of the Usefulness of NAEP

Participants' perceptions of several features of the NAEP program in terms of their usefulness were identified through an analysis of their responses to interview questions. Questions asking about their perspectives on several aspects of NAEP were developed based primarily on the literature review on the utility of NAEP. Interviewees' responses to different questions (perhaps indirectly related) were sometimes overlapping. Hence, their responses were organized and presented around five categories: 1) policy relevance of NAEP data; 2) state-to-state comparisons; 3) NAEP under NCLB; 4) NAEP performance levels; and 5) improvement of NAEP (integration of NAEP results with other datasets, higher-order thinking skills, NAEP inclusion, and use of web's features).

These aspects of NAEP were focused on since most of them were addressed as major issues associated with the NAEP program in the literature. For example, NAEP collects background and contextual information thought to influence academic performance. This information is intended to provide an interpretive context for NAEP results in support of an in-depth understanding of student achievement, which leads to improvement of its policy relevance. Accordingly, questions were asked to examine how state education staff actually perceive NAEP data in terms of their policy implications. Secondly, State NAEP was designed primarily to allow states to compare the performance of their students with that of students in other states. How do state education personnel perceive the usefulness of state comparisons? What concerns do they have about such comparisons?

NCLB encourages states to verify state gains with trends on NAEP, although the specifics are not indicated in the legislation. How do state education personnel view

NAEP's role as confirmation? It seems informative to probe what they think about this issue. Fourthly, NAEP performance levels are intended to be useful in communicating NAEP results, and many states have taken NAEP's lead in reporting student achievement by performance levels. Nonetheless, there has been a controversy around the achievement-level setting method and the results. Over the last decade NAEP has faced considerable criticisms that the achievement levels are flawed. For example, the most recent NAEP evaluation criticized that NAEP item pools do not match well the types of knowledge and skills portrayed by NAEP achievement-level descriptions (Pellegrino et al., 1999). Thus, there was a need to ask questions about the usefulness of the NAEP achievement levels.

Lastly, some of the issues surrounding the improvement of NAEP addressed by pervious NAEP evaluations were asked, including integration of NAEP results with other datasets, reporting divers aspects of achievement, NAEP inclusion of students with disabilities and English language learners, and use of web's features. This section describes how staff members in different positions at the SEA perceive each of these aspects of the NAEP program.

Prior to illustrations of their perceptions of NAEP, a description of what purposes participants expect assessments to serve in general is presented. Quotations are presented to support the arguments made by participants. Alphabets ranging A from G were used to indicate the seven participants interviewed to protect their confidentiality.

Assessment Purposes and the Appropriate Use of NAEP Data

An analysis of responses to a question regarding assessment purposes elicited seven different types of purposes that participants expect assessments to serve. Most participants responded that assessments serve multiple purposes. The assessment purposes identified include:

- describing student performance against academic standards
- informing instructional decisions
- providing information on student performance to students, teachers, parents, and the public
- holding school districts and schools accountable for student performance
- providing information on strengths and weaknesses of programs
- certifying students as having attained specified levels of accomplishment
- describing the status of the education system

There were no distinct differences among participants in opinions on the purposes of assessments depending on their positions. Most respondents contended that assessments should measure student performance against academic standards. It is speculated that their support of the use of assessment for this purpose is closely associated with recent standards-based reform efforts. In particular, reading and science specialists agreed that assessments must be aligned with standards so that they can measure how students make progress on meeting prescribed standards. A reading specialist spoke of this use in this way:

> Now they [assessments] are used as individual student measurement to make sure students are making progress on meeting standards so they got responsibility I guess [for] measuring the student progress in terms of meeting state performance and academic standards. (A-Interview)

Not surprisingly, a chief policy officer, who is responsible for establishing the accountability system for the state, viewed accountability as a major purpose of assessments. That is, assessments should be used as a tool for holding school districts and schools accountable for student performance. The policymaker argued that assessments should provide information on the performance of schools and school districts in terms of meeting academic goals for students so that assessment results can be used for accountability purposes. He emphasized that having high-quality assessment data is the key to making it possible for the state education agency to make decisions about consequences in a fair and objective manner. In addition, he briefly mentioned the relationship between school funding and assessment information. There is a general feeling that the public are reluctant to support any tax increases, regardless of the advantages and thus states have focused their attention on improving student achievement and on adding the necessary assessment measures to identify the effect of increased educational expenditures (Whitney, 1993). It is also voters' right as citizens to demand increased accountability for how their tax dollars are being spent to improve education. In this sense, he noted that it is important to provide information on student performance to the public since tax dollars spent on education should be accompanied by greater accountability and improved results. He commented on the assessment's role in this way:

> We have responsibility for determining whether or not schools and school districts are achieving the academic goals for [US State] students that have been established by the legislature and State Board of Education. … Only about 20% of registered voters actually have children in school [and] the vast majority of the registered voters do not have any direct connection to the school system. So having information that is public, because particularly for financial reasons people have to vote on things like taxes [and] pay for schools, they want to know our [*sic*] my [*sic*] tax dollars are going to an effective organization. One way of answering that question is to have assessments of student performance. (E-Interview)

Subject-area specialists and the chief policy officer also indicated the use of assessment in instructional decision-making. They stressed that assessments should provide classroom teachers with useful information on how their students are performing, which is essential in making instructional decisions. A science specialist maintained that information on student performance also could help state education personnel inform instruction and policy decisions about programs. A reading specialist also emphasized that assessment results must inform instruction:

> I also think that to some extent performance on state assessments should inform instruction for teachers now [about] not only what to teach but [also] what to concentrate on. (A-Interview)

The state NAEP coordinator asserted that assessment should be part of a bigger picture of indicators of what students know and can do. She indicated that assessments should provide information about the quality of the education system currently in place, that is, the overall status of the education system. Her remark implied that even the best assessments are imperfect measurement tools with limits on their generalizability and appropriate use (McDonnell, 1994). Thus, her point was that multiple indicators for the

quality of teaching and learning should be sought to improve the operation of the education system as well as student learning. Also, it is important to note that the use of multiple indicators increases the validity of inferences based on observed gains in achievement. She made her point in this way:

> They [assessments] only give a glimpse on what might be going on at some point for particular students. … Assessment also needs to be part of a bigger picture of indicators of what students are capable of in knowledge and skills. (C-Interview)

A mathematics specialist argued that the main purpose of assessments is to provide information about strengths and weaknesses of programs, while the reading specialist commented that assessments used to mostly look at the effectiveness of programs but have expanded to serve a variety of purposes today. There does not appear to exist a conflict between the two's views on expectations about assessments. In general, large-scale assessments serve multiple purposes today. Kiefer (2001) labeled three general purposes for large-scale assessments: measuring achievement, providing accountability information, and improving instruction. He argued that a particular assessment might have one or more of those purposes and that the emphasis could be variously weighted across purposes.

In particular, the mathematics specialist commented on important aspects of assessments, which is considered to be worth paying attention to. She pointed out that one of important aspects of an assessment is a clear communication of what it is designed to measure and not to measure. She argued that assessment information can be

inappropriately used without an understanding of the assessment itself. She argued the important aspects of assessments in this way:

> Assessments are very complex. So one important component of assessments is having information about the assessment: how it was put together, how it is designed to measure, and what limitations are for using particular data points that come out. I found most assessments don't provide that information. Because of that, people misuse [assessment] information on a regular basis. (B-Interview)

Her argument has some implications for NAEP data use since NAEP results are interpreted by some users beyond the data and the design used to generate them especially in attempts to explain poor performance (Pellegrino, 1999). For example, when student performance was disappointing in California, policymakers and newspaper reporters argued that poor performance had resulted from several reasons including overcrowded classrooms and the state's whole-language reading curriculum, which were not the variables collected by NAEP (Glaser & Linn, 1996). Also, users might misinterpret suggested relationships between achievement and background factors in a causal sense because of their lack of an understanding that NAEP design does not support causal relationships. Clearly, these misinterpretations are assumed to be due to their lack of knowledge about what NAEP measures, its purposes, and its uses.

Policy Relevance of NAEP data

NAEP collects a variety of demographic, background, and contextual information on students, teachers, and schools. Accordingly, NAEP presents disaggregated results by gender, race/ethnicity, SES, and additional variables, including language spoken in the

home, study and homework habits, instructional practice, teacher preparation, course-taking patterns, and technology use. These data are intended to provide an interpretive context for NAEP data. NAEP is a unique source of background information on factors that relate to student achievement, which might help inform educational policy (National Research Council, 2000).

Regarding a question asking how policy-relevant participants find these disaggregated data, most respondents indicated that basically those data might be policy-relevant especially in terms of closing the achievement gap in some way. However, some claimed that they are not as useful as state assessment data because of the nature of the NAEP data based on the representative sample of these subpopulations. For example:

> Well it helps states to determine how students in various disaggregating groups are doing. In addition, we would use state assessment data. I mean probably we would use state assessment data and local data more to determine that than we would with NAEP data. … the way the NAEP test is given, it's not given you know to every student every year like our state and local assessments would be. So I don't know we would use it as much as that. … probably we would get a little bit more detail, but usually it just confirms what we already know. (D-Interview)

Some participants argued that these kinds of NAEP data are not very helpful in making decisions at the state level because of the small sample size for subgroups. They stressed that in particular, the small sample size of the subgroup populations does not produce reliable estimates for these groups with NAEP. In fact, although private schools and schools with high-minority students are deliberately sampled at a higher rate to assure a sufficient sample size for increased precision of estimates (Johnson, 1989, 1992),

sometimes the sample size is still so small that it does not provide meaningful results for these subpopulations. Two respondents made their points as follows:

> Closing the achievement gap is an agency's priority, and it's an important policy. Having information about how students with different backgrounds perform in schools is very important. But the difficulty with NAEP data is [that] it's such a small pool. …when we have many schools where groups that fall in these demographic categories are so, the size is so small they do not report it at all. So it's not very helpful to most decisions that we make. However, it's an important view and if perhaps they increased the sample size, that might be more useful to us. (B-Interview)

> Well we make a lot of use of disaggregated data, and it's hard with NAEP because of the relatively small sample size. … the problem is [that] if you try to assess that using random sample methodology, you don't always get a sufficient sample size for an adequate analysis of this disaggregated data. (E-Interview)

It needs to be pointed out that the interviewee E above made some incorrect statement regarding NAEP sampling methodology and that he actually made the same argument concerning NAEP sampling several times during interviews. To sample students representative of the entire population and subgroups of this population, a complex sampling scheme such as a stratified multistage probability sampling design, rather than simple random sampling, is used to collect NAEP data. For example, in state NAEP, schools are first stratified based on urbanization, size, area income, and percent minority and then schools are selected at random within strata. Finally, the students within a school chosen are randomly sampled from a list of students within a grade.

An education program specialist insisted that she does not consider these disaggregated data to be policy-related because they are not provided at the local level and because they are simply statistical estimates. For example, she contended that NAEP does not provide specific information for districts and schools to use in making

educational decisions, which are clearly linked to curriculum and instruction. At the same time, she maintained that these data are not useful because they are merely statistical estimates of student performance. She made such arguments in this way:

> I don't find them particularly policy-related at all because they don't deal with specific data representative of individual schools or individual districts so they can be used at the local level. … It's a representative sample, so there is no way to use it to show how particular schools are doing because not all of children in any one school or even in one district are assessed. It's an estimate of what students know and can do on the NAEP assessment ... if you rank states using NAEP data, you find there's a huge group within the middle that has no statistically significant differences between their results. So it really does not tell us a lot…. You cannot make policy decisions on something that is only presented at the state level. (G-Interview)

Furthermore, two participants meaningfully responded to a follow-up question asking about how NAEP could provide better policy implications of NAEP results. The mathematics specialist suggested that it would be helpful from a state perspective if all of the NAEP items administered could be released in conjunction with a comprehensive analysis of all the student performance. She added that it would be useful to have an in-depth analysis of how students performed from student responses. For example, one of important aspects of assessments in doing her job, according to her, is the release of items and analysis of student responses so that she can conduct some analysis on where programs may or may not be successful in developing misconceptions or whether or not the assessment items are working well to help shed light on what students know and can do. She argued that the interpretive look at NAEP results from students' responses would provide more useful and policy-relevant information.

According to the NAEP website, only a small portion of each NAEP assessment is released since some questions must be kept secure for use in future NAEP assessments. Consultation with a project officer for design, analysis, and reporting at NCES confirmed that NAEP keeps approximately two-thirds of the items in each assessment secure for use in the next assessment in a subject since it is very expensive to develop new test questions. Yet, the mathematics specialist asserted that all test items should be released to fully understand the level of content knowledge assessed and student responses at each grade level, although she understands the resultant economic concern. She also stressed that increasing a sample size for subgroups would be helpful in reporting stable estimates of performance for relevant subpopulations. She made this point strongly as follows:

> There's no need to use the questions again. So you know … it's more expensive… If the bottom line is what's the best test that we can do [and] the best data we can release for the least amount of cost, then what they are saying is you know that cost factor. I understand the economic concern, but that cost factor is the determining factor for the quality of the test, then they have to decide how important education is. And if they can justify that well you know we just release a little bit [of] information [and that] we want to save money, [it is] fine. Save money, [but] stop doing it because it's not useful the way the data is [*sic*] not useful the way as it is. …So if it's just too expensive to do it well, stop doing it altogether. If you are going to continue doing it, do it correctly, [and] do it well so it can be used [and] so it is meaningful. But you know in this some middle of the road it's not helping anyone, it's confusing the public, it's confusing policy makers, [and] it's confusing students and parents. (B-Interview)

A chief policy officer brought up the technical aspects of NAEP scales, arguing that the NAEP achievement levels need to be improved since they are misleading. In particular, he claimed that the definition of being proficient on NAEP is not consistent with what most states consider to be proficient on their state assessments. For example,

he stressed that NAEP's definition of Proficient is actually higher than the state's definition of the “meet” standard. Because of this inconsistency of the two scales, he contended that it is misleading to compare the percentage of students at or above the NAEP proficient level to that of students meeting standards on the state assessment. An analysis of NAEP data confirms that the US State’s trend line for grade 8 mathematics falls between the NAEP Basic and Proficient lines with being closer to the trend in the percent of students at the Basic level or above.

He used a metaphor that comparing NAEP and statewide results in terms of achievement levels is the same as comparing apples and oranges. Thus, he suggested that NAEP results be reported only by scale scores unless NCES conducts an alignment study for each of all the states participating in NAEP to link the two scales. Here is some excerpt from his statements:

> It is misleading to compare percent meeting NAEP and percent meeting on the state assessment. It’s simply inaccurate and the degree to which you keep doing that with NAEP, you are creating a problem. … If you put the two scales side by side, being proficient under NAEP is really kind of equivalent to being exceeding standard in our state. So this has to do with a technical issue of what is the performance standard. So what happens is [that] people don’t understand this statistical problem with that. They just look at state information and it says 60 % meet standards, [and] they look at NAEP and it says 40 % are proficient. And they conclude that somebody is not telling them the truth. And what typically happens is that they say the state test is too easy. That’s the conclusion. That is totally inaccurate and I think [it is] extremely misleading. So my suggestion was that we not really report percent meeting or percent basic, proficient or advanced but that we report scale scores. …. You [NAEP] could impose alignment of state scale and NAEP scale and do an equivalency study of the two scales and actually create a way to compare apples and apples, instead of apples and oranges. (E-Interview)

With regard to the issue of policy relevance of NAEP data, three major points were made by participants. First, a sample size of minority students is sometimes too small to provide reliable estimates. Although schools with high-minority students are deliberately sampled at a higher rate to assure a sufficient sample size for increased precision of estimates for these subpopulations, sometimes the number of these students selected does not meet reporting standards. For example, in US State, NAEP results for American Indian students are not reported frequently because of the small sample size. Moreover, a small sample of African Americans in the state selected to participate in NAEP caused the same problem until 2000. Further, Grissmer et al. (2000) raise the concern that results from small samples of state NAEP themselves can be more vulnerable to statistical assumptions, estimation procedures, and the influence of a few outliers. Therefore, participants' suggestions of increasing a sample size for these subgroups are considered to be reasonable.

Secondly, an education program specialist maintained that she does not find disaggregated data provided by NAEP policy-relevant since those results are simply statistical estimates. It might be true that some educators and policymakers lack confidence in survey-based results, but NAEP is not designed to assess every student at grades 4, 8, and 12 in the nation. NAEP provides interpretive information such as disaggregated results to help NAEP users better understand achievement results and think about their policy implications. Hence, it appears more reasonable to argue that NAEP should provide more meaningful disaggregated data to inform policy decisions through increasing a sample size of minority groups or collecting more accurate family characteristics.

Lastly, the concern about NAEP achievement levels was raised by one participant. Standards-based reporting is intended to be useful in communicating student results, but NAEP achievement levels have been controversial regarding the process and the outcome of the undertaking (Baker & Linn, 1997; Koretz & Deibert, 1995/1996; Pellegrino et al., 1999). For example, the evaluation studies of NAEP achievement levels by Koretz and Deibert (1995/1996), the National Academy of Education (1996), and Pellegrino et al. (1999) have judged the current achievement-level setting model and results to be flawed.

Nonetheless, it has been widely perceived that the standards-based reporting is likely to enhance the communication of NAEP results to the public and policy makers. Therefore, it seems more urgent and cost effective to improve the methodology for setting achievement levels as suggested by those NAEP evaluation studies, rather than not reporting achievement level results or conducting an equivalent study of the two scales. Of course, it seems persuasive to conduct an alignment study of the NAEP scale with each of all the states' scales so that NAEP results can be validly compared to state results when necessary. Yet, before doing those alignment studies, there appears to be a need to carefully weigh the cost and the benefit of developing equivalency scales to link results from state assessments to NAEP.

State-to-State Comparisons

State NAEP was designed to provide states with information about their students' achievement and to allow states to compare their students' performance with performance of other states. Regardless of early concerns about implementation of state NAEP including the negative effects of increasing the stakes associated with NAEP,

reviews of the TSA (Trial State Assessment) indicated several positive impacts of the program (cited in National Research Council, 2000, p. 9). For example, state NAEP had positive influences on instruction and assessment in the state: 1) increased emphasis on higher-order thinking skills; 2) development of standards-based curricula, and 3) alignment of assessment and instruction.

A question about state-to-state comparisons was intended to probe whether state education staff's perceptions of such comparisons are consistent with the NAEP's intention of facilitating state reform efforts. The majority of participants responded to that question in a negative manner. In particular, they are concerned about the way state-to-state comparisons are reported from the results of state NAEP. For example, demographic differences among states, which are found to be a primary indicator of student achievement, are not taken into consideration in achievement results. In this sense, they asserted that state comparisons are misleading and unfair. A reading specialist contended that state-to-state comparisons are unreasonable without demographic information provided because of their tendency to provide an overly negative portrayal of education systems in states with disadvantaged student populations. He expressed his concern about such comparisons in this way:

> I think the demographic population of states certainly not taken [into] account. So when you see Connecticut year and year out really high, [it] probably has a lot to do with demographics of the students in that state, not necessarily that they have much more effective educational programs than other states. … there are states I am sure that also work very hard but come out low in state-to-state comparisons who have just more challenging student population. …. I think people look at it and say, oh that state is doing much more than [other] states. Without demographic information, I don't think the comparisons are fair. (A-Interview)

Unlike his statement, information on the different demographics of the states' student populations is actually provided along with NAEP results in reporting. The demographic information provided includes the percentage distribution of students along with their achievement data by race/ethnicity, by eligibility for the free/reduced-price lunch program, by English language learners, and by students with disabilities, as well as by exclusion rates. In addition, each state's key demographic information including racial/ethnic background, per-pupil expenditures, and pupil/teacher ratio is available on the State Profiles page of the NAEP website.

An analysis of the website basically supports his argument. For instance, when the performance of US State's students at grades 4 and 8 in mathematics were compared to that of its peer state with similar demographics, there was no significant difference in performance over time. However, even in the case of comparing states with similar student characteristics, demographic information is still needed to verify that such comparisons are appropriate.

Some asserted that state-to-state comparisons are not very meaningful since students in the state are being taught to meet their state standards, not NAEP frameworks. Further, they pointed out that state standards vary from state to state, but that NAEP results do not reflect variances in state curriculums. In this sense, they expressed their concerns about the comparability of the state-to-state comparisons. An education program specialist also noted that a typical use of state-to-state comparisons is rank-ordering of states, which she considers inappropriate. Some excerpts include the following:

> I personally don't find them very useful because I mean it's kind of nice to know how we do in a general sense. But beyond that, it's not really useful because each state has

> you know different standards or maybe you know different things they expect students to know and do. And we are more concerned about whether our students can meet our standards… (D-Interview)

> Every state has its own standards, so NAEP is not measuring those standards. They are measuring their own framework and to compare one state to another based on NAEP framework, you know the most I've seen it used for state to state is for ranking the states. I think that's a very inappropriate use of NAEP data. (G-Interview)

Only two participants (a chief policy officer and a NAEP state coordinator) maintained that state-to-state comparisons are of some value to the state in terms of providing some insight on how their state is doing compared to other states. Not surprisingly, the state-level policymaker placed more value on such state comparisons. However, none of them indicated that such comparisons had had an actual influence on state policy. The remarks of these participants are presented below:

> NAEP data really help us compare our overall performance as the state to other states. That's very helpful to us that we can see how our state is doing in comparison to neighboring State A or State B. …Because the scale is common across all the states, you can look at the performance from state to state and it's very useful. (E-Interview)

> Well, I think they [state-to-state-comparisons] have a value. It gives you a sense of where our state really sits in a bigger picture. (C-Interview)

Interestingly, the chief policy officer commented on the attention that state NAEP results receive from the press in the state. He argued that US State pays a large amount of attention to state-to-state comparisons, particularly to comparisons to State A because of the press's special interest in them. State A is its neighboring state and has demographic characteristics similar to US State's. His remark is presented below:

> We sometimes pay lots of attention to how we do compared to State A because our demographics are very similar. And for example, another measure is the Scholastic Aptitude Test for college entry, the SAT exam, and we always look at how US State's graduates from high school do on that exam in comparison to how State A graduates do…. One of the things that happens with the press is that they will do that anyway. You know newspapers will always ask you "how would you do compared to State A?" So even though it's not a required part of it, we do pay very careful attention to it. (Interview-E)

In particular, the mathematics specialist argued that given some states exempt a large number of students who are in special education or second-language learners from participating in NAEP, state-to-state comparisons are not meaningful. In addition, she stressed that performance comparisons especially between states in different contexts are of no use for enhancing the state educational system, and that comparing states without an in-depth analysis of student performance is not helpful to states. She remarked:

> Perhaps states that are very similar could be compared more reasonably, … why we would want to be compared to Mississippi, why Texas would want to be compared to Washington. It just doesn't make any sense. … The context under which students learn is so different. …I mean I am sure that [state comparison] is interesting, but [if] the question is whether it's useful, the answer is no. Interesting? Yes, useful? No. …. So how would I use that information [that] our students are doing better than [students in] California? How could I use that information to improve our system? I can't. There's no context to use that information. (B-Interview)

Further, two participants (a chief policy officer and an education program specialist) raised concerns about rank ordering of states based on state NAEP results. NAEP results are often reported as rank ordered lists of the 50 states based on mean values through diverse sources including NAGB (Stoneberg, 2005), and state comparisons also have prompted the press to rank states. However, the use of rank order statistics (estimated

scores only) is inappropriate unless standard errors are included in the analysis of NAEP scores (Stoneberg, 2005). The chief policy officer contended that the use of rank order statistics reflects simply the differences of demographics among states, and that thus the rank order can be easily predicted. However, he overlooked some important point that educational factors in addition to student demographics affect student achievement and that thus comparisons made after controlling for the social and economic factors could be useful in examining policy indicators that might have led to improvement in achievement in other states.

> [Rank order of states is] totally meaningless. … unless you are gonna control for all the demographic variables and differences in per capita spending on schools, balance between rural, urban, and suburban schools, I mean that is such a complicated thing and just to put out a rank order list of states is meaningless. … it [rank ordering of states] is predictable based upon the demographics of the states. Small states in the northeast part of the United States tend to have little diversity and large number of middle-class families. And middle-class families have money and well-funded schools in those states. It's statistically very likely that those students are gonna do well on standardized tests. … if you do the same test in a high poverty state like Louisiana or Mississippi with high diversity….. just the impact of demographics is huge in terms of predicting the outcome of rank order. (E-Interview)

In particular, an education program specialist stressed that state-to-state comparisons and the rank order of states promote misunderstanding through the media. She argued that such misunderstandings result from the media's lack of knowledge of major differences between NAEP and the state assessment and of statistics. Her concern was that the misunderstanding or misinterpretation of NAEP results by the media might cause more serious problems since most people such as policymakers, educators, and the public obtain NAEP information from the media. Her remarks are:

> I think it [rank order of states] is inappropriate because it gives the media a false representation of what NAEP is. And the media does not seem to know enough about assessment and specially enough about the differences between NAEP and state assessments to be able to make any kind of significant comments using the data. In fact it promotes misunderstanding. …they [the media] don't understand the very concept [of] statistically significant differences in the first place and they don't understand what the test is designed to measure. And each state has its own standards its curriculum, instruction and assessments are based on, and that is what our students are being taught. … where it [the media] takes the public and state governing agencies, [and] the legislatures, you know the whole thing. It causes a lot of problems sometimes rather than resolving problems. A lot of time it doesn't open a dialogue that we need (G-Interview)

State NAEP was designed to serve as a broad indicator of student achievement at the state level. States might obtain benefits from this national assessment, which permits states to compare to national trends and each other. However, almost all participants expressed more concerns about state-to-state comparisons than appreciation of these comparisons. They argued that NAEP results do not reflect variances in state processes for establishing curriculum and in student backgrounds.

Firstly, they noted that most state assessments measure student performance on their own curriculum standards, say, on what policymakers and citizens in the state consider important for students to know and be able to do. Yet, given that NAEP frameworks are developed based on various sources including state standards and that many states including US State consider NAEP frameworks when developing or revising their own standards, it does not appear reasonable to argue strongly that NAEP and statewide assessments measure different things. Nonetheless, the degree of alignment between state standards and NAEP frameworks varies from state to state and thus it is evident that state comparisons must be interpreted with caution.

As noted by some participants, in state-to-state comparisons different state demographics are not taken into consideration. Therefore, state NAEP results should be considered in context, and comparing the performance of a state's students with that of students in other states with similar resources or students are important and more meaningful. In addition, comparisons made after controlling for social and economic factors might be useful in identifying which states are doing well and what they are doing that works (National Research Council, 1999; Grissmer et al., 2000; Swanson & Stevenson, 2002).

One participant raised concerns that state-to-state comparisons promote a misunderstanding of NAEP results by the press, which is passed on to the public. In fact, research has shown that the meaning of achievement levels continues to be misinterpreted by the press, that statistical misinterpretations remain, and that statistical significance has been sometimes interpreted as substantial significance (Hambleton & Slater, 1997; Koretz & Deibert, 1995/1996; NAGB, 2000). Recently, NCES has placed a page for the media on the NAEP website possibly to facilitate press members' understanding of NAEP and correct interpretations of NAEP results. Further, it would be more helpful for NCES to offer example explanations of NAEP results on the NAEP website in a way that highlights both correct and incorrect interpretations.

NAEP under NCLB

Beginning in the 2002-2003 school year, under the NCLB act any state that wishes to receive a Title I grant must participate in the biennial state-level NAEP in reading and mathematics at grades 4 and 8. State participation in NAEP other than reading and

mathematics in grades 4 and 8 is voluntary. In addition, NCLB encourages states to verify state gains with trends on NAEP. Furthermore, although the specific role of NAEP in the NCLB accountability system has not been determined, some interpret the NCLB legislation in relation to NAEP in this way: NAEP might be used to confirm state performance on their state assessment or further to audit state measures of yearly educational progress. NAGB studies (2002b) recommend that "NAEP results be used to confirm overall improvement but that annual improvement targets on NAEP not be established for states nor used in confirming state results (p. 13)."

Not surprisingly, all the participants responded to a question on this issue in a negative manner. Some argued that comparing the two tests is not valid unless they are highly aligned. Again, they stressed that the state assessment is directly tied to the state standards that teachers are required to teach in classrooms, not NAEP frameworks and thus the two tests are not comparable. They addressed their concerns in this way:

> I would be very much against that [doing an audit]. I wouldn't be supportive about that at all simply because you know each state has their [*sic*] own curriculum standards. Now when we want to go to national standards [or national curriculum], then I can certainly feel that can be used. But we don't have national standards right now… (D-Interview)

> I think I would have concerns about how valid the comparison would be… because I'm not sure how closely the alignment is between NAEP framework and … you know what the state is requiring as far as teaching the standards… If our standards are not aligned with NAEP tests, then I think our standards are better to judge how well students are doing because NAEP tests are not aligned to those standards… so then my concern would be that's not a valid use of that. (F-Interview)

Two participants brought up some technical issues involved. They contended that each state has its own testing program and that each state's scale and the NAEP scale are

not aligned. Moreover, the mathematics specialist stated, “Anyway, it’s premature to use NAEP to do an audit because there is too much variation among states.” She stressed that this kind of analysis is not statistically valid since each state’s assessment model is completely different, adding that to conduct this analysis NAEP and each state must develop an audit model. A chief policy officer also contended that this would not happen unless equivalency scales are developed. He commented that alignment studies by the federal government are necessary to link scores from state assessments to NAEP for an audit. Their arguments were made in this way:

> Each state has its own test, [and] they [NAEP] would have to come up with 50 different ways to do that analysis. If they pick one and they pilot it, say, it’s Massachusetts, for example, they do everything every year. They pick Massachusetts as a pilot, then the states like Massachusetts would do well and audit it. And states not like Massachusetts would not. So it’s not a true indicator to audit each state’s process or test. It’s an audit of how the model worked that they used. So I’m not sure if that can be done unless it’s done concurrently with … side by side state people and NAEP people developing a viable audit model. (B-Interview)

> I think it [using NAEP to validate state performance on state tests] wouldn’t happen unless there’s aligned scales. That’s real simple. If they will compare state assessment performance and NAEP performance on the common scale or linked scale, you are able to know what you are talking about. As long as they do percent meeting at the state level and percent Proficient at the national level without defining the differences between those two, it’s not useful. (E-Interview)

Interestingly, while explaining the relationship between NAEP and the state assessment, the mathematics specialist maintained that NCLB is more aligned with state assessments than NAEP. She claimed that NCLB requires annual testing of grades 3 through 8, but NAEP tests are administered every four years except mathematics and reading. Further, she added that it is impossible to split the NAEP standards into 5^{th}-, 6^{th}-,

and 7th-grade statements since NAEP's 4th-grade statements are the same as 8th-grade statements for mathematics. Therefore, she contended that it is not reasonable to use NAEP results in order to audit state measures of AYP (Adequate Yearly Progress). A statewide accountability system mandated by the No Child Left Behind Act of 2001 requires states to ensure that all schools make Adequate Yearly Progress. AYP is a measure of year-to-year student achievement on statewide assessments. In this context, she argued that NAEP results could be compared only with themselves over time, not with state test results. Her argument was:

> The NCLB requires that assessment and standards are aligned and [assessment is] administered every year. So the level of detail about differences between the 3rd and 2nd grades and 3rd and 4th grades must be described in the standards so [that] they can be assessed. For NAEP, 4th-grade statements are exactly the same as the 8th-grade statements, so states could not use that information in order to break down the 4th grade, 5th grade, 6th grade, 7th grade and 8th grade. … I find that NAEP, my confidence in NAEP is for using it to compare how students are doing on NAEP over time… I don't use other data sources for comparison. (B-Interview)

The NAGB study (2002b) indicates that general trends in the same direction or lack of directly contradictory information should be considered as confirmation. In addition, the NAGB report states that NAEP is attractive as a source of confirmatory evidence since its frameworks are developed considering state curriculum documents. The state NAEP coordinator to some extent agreed, stating, "…in a larger context… you can use NAEP to confirm trends in state performance on the state assessment, but it's a very loose comparison." The state NAEP coordinator seemed to agree to this position:

> There's some real difference between these assessments [NAEP and state tests], [and] so they probably should not be compared directly on one hand. On the other hand, if both NAEP data and state data are going in the same direction, that's a good kind of double check. … I think it's worth looking at NAEP and state assessment data in terms of general trends… If they are not [going in the same direction], then you need to do some more work on that to figure out what might be going on. (Interview-C)

In contrast, other participants raised concerns even about NAEP being used to confirm state test results in a very general way. In particular, they viewed the comparisons as problematic when comparisons of student performance on NAEP and state tests are not congruent. Furthermore, the mathematics specialist argued that the two tests can not be compared at all since they do not measure the same skills: " If I wanted to say, basically speaking, what percentage of students demonstrates at the proficient level in reading on the NAEP test, then I look at the NAEP test. If I want to know what percentage of students are Proficient in reading on the state assessment, I would look at the state's own assessment."

In order for NAGB and NCES to convince state education officials about the value of using NAEP as confirmatory evidence and to help them better develop a more comprehensive picture of student performance, more research needs to be pursued that examines whether and/or how NAEP data could be used for this purpose, probably by undertaking a systematic state-by-state analysis of NAEP and state test results.

NAEP Achievement Levels

NAEP results are reported not only in descriptive terms (summary scale scores) but also in evaluative terms (percentages of students that reach specific levels of performance defined by what students should know and be able to do). Reporting of NAEP

performance by achievement levels is intended to make NAEP results more understandable to policy makers and the public. Participants answered some questions associated with these NAEP performance standards.

First, interviewees' responses to a question about how they felt about the achievement levels were mixed. Three participants commented that in general, these NAEP achievement level descriptors are fine and fairly easy to understand. In addition, a reading specialist commented that NAEP performance standards are higher than US State's.

> They [NAEP performance standards] are fine. I don't have an issue with them. … I think they … have higher expectations than perhaps our state performance standards… I think they are easy [to understand]. (A-Interview)

Two respondents stated that these descriptions appeared clear, but that they were not very familiar with them. According to the state NAEP coordinator, she did not have a chance to read these descriptors carefully before but felt that they were fine, stating that "you know I haven't looked at them really closely, but I think they give quite a detail for science. .....you know these are pretty clear." She stated that she had been given a lot of training about NAEP since positioned as a NAEP state coordinator three months before. It is understandable that she had to learn many aspects of the complex NAEP program over the short period of time. Yet, to promote an understanding of NAEP among different audiences in the state, which was one of her primary responsibilities as the NAEP coordinator, she should have been familiar with NAEP achievement levels. Similarly, a science specialist, who has been in that position since 2004, commented that she is not

very familiar with the achievement level descriptions, but seemed fine with them, stating, “I don't know, I have to look more in depth at framework and then I mean it seemed ok, but I don't know how exactly they are described and what the break points are.”

In contrast, two participants maintained that the NAEP descriptions are unclear. A chief policy officer contended that the achievement-level descriptions are confusing since the definitions of the NAEP achievement levels are unclear on what they mean. He added that information on the achievement-level setting procedures is lacking in reporting. That is, information is unclear on how the NAEP arrived at the definitions of Basic, Proficient, and Advanced and on what criteria were used for setting those performance standards on the scale. Also, an education specialist, who had received a few NAEP trainings and were not very familiar with NAEP, remarked that the NAEP achievement-level descriptors are unclear and too broad. It is likely that a related difficulty that the education specialist saw in the NAEP reporting categories is that at least the Below Basic and Basic categories are simply too broad. In fact, a large percentage of students in most states are generally in the two categories, but no “Below Basic” descriptor was developed and is provided as an achievement level. For example, in the 2005 mathematics assessment 66 % of grade 8 students in US State were classified in the Basic and Below-Basic categories. Thus, it seems reasonable to argue that the Basic category is broad. Even though he did not provide specifics, he stated:

> I remember at that time that I first was introduced to them, I had some difficulty trying to identify in my mind the vocabulary used. Basic, Proficient, and Advanced did not seem to match what in my mind what it was. I didn’t think they communicated very well what they were trying to say. … it [the basic level] didn’t really to me go far enough in showing determining whether students had… the

> knowledge or achievement in that particular category… that particular level was awfully broad, actually they were all broad." (Interview-D)

Interestingly, the chief policy officer raised the concern that the NAEP scale does not help understand whether student performance is improving from grade 4 to grade 8. He argued that as a policy maker it is important for him to know whether or not students make progress from grade 4 to grade 8 and whether or not schools are getting better in educating students. He emphasized that a test must tell how much students learned in addition to how much they know. He commented:

> The other thing that appears to be missing from NAEP, as far as I know, is [that] the scale is not a vertical scale from the 3rd grade through high school. It's not a single scale, [but] it's multiple scales… it's [US State scale] like using a tape measure rather than a bell curve. We don't use bell curves. We are not looking at a random distribution underneath the bell curve, [but] we are looking at actual student growth on a specific scale. … it [NAEP scale] doesn't tell anything about the growth or improvement from 3rd grade through 8th grade. Basically it doesn't tell you if schools are getting better or not. What is the connection between the 4th grade and the 8th grade? Without knowing the scale is connected, it doesn't help you to look at 4th grade data and 8th grade data and to talk about whether or not students are growing or they are learning between the 4th grade and the 8th grade. They only tell you here's how one population of 4th graders did, [and] there's another population of 8th graders did… It's a measurement problem. (E-Interview)

Cohort growth analysis sheds light on the question of how much improvement students make over the four years of schooling from grade 4 to grade 8 (Wilson & Blank, 1999). Consultation with a project officer for design, analysis, and reporting at NCES revealed that the assessments such as reading and mathematics based on frameworks developed in the early 1990's use a cross-grade scale of 0-500, thus allowing for the cohort growth analysis. In contrast, assessments based on more recently developed

frameworks, such as science and writing, use within-grade scales, and thus one cannot estimate growth between grades and how that growth has changed over time. He added that NAEP started using within-grade scales (with the NAEP 1996 science assessment) since it was felt that the cross-grade scales are not based firmly enough to make reliable comparisons between the grades and that the material covered in the assessments is very different for the different grades.

Four participants provided their responses to a question about whether the percentages of students at the proficient or above on NAEP are consistent with those of students meeting the state standards. Two respondents commented that more students meet the state standards than reaching proficient level on NAEP. They argued that this inconsisteny results from the fact that NAEP performance standards are set higher than the state standards. For example, a reading specialist stated, " I think NAEP has much harsher scales associated with it …, but state tests I think show more students in the proficient category than NAEP does. I think there's a higher performance level expected in the NAEP test." A chief policy officer asserted that comparing the two levels is like comparing apples and oranges since these two scales are not comparable, stating, "NAEP's definition of Proficient is actually higher than the state's definition of Meet standard…. the scales are not aligned, [and] then it is misleading to compare percent meeting NAEP and percent meeting on the state's own assessment." This argument was confirmed by an analysis of NAEP data provided on the NAEP website, as noted previously.

The other two participants also contended that the lack of comparability between the two standards does not allow for these comparisons. In particular, the mathematics

specialist argued that the two standards are not comparable in several aspects. She explained that the two assessments measure completely different things and that the methodology for setting the NAEP performance standards is different from that used by the state. However, she did not describe the specifics of the differences in standard-setting methods. She made her point in this way:

> I would guess because state tests again measure completely different things than NAEP does. And the methodology that we used to pick the number, set the number, [and] what the number is, which determines whether or not students meet the standards, is completely different from the methods NAEP uses to set what's Basic, what is Proficient, [and] what is Advanced. … Another component of that is NAEP has three levels, [and] the state assessment has four. So if you separate students into quartiles, it's going to look different than if you use the NAEP system where they separate three groups. So there are a lot of things that statistically don't make sense comparing them. So we don't expect the numbers lined up. (B-Interview)

It should be noted that the comparability of the two tests in terms of performance levels depends on several factors, including the extent of alignment of the standards, standard-setting methods, definitions of the performance standards, stakes attached to test results, the purpose of their performance standards, test formats, test difficulty, etc. (Linn, 2000). In fact, NAEP and the state's own assessment are not similar in terms of standard-setting methodology, test formats (the state test is 100 % multiple choice), and test stakes (the state assessment is a high stakes test for schools). Moreover, the state's content standards are not in tight alignment with NAEP frameworks.

As one participant pointed out, the NAEP standard-setting process is not described in detail either in NAEP reports or on the achievement levels page of the NAEP website. Users need to search relevant reports probably under the publication section on the

website to find out relevant information in order to understand the procedures appropriately and then interpret NAEP results correctly. Unfortunately, to date there is no report provided on the website describing the achievement-level setting process in detail. Thus, NCES needs to provide relevant information on the website to facilitate accurate interpretation of NAEP achievement-level results.

Improvement of NAEP

The final section of the interview protocol included questions asking about reporting data on the various aspects of achievement, integration of NAEP results with education inputs from non-NAEP sources, inclusion of students with disabilities and English-language learners, and making NAEP results more visible and informative. These issues were addressed for the purpose of improving the NAEP program by previous NAEP evaluation studies (Glaser et al., 1997; Pellegrino et al., 1999).

Some of the suggestions made by these studies were that NAEP should: 1) enhance the participation, appropriate assessment, and meaningful interpretation of data for students with disabilities and English-language learners; 2) integrate NAEP results with other NCES datasets to provide a better interpretive context for NAEP data, including the Schools and Staffing Survey, the National Education Longitudinal Study, and the Common Core of Data; 3) report data on divers aspects of achievement; and 4) make NAEP results more visible and informative. Participants' views on these issues are followed. Readers are reminded that all interviewees were not asked about all of these questions due to time constraints.

Integration of NAEP results with other datasets Recent NAEP evaluations suggest that NAEP student, teacher, and school background questionnaire results are not sufficient to meet NAEP users' interpretive needs (Glaser et al., 1997; Pellegrino et al., 1999). A study examined how adequate NAEP's social context measures are for the purpose of reporting adjusted test score differences among population groups (i.e., student test score differences that statistically control for dissimilarities in social context). The study found that NAEP does not provide enough social context measures affecting achievement (Berends & Koretz, 1995/1996) and that the quality of some of NAEP's measures is low because of its reliance on student self-reports.

In order to enable achievement results to be used in a more policy-relevant manner, it would be very useful to integrate NAEP results with data on education inputs by non-NAEP sources collected by NCES (Glaser et al., 1997; Pellegrino et al., 1999). For example, the Common Core of Data (CCD) is a comprehensive data base of all American schools that contains information on school variables such as enrollment size and type of school. The School and Staffing Survey (SASS) surveys samples of U.S. schools to collect extensive data on a wide range of school resources and practices, including fiscal and administrative arrangements, teacher pay, and teacher preparation. In addition, National Education Longitudinal Study (NELS) is another NCES dataset that gathers measures of both system inputs and student outputs. The most recent NAEP evaluation indicates that if those data are coordinated with NAEP data on student achievement, a more comprehensive view of the inputs and outputs of American education would result (Pellegrino et al., 1999).

In fact, as a result of the No Child Left Behind Act, NAGB is responsible for selecting and approving all of NAEP's background questions and actually initiated a process to prepare a general framework to guide the collection and reporting of background data. The framework will guide the development and selection of noncognitive topics and questions, starting with the NAEP 2006 assessment. In particular, NAGB considers using some specific sources of data collected outside of NAEP, including the U.S. Census, Quality Education Data, Inc. (QED), CCD, and SASS (NAGB, 2003).

Five participants responded a question about the issue of integration of NAEP results with education inputs at the student, classroom, school, and state level from non-NAEP sources. However, their responses were not specific enough, which might be because they did not fully understand this issue or did not simply view it as being important. One participant claimed that NAEP is merely one piece of the puzzle, and therefore a broader picture can be obtained through the integration of NAEP results with other indicators collected by non-NAEP sources. The state NAEP coordinator described the big picture in this way:

> I'm just studying what the big picture looks like, [and] so how NAEP could be part of that, the big picture in terms of all the policy decisions the state Department of Education makes and all of the work that the state education agency does to help support education in US State. And so what kind of data from NAEP could help inform those policy decisions about programs, you know how we spend our money, [and] where we put resources. (interview-C)

A science specialist basically agreed to that idea, but was concerned about the cost factor. She stated, "I guess I'm not sure about the benefit versus the cost. It sounds to me

like that would be very time intensive and perhaps expensive to make those connections to be able to correlate and pull data in…" An education program specialist simply commented that all of the background factors collected by NAEP are student-reported, and thus they are suspect. On the one hand, her argument is not correct since background information is collected from not only students but also schools and teachers. Researchers rely on their responses to measure policy variables. On the other hand, there has been some concern that data reported by 4th- and 8th-grade students might make their quality problematic.

The mathematics specialist indicated that it would be better to integrate NAEP and state data to provide a better interpretive context for NAEP results. She remarked that the integration might have to be done in partnership with states because states have more capacity to gather background information on their students that might relate to student achievement. A chief policy officer raised concerns about privacy violation that might occur when integrating different data sets. He argued that a sample size is an issue in survey-based results and that privacy violation depends on how much disaggregation is provided in data.

In fact, the NAEP website provides some information on state characteristics collected by non-NAEP that might affect performance. For example, it presents data on expenditures per pupil, average teacher salary, and pupil/teacher ratio by state in public schools, which are obtained from the NCES Common Core of Data (CCD), Statistics of State School Systems, National Public Education Financial Survey, State Nonfiscal Survey of Public Elementary/Secondary Education, and Estimates of School Statistics.

Yet, these data are not current (collected from school years 2000-2001and 2002-03) and the information needs to be updated accordingly.

In addition, NAGB has established a general framework to guide the collection and reporting of background information including other sources of non-cognitive data (NAGB, 2003). The NAGB report states that the use of data from non-NAEP sources should be increased. However, research has been rarely conducted that combined NAEP's large data sets with other information to tease out policy implications. NCES and NAGB need to make an effort to pursue this line of research and then provide it on the NAEP website.

Reporting data on various aspects of achievement Previous NAEP evaluation studies recommend that NAEP should employ a broader definition of achievement to reflect more complex, rich conceptions of achievement and track this full range of achievement (Glaser et al., 1997; Pellegrino et al., 1999). Further, Pellegrino et al. (1999) suggest that NAEP use mixed methods of data collection (not simply large-scale surveys) to collect information on all aspects of student achievement. An analysis of the NAEP website shows that current NAEP provides overall average scale scores and content subscale scores, but does not provide data on the various dimensions of achievement (e.g., knowledge and skills, problem-solving skills).

An interview question associated with this topic focused on reporting data on diverse aspects of student achievement. Five participants responded to a question about whether NAEP reporting of subscores on higher-order thinking skills is useful. A science specialist commented that NAEP needs to provide performance data on higher-order

thinking skills, which is an area US State does not have sufficient information in. She added that the diverse cognitive demand levels including higher-order thinking skills would be more explicitly articulated in the US State's science standards during the next revision process.

The mathematics specialist also remarked that the state assessment does not provide information on higher-order thinking skills. Therefore, she argued that those data would help the state look into the state education system by providing information on how students do in higher-order thinking skills regardless of their backgrounds including SES and parent education levels. She commented that these data might tell educators and policy makers in the state more about the educational enterprise than simply measuring students to sort in groups.

> Say, this student is someone who is [in] the upper socioeconomic status, both parents have PhDs and that student goes proficient. But a student who has low socioeconomic background where both parents or just one parent never graduated from high school and they [*sic*] scored advanced. What this tells about our system? … Perhaps each state has its own secure student identifier or has data wheel house on students and you know without violating confidentiality, which is basic info about students. What kind of courses they are enrolled in or what kind of schools they attend, [whether] those schools are in high poverty areas such and such, that kind of general information. And then NAEP results could be using you know whatever, put together to come up with a picture of how statewide and nationwide students do in higher-order thinking skills regardless of what their SES is, regardless of what their background is. (Interview-B)

A chief policy officer agreed that NAEP does not provide sufficient information on higher-order thinking skills. He argued that what he sees from NAEP results is the usual percentage of students at the basic, proficient, and advanced levels but that NAEP does not necessarily interpret Advanced as what specific skills have been mastered. He added

that especially in the writing area NAEP does not report enough information on what kinds of writing tasks have been given to students, while the state's writing assessment provides more information on the quality of student writing performance than NAEP does:

> On our state writing assessment, we know [whether] the student was doing personal narrative, or creative or expository or technical writing and various modes of writing at different requirements or what you would consider quality. … So we don't get a lot of information back from NAEP about what kind of writing was provided to students. (Interview-E)

In fact, the NAEP writing framework prescribes that NAEP writing tasks focus on three purposes for writing: narrative (telling a story), informative (informing the reader), and persuasive (persuading the reader). Descriptions of narrative, informative, and persuasive writing are provided in NAEP results unlike the chief policy officer's argument, but separate data on the three writing tasks are not presented.

Two participants viewed this issue somewhat differently. A reading specialist claimed that data on the various dimensions of achievement provided at the state level are not useful and that these data would be useful in making instructional decisions only if given at school or student levels. The state NAEP coordinator simply maintained that NAEP is not designed to provide these data and that a different test designed to serve that purpose would be administered if these kinds of data were necessary. Those data might be more instructionally helpful if provided at the school and student level, but state-level data still are useful in identifying whether students in the state are able to use higher-order thinking skills and whether these skills are improving over time. In addition, NAEP measures

diverse dimensions of achievement, and thus it is possible that NAEP provides data on each dimension.

In particular, one respondent remarked that NAEP does not specify the Advanced level in terms of what specific skills have been accomplished, when reporting achievement-level results. In fact, the achievement-level descriptions are provided in NAEP reports (placed on the NAEP website as well), but the achievement-level results are reported only as the percentages of students performing at each achievement level. Therefore, it would be more useful to provide the results in terms of specific knowledge and skills mastered in conjunction with the percentages of students performing at each level.

Inclusion of students with disabilities and English-language learners Students with disabilities and students with limited English proficiency comprise a significant proportion of U.S. students (Pellegrino et al., 1999). Accordingly, NAEP has made an effort to include as many of these students as possible so that the results can be representative of the nation's students. For example, accommodations in the testing environment or administration procedures are provided for these students. NAEP accommodations include extra time, testing in small-group or one-on-one sessions, reading aloud to a student, and scribing a student's responses. Information on inclusion of these students such as inclusion policy, accommodation types and exclusion rates are provided on the NAEP website.

Five participants provided their responses to a question about this topic. They all agreed that all of these students eligible for meaningful participation should be included

in the NAEP assessment. A NAEP state coordinator claimed that the more inclusive, the more accurate the data are and that students in special needs included in the statewide assessment should participate in NAEP. A science specialist also agreed to the notion that the more inclusive NAEP is, the more valid the findings are, but was concerned that if there is an inconsistency about the inclusion of these students across the states, state-to-state comparisons might be invalid.

In reality, the difference between state-developed and NAEP-established inclusion policies has caused some students of these subgroups to be included in state testing programs but precluded from participation in NAEP (National, Research Council, 2000). Moreover, school personnel are encouraged to use inclusion criteria provided by NAEP, but the application of the inclusion criteria varies from state to state (Pellegrino et al., 1999). In fact, it is generally agreed that high exclusion rates might affect the accuracy of NAEP results and the validity of state comparisons (Olson, 2003).

Further, the education program specialist stressed that sample sizes for those students are too small to draw generalized conclusions about the subgroups as a whole. In addition, the achievements and educational needs of students with learning disabilities might differ greatly from those of students with physical disabilities. Also, there are differences in the achievements and educational needs of English-language learners based on native language, as well as within native language groups depending on how much English is spoken within their home or their specific ethnic identification within their language group (Pellegrino et al. 1999). Therefore, achievement results for these subgroups must be interpreted with caution because of the lack of certainty of their generalizability.

A chief policy officer also argued that it is important to put together samples in a way that a representative sample of these student populations can participate in the NAEP assessment. In addition, he raised concerns that some schools might be tempted to exclude some of students with disabilities and English language learners who can participate with adequate accommodations provided. He stated, “I think the element of manipulating the data sample is a serious potential problem because local schools do not like to receive negative publicity… so if they think their schools are gonna look better by maybe not including as many students with disabilities or limited English proficiency students, there’s a temptation to do that.”

The mathematics specialist emphasized the importance of accommodations tailored to the needs of students with disabilities and those with limited English proficiency to have them fully demonstrate their knowledge and skills. She maintained that if appropriate accommodations are given, special education and limited English proficiency students perform much better than what people think. In this sense, she remarked that if there is a student with disability in processing information, someone should communicate that question to the student as concisely as possible and then glean whatever answer can be obtained from the student. She also contended that many students with disabilities in reading or writing do not have disabilities in mathematics:

> I think they should be fully included …whatever combinations are necessary to help them successfully dispatch their ability… If accommodations are made, students with disabilities do much better than people think. And many students that have disabilities in reading or writing don’t have that disability in math. But because they have to engage in reading and writing in order to do a math test, they are not able to demonstrate that as well. (Interview-B)

It is important to note that accommodations must be provided through modified versions of the assessment or altered administration procedures that do not change the nature of the construct being measured, but that the validity of results obtained under accommodated testing situations in large-scale assessments has not been fully established (Pellegrino et al., 1999). None of participants raised concerns about the validity of different types of accommodations.

NAEP results made more visible A variety of NAEP reports are made available through the World Wide Web. Glaser et al. (1997) suggest that NAEP reporting be adapted to take advantage of the Web's interactive and multimedia features. For example, NAEP could place on the web videoclips of students problem-solving or videotapes of students working in groups to better understand student interactions. In fact, on the NAEP web site, NCES have placed some audio and video files of sample performances from the music and visual arts assessments. Hence, a question was focused on the information that takes advantage of the web's multimedia features.

Four participants responded. Three respondents argued that providing videotapes of student performance along with NAEP results would be informative, while one interviewee commented that she does not view that as significant. A reading specialist remarked that videotapes of having students read aloud might be something that NAEP can do and be useful. A science specialist commented that videoclips of students' doing scientific inquiry using science kits would be valuable. The mathematics specialist also agreed that making videotapes of actual student performance available on the web would help to better understand NAEP results. She stated:

> Think about assessing students, it's nice to have some model something to look at. I haven't seen that yet, so I had no knowledge of that. Having something like that in mathematics and science would be very accessible. … Having students identify what students say, this is a correct one [and] this one isn't a correct one, that would be a great combination of student performance. (Interview-B)

NAEP Data Use by State Education Personnel

The term " NAEP data" was meant in this study to include all relevant NAEP information involving NAEP methodology and resources as well as the results from NAEP assessments, and its meaning was explained to participants during the interviews. The state-level use of NAEP data was examined through interviews, document analyses, and content analyses of the SEA's web site. The focus of the documentary analysis was on corroborating information gained in interviews, as well as on occasions that might not be found through interviews. The NAEP page of the SEA's site was analyzed focusing on how the SEA communicates NAEP information to different audiences across the state and facilitates NAEP data use in the state. Data from these three sources were analyzed independently and the analyses were then integrated. Taken together, the results of those analyses are presented in this section.

Interview questions about the use of NAEP data were mostly categorized into: 1) providing information to the general public, educators, or legislators; 2) informing policy decisions at the state level; 3) informing educational programs at the state level; and 4) accountability purposes. Participants were first asked about how they had used NAEP data, and these pre-categorized uses were asked when all of them were not mentioned. A question asking about the source of NAEP information was followed by these questions.

Responses to a question about how participants obtained NAEP information were mostly directed to the NAEP state coordinator and the NAEP website, as expected. Most respondents remarked that it is very helpful to have someone at the SEA who is very knowledgeable about the NAEP program and works as the liaison between NAEP and the SEA. A mathematics specialist stated that the NAEP coordinator is helpful especially in ensuring that she has the right and most recent information. Five respondents, who indicated that they were familiar with the NAEP website, commented that they acquire NAEP information mostly from the website.

Additional sources were closely related to the jobs they were doing. A reading specialist responded that he sometimes acquires NAEP information from NAEP meetings (e.g., review meetings, release meetings) that he attends on occasion and from press releases. A science specialist commented that she sometimes receives the information through the Council of State Science Supervisors that she serves. She added that recently its members including her were involved in the revision of NAEP science frameworks. The NAEP state coordinator stated that she obtains lots of NAEP information though an internal network for NAEP state coordinators called “My NAEP” that is provided by the NAEP State Service Center. The chief policy officer responded that he obtains information on state NAEP results from a NAEP ambassador as well who visits the SEA before their nationwide release. He stated:

> We also get a visit from someone called NAEP ambassador who comes in and visits usually with me with our measurement person and the person responsible for curriculum and instruction. Three of us plus the NAEP state coordinator, we sit down with this ambassador, the person who comes, and they usually come just before results are going to be released nationally. They bring us our state results so we have a chance to look at them and have some time to analyze how we’ve done before it

> gets to newspapers, because as soon as they release the data publicly the newspaper starts calling. They want our analysis, they want comment, and they want us to be able to respond. So fortunately we have this opportunity to look at them first literally. (Interview-E).

Overall, NAEP data use was found to be limited at the SEA. The main use of NAEP was to provide NAEP information to legislators, policymakers, the public, and educators across the state. NAEP frameworks were also used at the agency to inform policy decisions such as the revision of the state standards. Of course, most participants indicated that they look at NAEP results to see how students in the state have performed when the data are released. For accountability purposes, none of the participants indicated use. All participants argued that using NAEP results for accountability is inappropriate since they are on the basis of a small sample of students in the state. A chief policy officer contended that because state assessment system provides much more data on student performance than NAEP does, NAEP results are not used from an accountability point of view in the same way the state assessment data are utilized. He made his argument in this way:

> NAEP is not mandatory for every student and it's a random sample. So it's not every school, it's not even every district, and certainly not every classroom. So it is not as comprehensive as the state assessment system, which tests every student in every classroom, 3th grade through 8th grade and at least once in high school…So as a result NAEP information is not used from the accountability point of view. (Interview-E)

> It's not [appropriate] because it's a sampling and a very small sampling. It's not appropriate for accountability. And also given that some states are chosen to exempt large number of students who are in special education or second-language learners and kind of disregard federal guidance on that, it's difficult to use [NAEP] for accountability purposes to compare results from state to state. … I wouldn't find it useful in that regard. (Interview-B)

In summary, it was found that NAEP results and resources were used to:

- describe the performance and progress of students and subgroups of students
- make state-to-state and state-to-the nation comparisons
- revise the state standards
- discuss relationships among achievement and background factors
- link state assessment results to NAEP

The following are comprehensive and detailed descriptions of NAEP data use found through analyses of the three data sources including interviews, documents, and the NAEP section of the SEA's web site. Each quotation is followed by alphabets ranging from A to G in parentheses to indicate an individual participant. Data sources (e.g., Interview, Document, Website) might be also indicated when necessary.

Provision of NAEP Information

NAEP has made every effort to communicate effectively its results to intended audiences. As part of the effort, NAEP has assisted states in articulating NAEP results to the general public, educators, policy makers, and those who might be interested. All participants commented that NAEP information was used to share results from recent NAEP assessments with the public, legislators, educators, and parents in the state and/or to promote an understanding of NAEP among them. For example, a chief policy officer reported that he had shared NAEP data with a variety of legislator committees and made presentations to the legislature and the State Board of Education. He added that to disseminate NAEP data, the agency also had placed NAEP information on it's website

and published NAEP results in US State annual report cards, which was confirmed through analyses of the relevant documents and the website.

> That [provision of NAEP information] is probably the main use of the data actually because as I talked about that before it's not very helpful for accountability purposes because it's a random sample. But for talking to groups or giving the general public information on how [the state] schools are doing, I have shared it with a variety of legislator committees [and] made presentations to the legislature. … We put all the information on our website and we try to make the information available. We also put the information into the "State Report Cards," the booklet that comes out every year on the overall performance of [US State] students. (Interview-E)

An education program specialist stated that she had provided the performance of students as a whole and by subgroup for the state. The NAEP coordinator commented that she had not used NAEP data yet (she took over the position 3 months before), but that she planned to work with science teachers in the state to help them understand what they can really do with the 2005 science NAEP data (to be released soon), by making presentations at workshops for them across the state and even at local schools if necessary. She added that the information was also provided through the state publications such as the *Superintendent's Pipeline* and the *Assessment Update*, which was verified through analyses of the relevant documents.

> What I'd like to do with science data is to work with science teachers primarily to help them understand what really they can do with that data. So there are some science groups here, I understand, in US State that might be good to work with. So I am gonna talk to science folks here in the beginning. We have best to do that because I think they have some workshops like in August so I maybe can be there and do some presentation. I'd certainly like to make data user-friendly and give science teachers an opportunity to begin to see what they can do for them. (Interview-C)

This use of NAEP data was also found in documents and reports, most of which were posted on the agency's website. Most documents reviewed provide brief information on NAEP assessments, NAEP testing schedules, and NAEP results for the state. For example, the *Superintendent's Pipeline*, a monthly e-mail newsletter (6-10 pages) published by the agency to alert school districts about upcoming the department requests, activities and opportunities, provided brief information on NAEP results, the difference between NAEP and the state test, and NAEP schedules and placed links to the NAEP website or the state's NAEP site. This newsletter is also provided on the agency's web site.

Press Releases presented the state students' performance on NAEP briefly through statements made by the state superintendent when NAEP results were released nationally. The *Test Administration Manual* for 2005-2006 included an overview of the state's statewide assessment, assessment options, and test administration. The 16-page manual included NAEP information only in the one-page statewide testing schedule. *NAEP Newsletters* provided on the state NAEP site also reported the latest NAEP results for the state and the NAEP schedule for the upcoming assessments. Yet, *Minutes of State Board of Education meetings* did not address any issues related to NAEP.

The *Statewide Report Card* is an annual publication required by law that reports on the status of public schools and their progress towards the goals of the US State Educational Act for the 21st Century. This report card provides student performance assessed through statewide tests, NAEP, international achievement tests, and college admissions tests such as the SAT and ACT, in addition to alternative education programs, school staff, enrollment, and school and district report cards. For instance, the 2004-2005

issue (96-page report) assigned six pages to a short description of NAEP results in mathematics, reading, writing, and science, along with an introduction of what NAEP is, inclusion policies, why NAEP does not report scores for schools and districts, and the policy definitions of NAEP achievement levels, as well as the history of the state NAEP participation. For each subject assessed, the NAEP section briefly described the state's results including the percentages of 4th and 8th graders who met or exceeded the NAEP Basic level.

The *Assessment Update* is a one-page newsletter that provides the latest information on the state's tests including development and administration, scoring and reporting, and psychometrics and validity. It is published biweekly and specifically targeted for assessment people in the state. Almost every issue has recent information on NAEP in it. For instance, it presents NAEP information including assessment schedules, the number of the state schools that have participated, results release dates, and information collected from NAEP High School Transcript Study or NAEP background questionnaires. It also includes brief descriptions of the long-term trend (LTT) NAEP results and the difference between LTT and main NAEP, as well as links to Snapshot Reports for 4th and 8th reading and mathematics, the NAEP Questions Tool, and the State Profile on the NAEP website.

In addition, US State has posted NAEP information and tools on its website to promote correct interpretation and understanding of NAEP data. The NAEP page is composed of five sections: news announcements, pages, miscellaneous internal links, external links, and contacts. The "pages" section includes NAEP frameworks, glossary terms, newsletters, resources, NAEP results, sample questions, and Frequently Asked

Questions. NAEP newsletters generated by the coordinator are two or three-page summary reports that describe the latest NAEP results for the state. NAEP Resources is linked to the NAEP homepage, the NAGB site, and the NCLB website, while NAEP frameworks linked to corresponding pages on the NAEP site. NAEP Results is linked to Nation's Report Card, highlight reports, and state Snapshot reports for the state (1998 through 2005 results) on the NAEP website. In addition, the NAEP page is internally linked to the state's Assessment/Testing page. The external links are linked to the state profiles page and NAEP schedule from 2005–2017 on the NAEP site.

In summary, it was found that NAEP information was primarily used at the SEA to promote an understanding of what NAEP is and how it works and to describe how students and subgroups of students in the state are performing against NAEP criteria. This use was consistent through the three data sources collected.

Making Comparisons with Other States and the Nation

NAEP is considered by education stakeholders to provide valuable information for national and state comparisons (NRC, 1999). In fact, state NAEP was developed with these comparisons in mind. However, most participants expressed concerns about state-to-state comparisons stating limitations associated with such comparisons, as described in the earlier section.

It was found that NAEP data were used to examine how US State students performed compared to other states and the nation, but the focus was primarily on state-to-the nation comparisons. During interviews, only two participants (a chief policy officer and an education program specialist) remarked that they used NAEP data for state comparisons.

For example, an education program specialist stated that while in the previous position, she provided NAEP results for the state in comparison with the nation and other states. The chief policy officer contended that NAEP data really helped the state compare its overall performance to other states and that he frequently provided presentations to legislators and the State Board of Education regarding how students were doing in comparison with other states and the nation.

Analysis of documents confirms that the state NAEP results are used to make comparisons with national and regional results, but not with other states. For example, the November 2005 issue of the *Superintendent's Pipeline* briefly provides the 2005 US State results at grades 4 and 8 in reading and mathematics, comparing them to the nation but not to other states. *Press Releases* also present NAEP results compared to previous assessments and the nation. Interestingly, although stating NAEP is the best national comparison of how well students perform from state to state, Press Releases do not make state comparisons. The *Annual State Report Cards* and NAEP newsletters also provide state results in comparison to both the nation and the region.

However, a special report developed by the assessment office, entitled "NAEP State writing report for the 2002 writing assessment," provides comparisons between US State and other states in terms of scale scores. This document presents results for US State's students at grades 4 and 8 in writing, in comparison to the nation, the region, and other states.

In summary, US State used NAEP data primarily to compare its performance with the nation and state-to-state comparisons were found to be infrequent. Comparisons to other states with similar characteristics such as State A were not found in documents, which

conflicts with the argument made by the chief policy officer during interviews. Yet it might be possible that although he used NAEP data for such comparisons, those uses were not merely documented. The rare use of state-to-state comparisons might be related to the notion by most participants that such comparisons are misleading and not useful, but this study was not designed to identify that connection.

Revision of the State Standards

Since 1990, NAEP has used its performance standards to chart the progress of the nation's students toward high academic achievement (Pellegrino et al., 1999). Reporting achievement-level results is intended to understand whether student performance is "good enough," that is, reporting of NAEP results in evaluative terms. Reporting of NAEP performance by achievement levels has sometimes driven educational debate in the state and/or changes in a state's policy (Glaser et al., 1997; National Research Council, 2000; Pellegrino et al., 1999).

According to a chief policy officer, US State was one of the states that changed state policy due to NAEP performance. He argued that after examining NAEP results, the state revised its English (in 2005) and mathematics (in 2002) standards with inclusion of the structure of NAEP standards as part of that work to ensure that the state standards are rigorous enough. Yet he did not mention what he meant by "rigor." He explained that the SEA conducted a match gap analysis between what the existing standards included and what was in NAEP frameworks and then incorporated missing information into the state standards. Below is part of his remarks:

> We look at the NAEP results to help us look at our own standards to make sure our standards are rigorous enough and [that] the standards contain the material that is measured under the NAEP exam. And so we have revised both our English and mathematics standards after looking at the results we achieved under the NAEP test. (Interview-E)

Other participants also remarked that the state's standards were revised on the basis of NAEP frameworks. For example, the mathematics specialist asserted that she had used NAEP frameworks and NCTM standards of 2000 in 2002 to help complete the final version of the mathematics standards. According to her, the revised standards were presented at the State Board of Education with evidence of their alignment with NAEP frameworks and then adopted by the board.

> The State Board must approve the mathematics standards. So they were concerned that [the revised] mathematics standards are well aligned with NAEP frameworks so we provided evidence and they did. (Interview-B)

A reading specialist stated that the state had referred to NAEP reading frameworks when refining its reading standards in 2005, which supported the argument by the chief policy officer. For the state's science standards, a science specialist remarked that a match gap analysis had been performed when revising the standards in 2001, but that she had not been in her current position then. She argued that she would utilize NAEP frameworks as a resource during the next revision of the state curriculum standards that would be conducted in a few years.

A document entitled "Academic Content Standards: Creating consistency across US State" confirms that a match gap analysis is done when revising state standards. It states

that a match gap analysis is conducted to identify any possible gaps between the existing state's standards and national standards (including NAEP frameworks) and that identified gaps are then closely examined by the Content and Assessment panel for the given content area (USDE, 2005). "Because the state's standards are prioritized to address content most important for students to know and be able to do, there may be material in the national standards not included in the state content standards" (USDE, 2005, p. 5). According to the document, a gap analysis between the state standards and NAEP framework is important since NAEP is the basis for making state-to-state comparisons of the content area programs of statewide educational systems.

A match gap analysis performed for the 2002 revision of the state's mathematics standards was also verified by another document. The document has a side-by-side analysis of the NAEP and the state standards by grade level and by content area. Each grade-level standard contains a reference to NAEP frameworks and/or NCTM (National Council of Teachers of Mathematics) standards. However, the state did not conduct an alignment study to identify the exact extent of alignment between the two. The mathematics specialist responded to the relevant follow-up question that the two standards are not similar in several aspects enough to allow for quantification of the degree of alignment. She responded:

> The NAEP standards and state standards are not alike enough to be able to quantify a degree of alignment. The NAEP items are few, and [NAEP frameworks are] broad in scope compared to state standards. The state standards are more descriptive at each grade and are required to align to the NCLB requirements (for annual assessment) so a broadly stated generalized alignment is the best scale of alignment possible. (Follow-up comments, B)

The “US State Standards Newspaper” also supports the notion that the state’s standards are to some extent aligned with relevant NAEP frameworks. It is an approximately100-page report published each year to provide teachers, administrators, and other education leaders in the state with the most updated information about the state’s academic content standards and performance standards. For example, the issue for 2005-06 School Year includes a one-page table showing that the state academic content standards are developed or revised based on both national standards developed by professional organizations and NAEP frameworks. Yet it does not provide specifics of the alignment issue including the degree of alignment. Minutes of State Board of Education meetings (October 2005-April 2006) placed on the Board website were also reviewed, but no issues related to NAEP including the adoption of revised state standards were addressed.

Relationships between Performance and Background Factors

The current NAEP background questionnaire results could be used as the primary source of data to meet NAEP user’s interpretive needs, namely, possible sources of promising or poor performance. This information is useful in examining the relationships between background or contextual factors and student achievement. Again, readers are reminded that no causal inferences should be made between achievement and these variables.

As noted previously, in US State NAEP data were primarily used for descriptive and evaluative purposes, and no use for interpretive purposes was indicated through inter views. However, documentary analysis revealed that the SEA had made an attempt to use

NAEP for the purpose of providing interpretive information to assist educators, policy makers, and the public in beginning to think about policy implications of NAEP results.

One document, entitled "NAEP-language and literacy issues: Factors related to student performance," addressed the relationships between achievement and variables that might relate to student performance. The report states that its purpose is to help communities, educators, and parents build capacities for using effective practices to improve students' academic achievement in US State (US State Department of Education, 2004). The document examined 4th and 8th grade students' performance on NAEP in the state in relation to their own perceptions of home and other factors outside of school, including language other than English spoken in the home and literacy materials in the home. The home factors were based on students' responses to background questionnaires for the NAEP 2002 reading and writing assessments and the 2003 reading and mathematics assessments.

State NAEP collects background information on major educational policy areas such as characteristics of teachers, instructional approaches, access to resources, or teacher preparation. Possible relationships indicated between achievement and variables could be further explored with a strong research model. In this way, policy makers could determine the sources of good or poor performance and then make policy decisions for improvement. Clearly, the state's attempt to link non-educational factors to student academic performance represents a first step toward utilizing NAEP data for interpretive purposes. Furthermore, the state efforts might be expanded to investigating policy implications of NAEP results in terms of policy-relevant correlates of proficiency. At the same time, NCES needs to place more effort into this line of research. These efforts could

help states refine current approaches in an ongoing process of improving education (Swanson & Stevenson, 2002).

Linking of State Test Results and NAEP results

Statistical linkages among existing assessments might provide a basis for comparability (National Research Council [NRC], 1999), and there has been growing interest in linking statewide test results to NAEP to enhance the comparability of student achievement between the two. In general, linking means putting the scores from two tests on the same scale and includes a variety of approaches to make results of one assessment comparable to those of another (Linn, 1993; NRC, 1999).

In developing an equivalency scale to compare the results of different achievement tests, one should consider multiple potential factors that affect the validity of inferences drawn from the linked scores. These factors include the content and format of the tests, test administration, the stakes attached to test results, etc. In particular, the unique character of NAEP, which is substantially different from state tests in design and implementation, presents significant challenges to linkage (NRC, 1999).

Despite the serious challenges that the linking of state test results to NAEP encounters, the benefits might outweigh the concerns (Mullis, 2003). Clearly, linking state tests to NAEP enhances the utility of NAEP information. For example, the linkage could enable states to evaluate their own assessments against a national criterion and facilitate state-to-state comparisons. Also, by comparing the performance of a small sample of students to the performance on the state's own test, results from more frequently administered state tests could be translated into estimates of results on the

NAEP scale. As a result, the linkage may lead to monitoring trends of student achievement on the NAEP metric on an annual basis without heavy reliance on NAEP. Further, linking might permit reporting of district and school results in the NAEP metric, and thus educators and policy makers could compare school and district results from statewide tests with NAEP results.

Several states have already begun to conduct partial linkages (NRC, 1999). Documentary analysis revealed that US State had conducted a linking study using the 1998 state NAEP in reading and its matched state test scores. The study was designed to produce a transformation that would turn the state's reading assessment scores into NAEP proficiency estimates. In fact, the US State assessment is not parallel to NAEP tests. For example, they have items in different formats: the state test is entirely multiple-choice, while the NAEP format has the approximately 50-50 distribution of multiple-choice and constructed-response items. The frameworks on which the items are based are somewhat overlapping, but different from each other.

According to the author, with those differences in mind, the linkage was conducted using the methodology of "projection plus variation," which is a variant of projection where statistical regression methods are applied (McLaughlin, 2000). Projections can be made to produce predictions of aggregated score distributions of NAEP on the basis of state test scores as long as there is a relation between the two measures (Linn, 1993; Mislevey, 1992). The methodology was used to produce unbiased links to both mean levels of group performance and percentages of groups performing at the NAEP achievement levels (McLaughlin, 2000). The study found that the linkage could be used to extrapolate to other samples of students in US State in 1998.

According to follow-up comments by the NAEP state coordinator, the original study was conducted to: (1) provide evidence of concurrent validity of the state reading assessment with other measures of the same general construct, and (2) provide an alternate means of reporting state assessment results (project onto NAEP scale). Yet she commented that because educators seem generally less familiar with the NAEP scale than with the state scale, the SEA has not tried to use the projected NAEP scores. She remarked that there is some potential confusion since the numerical scales overlap and that the state assessment staff are unsure of the effects of differences in the content standards and NAEP framework and differences in response mode. As a result, she pointed out that the agency would need to do some work to determine the above issues and their impact on the results.

Use of the NAEP Website by State Education Personnel

Several questions on the NAEP website were asked, including its usefulness and expectations for the website. As expected, the NAEP state coordinator has used the website the most frequently. In fact, NAEP state coordinators are given special training to use the website and part of their job is to collect and analyze data provided on it for reporting. A chief policy officer commented that he had used the website many times, while other participants reported infrequent use. Two participants indicated their one-time use of it, while one indicated no use. Overall, the NAEP website use by state education staff was found to be very limited. This section focuses on an analysis of

participants' responses to questions about the use of the website and their expectations about the site content.

The Website Use

Most used areas of the website were found to be NAEP frameworks, state profiles, NAEP Questions Tool, and the NAEP Data Explorer (NDE). Most participants stated that the website is user-friendly and that the quality of information is fine. However, one participant briefly expressed his uneasiness about using NDE, which is designed to search and display variable tables using all major national and state NAEP assessments since 1990. The NDE allows for the disaggregation of data by each or any combination of approximately 1000 variables. In particular, he appeared to be overwhelmed by such many variables to choose from:

> I can use the Data Explorer a little bit to do some comparisons from state to state…. I find the Data Explorer a little complicated to use. I wish this was [*sic*] a little easier. … It's setting up all the variables in it. (Interview-E)

An education program specialist raised concerns about the possibility of misinterpretation and misuse of data obtained from the website. She pointed out that some have difficulty in interpreting data they have gotten from the website and that there must be a differentiation between using the website to find needed information and what interpretation is made. In fact, studies indicated that misunderstandings and misinterpretations regarding NAEP results mostly result from user's unfamiliarity with the NAEP reporting scale and limited knowledge of statistics (Hambleton & Slater, 1997;

Koretz & Deibert, 1995/1996). Similarly, she maintained that some misunderstand statistical findings presented, thus leading to inappropriate use of the data. For example, she remarked that it is inappropriate to make generalizations of NAEP findings from students with disabilities to this subgroup population:

> For example, students who are in special education programs, NAEP calls it students with disabilities, … [to] make generalizations about that group of students is totally inappropriate. But there are people out there who would do that. And there are people out there who use it, who use any kind of statistical findings for their own purposes, rather than for the purpose that the data was designed to be useful for. (G-Interview)

Additionally, she pointed out that some information on the state profiles page is outdated. The state profiles section compiles data from the Common Core of Data collected by NCES and provides easy-to-use features for graphical presentation of results for each of states participating. The state profiles present key data about each state's student and school population, NAEP testing history, and NAEP results. She commented that information on student demographics provided for each state is not current (a year and half old), but is simply the most recent data collected.

> The problem with some of the information on the site is that it's outdated and it's not current. So it's not accurate. … when people see things on the web and newspapers wherever, they think it's absolutely true, and it's not. … I believe they are from the Common Core of Data pulled out by NCES. And even when NAEP does sampling of states, they are using data that is a year and half old. It's not what's really happening in the states, [and] it's not really the number of students in the schools and districts. It's a year and half old. Same thing with identifying, for example, limited English proficiency students. (Interview-G)

With regard to the infrequent use of the NAEP website, most participants responded that they simply feel that they do not have the reason to use it. A few stated that if they were involved in developing or refining state standards, they would probably refer to the website more often. Interestingly, the education program specialist stated that the person who uses the website in the SEA in most states is a NAEP state coordinator so that the coordinator can provide data to state education staff who need it.

These findings indicate that state education personnel do not have an in-depth knowledge about the NAEP program and the NAEP website. In particular, the participants, who reported rare use, appear to perceive that NAEP frameworks are a major source of NAEP information provided on the site they can use. Thus, they did not feel that they need to visit the website for NAEP information, unless they work on the revision of the state standards.

Moreover, none of the participants responded that they had made any request for information through a feedback form set up on the NAEP website. The researcher herself asked a NAEP question through the request form on the site and received a response by a relevant NAEP staff member via e-mail within 24 hours. She also asked specific questions directly of NAEP staff members using e-mail addresses provided on the site. The researcher feels that their responses were very professional and useful in better understanding issues associated with NAEP. It seems desirable that state education personnel use these resources actively to advance their understanding of NAEP and/or share their perceptions of NAEP issues from a state perspective with NCES staff.

Expectations for the NAEP Website

When asked about what questions participants expect to be answered through navigation of the NAEP website from a state perspective, their responses were somewhat diverse. Below is a list of NAEP information that they expect to obtain from the website:

- State data on their students' performance and progress over time
- Performance of students in different demographics groups in the state
- Student performance among states
- NAEP frameworks
- NAEP items and analysis of student responses
- NAEP scaling
- National trends for best practices
- Resources for classroom teachers
- Latest research on reading, writing, or mathematics
- In-depth data analysis of student performance
- Research on how to use NAEP results to inform instructional policy decisions

The following are some remarks from participants:

> One is that I expect to be able to get clear information on my own state very quickly and I'd like to be able to ask questions into clearly the system about my state in an easy manner. The second thing that I'd like to know is how other states are doing so I could compare US State to other states. And then the third thing is I am very interested in what you just told me about items themselves and how they are scaled. (Interview-E)

> I guess what I will be looking at is any kind of analysis that talks about how the [NAEP] results could be used to inform instructional policy decision. (Interview-F)

> Well certainly state [NAEP] assessment data and information. I think that looking at national trends for best practices is another use. (Interview-C)

The NAEP state coordinator remarked that she expected to see what practices work, what practices do not work, and national trends for the best practices. However, as noted in earlier sections, correlations of NAEP performance with instructional practices obtained from background information do not imply causal relationships because of the nature of NAEP design. Therefore, NAEP does not allow users to identify national trends for the best practices.

NAEP provides state-level estimated average scores and standard errors disaggregated by gender, race/ethnicity, eligibility for the free/reduced National Lunch Program, students with disabilities, English-language learners, etc. Again, the education program specialist raised concerns about NAEP's representative samples especially regarding disaggregated data. She contended that sometimes samples for minority students in the state (e.g., Native American students) are too small to produce meaningful results.

> We rarely have any kind of data reported for our Native American students because at the time of the sample they do, it doesn't do quite enough students of Native American origin in that category. It [NAEP] just doesn't capture enough of students taking NAEP to be able to reliably report for that subgroup. And this representative sample is always going to be a problem. (Interview G)

In reality, sometimes the sample size for minority groups is not sufficient to provide statistically reliable estimates. For instance, 4th-grade Native American students' performance in US State was not reported from 1996 through 2005, while performance of

8th graders in this subgroup was not available from 1990 through 2000. Similarly, Black students' data for grade 4 were not reported in 1996 and 2000, while 8th-grade Black students' performance was not provided from 1990 through 2000. The performance of private school students in US State was not reported as well. Further, she pointed out that standard errors of the estimates for these subgroups tend to be so high that the results are not really meaningful. In general, the estimates for these subpopulations have larger standard errors because of their smaller size (Stoneberg, 2005).

In addition, the specialist argued that it would be better if NAEP information were more timely. NCLB requires that results in the mandatory subjects (4th- and 8th- grade reading and mathematics) be released within six months of data collection, which is a far faster schedule than ever before. Nevertheless, when published, NAEP results are still for some previous year's fourth and eighth graders. In contrast, state and local results are timely, reporting data for a cohort while it is still in the particular grade (National Research Council, 2000). As a result, when policy makers attempt to compare cohorts across NAEP and state assessments, they realize that the results are for different cohorts.

Summary

This chapter presented the analyses of the NAEP website, interview responses, documents, and the SEA's website to answer the research questions guiding this study: 1) what is the nature of the NCES website in terms of NAEP as sources of data to inform state educational decisions?; 2) what are the state education personnel's perceptions of

the use of NAEP in making informed educational decisions?; and 3) how are NAEP data used in supporting the state in responding to current issues in education?

Firstly, an analysis of the NAEP website revealed that NAEP results and resources could be used to inform state educational decisions in several ways: 1) developing or revising state standards; 2) tracking student performance over time; 3) examining changes in achievement gaps; 4) making state-to-state comparisons; 5) comparing state test results with NAEP results; and 6) exploring possible relationships between achievement and background factors. In addition, NAEP performance standards and released NAEP items and student responses were found to be able to be used to inform educational or policy decisions in a state context.

Secondly, it was found that state education personnel's perceptions of the usefulness of NAEP overlapped among participants regarding some issues, but were sometimes different depending on their position at the SEA and their familiarity with NAEP. For example, a chief policy officer considered state-to-state comparisons useful, but almost all participants indicated that such comparisons are not useful since they do not reflect variances in state standards and student demographics. On the other hand, all participants indicated that comparing the US State's assessment results with NAEP results is not valid because the state's content standards are not highly aligned with NAEP frameworks and because the state's performance standards are different from NAEP's.

Lastly, it was found that NAEP data were used by the state education personnel in US State to: 1) describe the progress of students and subgroups of students; 2) make state-to-state and state-to-the nation comparisons; 3) revise the state's standards; 4) discuss relationships between performance and background variables; and 5) link the state

assessment results to NAEP. Overall, this study suggests that the use of NAEP data was limited in relation to their potential utility and that the main uses were limited to providing NAEP information to stakeholders in the state and to consider NAEP frameworks when revising the state standards.

CHAPTER V

DISCUSSION AND IMPLICATIONS

This chapter consists of three sections. The first section discusses major themes and patterns that emerged from this study, compares the findings of the study with those of previous studies when appropriate, and explores their implications for the use of NAEP. In the following sections, implications for future research and the limitations of the study are discussed separately.

Discussion

The NAEP website was created in an effort to make NAEP information more accessible and useful to policy makers, educators, and the general public. In general, participants' use of the NAEP website was found to be limited. One can argue that their limited use might raise a concern about the validity of their responses to interview questions. Frequent use of the website indicates that participants may be knowledgeable enough to address issues related to NAEP and that in turn, their responses could be more credible. Yet, it was found that the frequency of the use of the site varied among them and that they had at least basic understanding of the NAEP program through reading printed NAEP reports and attending NAEP training sessions or NAEP meetings. In addition, their interview responses were triangulated using other data sources such as documents and the SEA's website. Therefore, the findings of this study are considered to be valid.

This study suggests that the state education personnel's use of NAEP is limited in relation to its potential utility and that their perceptions of the use of NAEP are relatively diverse depending on their positions in the SEA or familiarity with NAEP. These findings evoke some interesting discussions, which follow in the below subsections. This section focuses primarily on a discussion around the major findings of the study that might stimulate educational debate regarding the issues relevant to the use of NAEP. The following are a discussion of major issues related to alignment of state standards and NAEP frameworks, closing achievement gaps, NAEP achievement levels, state-to-state comparisons, and NAEP data use by state education personnel.

Alignment

A close alignment of state standards with NAEP frameworks might add credibility to the state standards and facilitate comparisons of student performance on NAEP and state tests. Of course, it might be possible that although the standards are comparable, comparisons of student performance on NAEP and statewide tests are not congruent. In this case, other factors might have impacted the discrepancy. Future research needs to be conducted on this regard.

What happens if the two standards are not comparable? The findings of this study suggest that when the two standards are not in close alignment, such comparisons might cause a conflict. For example, US State considers relevant NAEP frameworks when revising the state standards, but the two standards were not in a high degree of alignment. In fact, the extent of alignment was identified through neither interviews nor document analysis. State education personnel in US State generally felt that there was no value of

comparing the state and NAEP test results since the two standards were not closely aligned. They thought that even confirming the general trend of performance on the state assessment with NAEP results was meaningless. For example, the mathematics specialist stated that " I can only compare the NAEP information to itself. … I don't use other data sources for comparison," which represented the general view of the state education staff on this issue.

However, given the imperfect nature of measurement, one can argue that NAEP results could be used to establish concurrent validity of state assessment results. Concurrent validity is demonstrated when scores on an instrument are well correlated with current performance on some other instrument that measures related constructs. When there are gains on a state test, other measures of achievement in the same content area need to show gains as well (Linn et al., 2002). In fact, an analysis of the US State's performance data on the grade 8 mathematics assessments indicated that the results of NAEP and the state assessment did not show the contradictory trend over time as a whole in terms of meeting standards. This finding is somewhat consistent with the conclusion made by the NAGB study (2002b) that NAEP could be considered to confirm the general trend of state test results in grades 4 and 8 reading and mathematics unless there exists evidence of directly contradictory information. On the other hand, the analysis indicated some inconsistency for major ethnicities in terms of changes in performance gaps.

The literature suggests that both NAEP and TIMSS results could be used to help improve mathematics education (Blank & Wilson, 1999). NAEP frameworks are similar to TIMSS frameworks in terms of the framework development process, which is based on

a consensus approach involving a wide range of experts such as subject experts, education professionals, and measurement specialists, except that in TIMSS the consensus process involves experts from many countries. TIMSS was designed to promote understanding of the educational context in which learning takes place for mathematics and science at grades 4, 8, and 12. For mathematics, the two frameworks are quite similar with respect to the broad structure of their content dimensions (Wilson & Blank, 1999). These two frameworks structure their cognitive dimensions differently, but there is considerable overlap in the specific process skills, abilities, and competencies that are considered important to be included in each assessment to demonstrate performance (NCES, 2006). Further, in recent years studies have been conducted that linked NAEP to TIMSS results to see where states would have performed if their students had been in TIMSS (NCES, 1999b; Johnson, Cohen, Chen, Jiang, & Zhang, 2005).

Blank and Wilson (1999) analyzed areas of strength and weakness in U.S. mathematics performance on NAEP and TIMSS and suggested that a more meaningful picture of mathematics education in the U.S. can be obtained by studying the meaning behind average scores and summary statistics reported about NAEP and TIMSS. State education personnel may interpret these data on U.S. students in the context of their states. However, the NAEP website does not provide any description of comparing the two frameworks in an effort to help users understand that NAEP frameworks reflect international demands in key subject areas. This information also might help NCES articulate the level at which NAEP tests the content and what students need to know and do to be successful in college, the workplace, and international competition. This kind of information could further encourage state education personnel to seek TIMSS data as

another indicator of student performance to inform state decisions on how to improve teaching curriculum.

Closing Achievement Gaps

The No Child Left Behind Act is intended to hold public schools accountable for closing performance gaps between different groups of students. Examination of the results of the statewide assessment and State NAEP among subgroups could provide a more comprehensive picture of how subgroups are doing in comparison to the state performance as a whole and of whether or not gaps in performance among these subgroups are narrowing. States have focused on closing achievement gaps among subgroups of students as a major reform effort. NCLB also stresses improved achievement by *all* students by requiring states to set AYP objectives. The NCLB of 2001 mandates statewide assessments to be given yearly to at least 95 percent of the student body at schools receiving public funding. The state test results are to be used to evaluate improvement in achievement overall and for specific groups: 1) economically disadvantaged students; 2) students from major racial/ethnic groups; 3) students with disabilities; and 4) English language learners. States must develop AYP statewide measurable objectives for improved achievement by all students and for the disadvantaged groups. The exact statistic used in measuring AYP is left up to the individual states.

Closing the achievement gap is also a priority initiative at the SEA in US State. An analysis of NAEP data for US State indicates that NAEP and the state's assessment results in terms of meeting standards were somewhat inconsistent with each other

regarding the change in achievement-level gaps for White-Black and White-Hispanic students in grade 8 mathematics. For example, NAEP data indicated that the gaps between White and minority students fluctuated over time but were somewhat widening, while the state data indicated the very gradual decrease of the gaps. The discrepancy might be due to the fact that the US State's performance standards are different from those adopted by NAEP. For example, NAEP's definition of "proficient" may encompass different skills from the state's definition of proficient. In addition, other factors might have had an impact on the discrepancy, such as test content, motivation, and test formats.

In general, almost all participants agreed that disaggregated data might inform state policy decisions regarding issues associated with closing achievement gaps among subgroups. Yet, an analysis of NAEP data suggests that a sample size of diverse demographic subgroups for NAEP is sometimes too small to produce statistically meaningful estimates for relevant subpopulations. For example, NAEP results for American Indian students in US State have never been reported for grade 4 because of the small sample size, while grade 8 results for this group began to be provided in 2003. Similarly, performance data for African Americans in the state have been provided since 2003. Thus, state education personnel felt that increasing a sample size for these subgroups would make NAEP data more policy-relevant.

State NAEP was designed to provide the academic performance of students as a whole and by subgroup. However, most of the state education personnel pointed out that insufficient information on the performance of subgroups is not useful in their decision and policy making. Thus, it is reasonable to demand that NCES make an appropriate decision regarding whether to increase the sample size for these subgroups to better serve

as an independent measure of the performance of groups and subgroups of students at the state level. Interestingly, some participants argued that NAEP's disaggregated data are simply not helpful because of their basis on a small sample size for subgroups. This argument is not reasonable since the disaggregated data are considered to be representative of the relevant subpopulations as long as the sample size meets statistical reporting standards.

This concern also applies to students with disabilities and English language learners. One participant argued that performance data for these subgroups could not be generalized to the subgroups as a whole. Her concern is somewhat consistent with the notion that achievement results for these subgroups lack certainty of their generalizability because the achievements of students with learning disabilities might differ greatly from those of students with physical disabilities and because there are differences in the achievements of LEP students based on native language (Pellegrino et al., 1999).

The concern raised regarding the lack of generalizability of performance data for those students in special needs is reasonable since the survey-based measure such as NAEP should provide estimates of the performance of subgroups of students representative of that of the subpopulations. First of all, NAEP reports need to convey careful caveats on the lack of the generalizability of the information to all types of students with disabilities or types of language groups. At the same time, NCES needs to conduct extensive research as to how different types of learning or physical disabilities affect student achievement and how differences in native language impact the performance of English language learners. Finally, it appears challenging to oversample students with disabilities and English language learners by type in order to make

generalizations possible. Hence, before making any decision, it is necessary for NCES to carefully determine whether the benefits of over-sampling justify the costs.

Achievement Levels

The current reform efforts place great emphasis on the goal of high performance standards set for all students. These performance standards are intended to specify "how good is good enough" (Linn, 2000). NAGB reports NAEP results by performance standards to make them more understandable to the public and more useful to those who make instructional and policy decisions. NAGB believes that its policy of reporting performance by achievement levels has been more useful to the public and policymakers than if results were reported only normatively for NAEP (NAGB, 2000). The question to be addressed now is whether NAEP users feel the same way and whether policy makers perceive that the achievement levels are useful in their decision and policy making. Little research has been conducted to answer these questions, while this study examined how state education personnel perceive the usefulness of the NAEP achievement levels. The following is a discussion of the findings of this study on the state education personnel's perceived values of NAEP achievement levels and of the controversy surrounding the achievement levels.

Participants' responses to the NAEP achievement levels were mixed. Some viewed the achievement levels as being easy to understand, while others considered them unclear and vague. In this study, some argued that the definitions of the NAEP achievement levels are unclear on what they mean and that the descriptions for the achievement levels, especially the descriptor for Basic, are too broad. The finding of this study that some

considered the NAEP achievement levels easy to understand is somewhat consistent with that of a study by Bullock and DeStefano (1998). They examined the perceptions of the usefulness of the 1992 TSA results in reading through interviews with state assessment directors and found that most participants viewed NAEP achievement levels as being understandable and useful.

The Bullock and DeStefano study (1998) also found that some of these NAEP users perceived the category of Below Basic as vague since the percentage of Below Basic students is provided with no description for that category. They suggested the need for NAEP to define a category of Below Basic so that students at the Below Basic level can also be provided meaningful reports on their status on the standards. That is, what are the knowledge and skills possessed by students at the Below Basic level? This finding is to some extent consistent with the finding of the present study that some considered the achievement-level descriptors too broad.

In fact, many students in most states are generally in the two categories, Basic and Below Basic. For example, the analysis of NAEP data showed that 66 % of grade 8 students in US State were classified in the Basic and Below-Basic categories in the 2005 mathematics NAEP assessment. Thus, it appears reasonable to say that these two categories are excessively broad. Establishment of the description for Below Basic could differentiate more accurately among approximately two-thirds of students generally classified as Basic or Below Basic and might correlate more closely with various levels of states' definitions of student performance (e.g., US State has four levels). Such descriptions might be created by describing the knowledge and skills needed to solve the items anchoring on the NAEP scale below the Basic-level cutscore (Pellegrino et al.,

1999). On the other hand, the development of the Below Basic descriptor might distract the emphasis placed on the performance goal set by NAGB in some way that all students ought to reach Proficient or above. Future work is needed to weigh the merits and demerits regarding whether to establish the descriptor of Below Basic.

NCLB requires states to set performance standards for their own statewide tests, and NAEP achievement levels might provide a useful benchmark for state efforts to define their own performance standards. In fact, many states have established performance standards that are similar to NAEP (NAGB, 2000). If the two scales are aligned, then the percentages of students at different levels on statewide tests could be compared with corresponding percentages from state NAEP. This comparison is speculated to validate student performance on state assessments with state NAEP results in terms of meeting standards. Yet, interpretation still needs to be made with caution since the extent of the consistency in the results reported by NAEP and state assessments is also affected by other factors such as what they measure and how they measure it.

In contrast, what happens if the two scales are not aligned? The US State's performance standards are different from NAEP's in terms of standard-setting methodology, the number of achievement levels, and the descriptions. There was consensus among the participants that NAEP achievement-level results cannot be compared with results from the state's assessment. Some argued that equivalent scales should be developed to compare the two test results, if comparisons are necessary. Two participants maintained that NAEP's achievement levels are set higher than US State's. An analysis of NAEP data supports their argument that the state's "Meets" level is closer to NAEP Basic in terms of meeting standards than NAEP Proficient. These findings

suggest that direct comparisons of the results of the two tests do not provide reasonable information for states to use in their decision making process.

There has been controversy over the model, the process, and the end result of setting NAEP achievement levels. Previous NAEP evaluation studies also concluded that the achievement-level setting procedures remain flawed (Glaser & Linn, 1996; Pellegrino et al., 1999). A representative panel of raters established achievement level descriptions for each grade in a specific subject area based on policy definitions defined by NAGB for three achievement levels (Basic, Proficient, and Advanced), which apply to all subjects and all grades assessed by NAEP. The policy definitions for standards are then translated into content-specific descriptions to guide the formal standard-setting process. Finally, cut-scores between the achievement level categories are estimated, by estimating the probability that a hypothetical student at the boundary of each achievement level will get an item correct. The process of translating the standards implied by the policy definitions onto the NAEP reporting scale for a particular content area is called the achievement levels setting process (Reckase, 1998).

A chief policy officer suggested that NAEP results should not be reported by achievement levels since the achievement-level results of the state's assessment and NAEP confuse the public due to the incomparability of NAEP achievement levels to the state's performance levels. Although conveying strong opposition to reporting NAEP results by achievement levels, his argument appears to somewhat reflect the conclusions made by the recent NAEP evaluations. In response to criticisms associated with the achievement-levels from reviewers, NAGB has continued to make refinements focusing on the training of the panelists since achievement-level descriptions and cut-scores set for

each achievement level depend primarily on the judgments of panelists (NAGB, 2000). However, NAGB still employs the same methodology for the achievement level setting and further discussion of the controversy surrounding various aspects of the NAEP achievement levels issue is beyond the purpose and scope of this study. In addition, there is no indication that NAGB has reflected the changing composition of student populations and the resultant diversity of cultures when revising NAEP performance standards. The reason for that might stem from NAGB's belief that *all* students must reach the Proficient level or over, regardless of race/ethnicity and/or SES.

State-to-State Comparisons

A chief policy officer and the NAEP state coordinator viewed state-to-state comparisons as being valuable since such comparisons provide insight into how US State is doing compared to other states, but they did not indicate that state comparisons had had any actual impact on state policy. One of essential benefits of state NAEP is that it permits states to compare to national trends and each other (Mullis, 2003). Unlike the purpose of State NAEP, however, almost all participants expressed concerns about state-to-state comparisons, arguing that state comparisons are misleading since NAEP results do not reflect variances in state curriculum standards and student backgrounds. Also, another concern was raised that state comparisons might promote a misunderstanding of NAEP results by the media.

Demographic characteristics of students are primary factors that have a substantial impact on student performance, but this information is not taken into consideration in state comparisons. NAEP reports have typically presented unadjusted differences among

population groups without attempting to adjust them for dissimilarities in social context (Berends & Koretz, 1995/1996; Raudenbush et al., 1998). That is, state NAEP comparisons are made based on simple, unadjusted cross-sectional differences among states. Most participants shared with other researchers the notion that reporting only unadjusted differences among population groups without accounting for student demographics may be misleading (Berends & Koretz, 1995/1996; Raudenbush et al., 1998).

NAEP has provided unadjusted scores to send states a strong message about the need for improving education. Some have argued that unadjusted test results simply exclude states with low-income students and those with a large number of immigrant and ethnic minority students from approval and provide little insight into ways in which policy changes might produce better performance (Raudenbush, Fotiu, & Cheong, 1998). Yet, NAGB has believed that adjusting for differences in social context might send an unacceptable message about educational standards since educators should not establish low expectations for more disadvantaged states (Berends & Koretz, 1995/1996; Raudenbush et al., 1998). Similarly, it seems evident that there has been no change in the NAEP standards as a result of the increasing growth and changing composition of the ethnic minority population.

However, simply devaluing state comparisons is not reasonable while attempting to attribute poor performance only to student demographics. It muse be kept in mind that state means reflect an unknown mix of contributions from student demographics, school organization and process, and state policy (Raudenbush et al., 1998). Clearly, differences in state means at least partially reflect differences in policy and practice among states.

For example, an analysis of NAEP data for US State revealed the possible positive relationship between performance in mathematics and high emphasis on such content areas as algebra and probability & statistics in instruction. Further, state comparisons made after controlling for the social and economic factors could be useful in examining key policy-relevant predictors that might have led to improvement in achievement in other states.

An education program specialist raised concerns that state-to-state comparisons promote a misunderstanding of NAEP results by the press and argued that such misunderstandings are due in part to the media's lack of knowledge of statistics. She added that the media's misunderstanding further causes more serious problems since it is passed on to the public. Her concern is somewhat consistent with the findings of previous studies that press writers are sometimes unlikely to interpret NAEP results properly (Glaser & Linn, 1996; Koretz & Deibert, 1995/1996), partly because of their lack of knowledge about statistics (Glaser & Linn, 1996; Hambleton & Slater, 1997). Koretz and Deibert (1995/1996) investigated the extent to which the reporting of the 1990 NAEP mathematics assessment in terms of achievement levels successfully communicated patterns of student performance to the print media. They found that widespread simplifications and frequent misinterpretations of NAEP results were made by the press. For example, many articles reviewed confused *p* values (percentage answering each item correctly) and the percentage of students reaching the achievement levels. Hambleton and Slater (1997) examined the extent to which NAEP reports are understandable to policymakers, educators, and the media. They found that misunderstandings of NAEP results were common because most were unfamiliar with the NAEP reporting scale and

had a limited knowledge of statistics. For example, the study found that statistical significance had been commonly interpreted as substantial significance. In other words, statistically significant differences were perceived to be big and important differences by NAEP users.

In particular, a mathematics specialist contended that one of the important aspects of an assessment is to clearly communicate what it is designed to measure and not to measure since assessment information can be inappropriately used without an understanding of the assessment itself. Her argument has implications for use of NAEP data because NAEP results are often interpreted beyond the data and the design used to generate them particularly when attempting to explain poor performance (Pellegrino, 1999). For example, Glaser and Linn of the National Academy of Education established a panel and conducted a series of independent evaluations of the TSA (Trial State Assessment) for 1990, 1992, and 1994 assessments to investigate quality and utility of NAEP results, which was mandated by Congress. In its fourth report (1996), the panel presented recommendations and findings specific to its evaluation of the 1994 TSA in fourth-grade reading. As part of the study, the panel reviewed NAEP-related articles included in the 50 most read newspapers in the nation and found that the press tended to draw invalid conclusions from NAEP results. For example, the study revealed that poor performance in California was described by newspaper reporters as a result of several factors including overcrowded classrooms, too little spending on education, and the state's whole-language reading curriculum, but they were not the variables collected by NAEP. Clearly, these misinterpretations are assumed to be due to users' limited knowledge about what NAEP measures, its purposes, and its uses.

It is likely that these findings also have some implications for the interpretation of state assessment results. Reports on test scores in the media often draw oversimplified conclusions especially when showing how poorly students or schools are performing (Koretz & Deibert, 1995/1996). Usually, the misinterpretation of test scores by the press is assumed to result from their lack of basic and fundamental information about the test it self. What was the test designed to measure? What was the test designed not to measure? What purpose(s) are the test results designed to serve? What criteria were used to determine proficiency in relation to performance on state assessments? Such questions could help the media interpret test results in perspective. Therefore, explicit reporting of this kind of information needs to be included in press briefing packages so that the media can better understand the purposes of state assessments and state test scores and in turn, the amount of misinterpretation of state test results by the press could be reduced.

Recently, NCES has placed a page for the media on the NAEP website possibly to facilitate their understanding of NAEP and correct interpretations of NAEP results. Further improvements in the communication of results possibly could be made by NCES. For example, NCES may provide caution on the media page that state comparisons should be interpreted in context (e.g., curriculum consideration, student demographics). NCES might also provide example explanations of NAEP results in a way that presents both correct and incorrect interpretations. In addition, on the media page NCES needs to point out the difficulties in making causal inferences from correlational data such as NAEP, especially given the absence of a measure of prior achievement in NAEP (Raudenbush et al., 1998).

NAEP Data Use by State Education Personnel

State education personnel interviewed expressed some concerns about the use of NAEP particularly for the purposes of accountability, state-to-state comparisons, and comparisons of state test results to NAEP results. Nonetheless, this study found that they had used NAEP data for state comparisons especially to see how their students are doing compared to other states with similar resources or student characteristics. They also attempted to compare their state test results with NAEP results by linking the two tests in order to provide evidence of concurrent validity of the state assessment. Overall, it is very likely that with the potential benefits of NAEP data use in mind, US State has attempted to make use of NAEP the state has available in the context of the state to inform educational decision-making. The following discussion is of the use of NAEP data by the state education personnel and the overall interpretation of the findings related to their use.

This study found that state education personnel in US State obtained NAEP information primarily from the NAEP state coordinator and the NAEP website, and sometimes from the press release, NAEP training sessions, NAEP meetings, federal associations, or a federal NAEP ambassador. This study also revealed that NAEP data had been utilized in US State to: 1) track student performance over time; 2) make state-to-state and state-to-the nation comparisons; 3) revise the state's standards; 4) discuss relationships between achievement and background factors; and 5) link state test results to NAEP.

A main use by the SEA was to provide NAEP information to stakeholders across the state as a barometer of students' achievement and their progress over time, which is consistent with the primary purpose of NAEP. Interview data revealed that the

information was disseminated though presentations, workshops, and publications developed by the education agency. The evidence from interviews was mostly confirmed by document analysis, and documents and the agency's website provided this use in considerable detail.

This use was consistent with the findings of a study by Bullock and DeStefano (1998). Their study examined the usefulness of results from the 1992 Trial State Assessment (TSA) in reading through interviews of state directors of assessment. The study also focused on participants' perceptions about the credibility and usefulness of various components of the TSA including the reading framework, the achievement-level descriptors, reporting, and dissemination. The study found that the TSA results in reading were used mainly to provide information to the general public, teachers, and administrators and to inform policymakers at state and district levels.

Interestingly, the findings of this study indicate that dissemination of NAEP information targeted at classroom teachers is relatively limited at the SEA. For example, according to the chief policy officer, the SEA held workshops to share NAEP results primarily with principals and superintendents rather than with teachers, and workshops for teachers tended to be more about the state assessment. Teachers play a central role in educating students, but are generally believed to know little about NAEP, its purposes, and its uses. Thus, there appears to be a need to provide them with information about what NAEP measures and how the state's assessment relates to NAEP, and strategies that teachers can use with NAEP data in classrooms need to be offered. The SEA may engage professional organizations of teachers to promote an understanding of NAEP and its use in cooperation with the state NAEP coordinator and subject specialists. Through

workshops, for instance, the agency can provide information on how to use NAEP information (e.g., NAEP Questions Tool) to classroom teachers.

This study found that another main use by US State was to consider NAEP frameworks when revising the state's standards. NAEP frameworks are believed to reflect a balance between current research findings and reform recommendations and current practice. Then, some questions need to be addressed. How do NAEP users perceive the NAEP frameworks? In what ways are the NAEP frameworks actually used? Limited research has been conducted to address these issues. Most participants remarked that the state revised its standards based primarily on NAEP frameworks. Interviews and documents indicate that those revisions were made to ensure that the state standards were rigorous enough and included the content measured by the NAEP assessment. In particular, relevant documents stated that the alignment is important since NAEP is the basis for making state comparisons. Although the extent of alignment between the two was not exactly identified, interview data and document analysis suggest that the alignment is not a close one in general. Unfortunately, this study did not specifically probe into why US State wants to refer to relevant NAEP frameworks in revising the state's standards. Why does US State attempt to make the partial alignment between the two standards rather than a tight one? Future research needs to be pursued to answer these questions.

It was also found that US State used NAEP data primarily to compare its performance with the nation but that comparisons with other states were rare. Some use for state comparisons identified in interviews was not corroborated by document analysis. Even comparisons to like-states such as State A indicated in interviews were not supported

through document analysis. This inconsistency might be due in part to the fact that some of this use was not recorded or occurred simply to understand issues.

Very limited use of state-to-state comparisons in US State is actually inconsistent with the findings of the study by Bullock and DeStefano (1998). The study explored the usefulness of the 1992 TSA in reading for state assessment directors and found that they viewed state comparisons useful and that TSA results served as a barometer of how the states were performing comparatively. Their findings might be attributed in part to the fact that participants for the study did not have a chance to identify disadvantages and limitations addressed in the current study regarding state comparisons at the time of interviews (in late1994 through early 1995) because the 1992 TSA was the first NAEP assessment for which reading results were provided at the state level. In fact, there exists almost eight years of distance between their study and the present study where changes in design and reporting for NAEP occurred.

Lastly, this study found that their use of the NAEP website is limited in general. For the reasons for the rare use, several commented that NAEP information is not needed in doing their jobs. Some felt that only staff members engaging in the development or revision of the state standards need to use NAEP information. These perceptions indicate that state education personnel might have relatively limited knowledge about NAEP, its use, and the NAEP website. For example, when asked about whether NAEP achievement-level descriptors are consistent with US State's, most participants responded that they are either quite similar or inconsistent. Yet, it was discovered that US State had established cut-scores for each of its performance levels for each grade in a subject, but that no achievement-level descriptions had been developed. Therefore, to facilitate use of

NAEP information provided on the website, it is important that NCES needs to make greater effort to advertise it to state education personnel especially in terms of what information is available on the site and what information NAEP provides to states that is superior to the information they already receive from state assessments.

Taken together, the use of NAEP by US State is considered to be limited in relation to its potential utility identified through the analysis of the NAEP website. The primary use of NAEP data by the state education personnel were focused on providing NAEP information to educators, legislators, administrators, and the general public across the state and considering NAEP frameworks in revising the state's standards. The state's reporting of NAEP results was mostly descriptive concerning how students are making progress over time, and little attempt was made to interpret the data in a way to relate student performance to educational or policy factors. In fact, their actual use of NAEP was mainly for the alignment purpose and NAEP data tended to be viewed by most of them as something to be referred to by curriculum specialists when revising the relevant state standards.

It was speculated that easier access to NAEP data through the Internet and the NCLB requirement for states might facilitate the use of NAEP data at the state level. The findings of this study suggest that the NCLB has affected the use of NAEP data more than the improved availability of NAEP data. For example, this study found that the use of the NAEP website by the state education personnel is very limited and that they have limited knowledge about what information and resources are available on the site. In contrary, interview data and document analysis revealed that the SEA's policy to revise the state's standards somewhat in alignment with NAEP frameworks had been

implemented since the legislation of NCLB. More information on NAEP has been also provided in the state's Annual State Report Cards required by law, which report academic performance on statewide assessments along with other indicators of achievement. In addition, the state's attempt to explore possible associations between achievement and context factors was also made in the NCLB system. Thus, although much information does not exist on how NAEP was used before NCLB, it appears reasonable to conclude that NAEP data have been utilized under NCLB in US State more than ever.

There appear to be several possible reasons for the limited use of NAEP by the state education personnel. Firstly, they do not necessarily have an in-depth understanding of the complex NAEP program including what information NAEP provides specifically and the subsequent use of NAEP data. For instance, some participants were found to hold misconceptions about NAEP sampling methodology or reporting of state NAEP results. In a broader sense, it might be possible that some have not necessarily developed comprehensive assessment literacy essential for fully understanding various aspects of the national assessment. Assessment literacy refers to having an adequate amount of information and understanding about how student learning is assessed and tested. Secondly, the SEA staff might have limited knowledge about statistics. For example, some simply argued that NAEP data are not useful because of their basis on a small sample of students, which overlooked the fact that the small sample of students represents the entire population statistically. Thirdly, sometimes NCES does not appear to provide enough of relevant information along with NAEP results in reporting to facilitate correct interpretations of the results. For example, NCES provides neither in reporting nor on the NAEP website detailed information on NAEP's achievement-level setting procedures

essential for state education personnel to accurately interpret NAEP results. Lastly, communication between NCES and the SEA may not be fully made. For example, this study found that although most participants knew about the existence of the NAEP website, they were not necessarily knowledgeable about what information is available on the site and how to use on-line tools provided in that system.

It is desirable that states make better use of NAEP data provided by the federal government to improve any aspect of the quality of education. First of all, to facilitate use of NAEP data by state education personnel, there appears to be a need for better communication between NAEP and the SEA. According to a project officer for NAEP state coordinators at NCES, the coordinators' major responsibilities are to promote an understanding of NAEP among different audiences in their state and to coordinate the NAEP assessment within the state. Their additional task is to enhance states' capacity to use NAEP data by analyzing NAEP data themselves. Therefore, there might be a need for NCES to encourage the NAEP state coordinator to more aggressively promote an in-depth understanding of NAEP among state education personnel. For example, NAEP state coordinators might attempt to find out what information their colleagues need in doing their jobs and then make suggestions on what aspects of NAEP could be informative. They may also make a presentation of a tour of the NAEP website from a state perspective to show their colleagues what is available and not available on the site and how to use web-based tools to search for the information needed. On the part of state education personnel, this study suggests that they need to better understand diverse aspects of NAEP and then try to benefit most from the information it provides in an effort to improve education as a whole in the state.

Implications for Future Research

This study found that NAEP information provided on the NAEP website has a potential to be used to inform state educational decisions in several ways, but that actual use by state education personnel was limited primarily to the provision of NAEP information to educators, administrators, and the public across the state and the partial alignment of the state standards to NAEP frameworks. The results of this study have some implications for future research regarding state education personnel's perceptions of the usefulness of NAEP and the subsequent utility of NAEP data.

All participants contended that the SEA had considered NAEP frameworks when revising the state standards in an effort to ensure that the standards are rigorous enough. Their argument was supported by document analysis. The next question to be addressed is: How rigorous do they find NAEP frameworks? Unfortunately, this study did not probe deeply into this issue. Future research should examine how rigorous state education personnel consider NAEP frameworks and how they define "rigorous enough." In addition, it is assumed that in some states their standards are to some extent aligned with the NAEP frameworks, while other states might develop or revise their standards in close alignment with the frameworks. Future work needs to be pursued regarding why the degree of the alignment might be different from state to state.

State NAEP is intended to provide information about how states are performing in comparison to other states and the nation. This study found that most of state education personnel in US State did not consider state-to-state comparisons meaningful. US State has stayed in the middle among states in terms of state NAEP performance, and its state

standards are not closely aligned with NAEP frameworks. Thus, research needs to be extended to investigate how state education personnel in higher-ranking and lower-ranking states view the usefulness of state-to-state comparisons. Similarly, future research needs to explore how state education personnel in states, where their own standards are tightly aligned with NAEP frameworks, perceive the usefulness of such comparisons. Further, in those states comparisons of student performance on NAEP and state tests are not in accord, how do state education personnel interpret the discrepancy?

An analysis of NAEP data revealed that the general trend of student achievement on the US State's assessment in the 8th-grade mathematics was likely consistent with that on NAEP as a whole in terms of meeting standards, but indicated some discrepancy for the trends in performance gaps between white and minority students. These findings suggest that the inconsistency might be due in part to the differences between the two testing programs in achievement-level setting methods. Thus, further research is needed to explore how differences in the standard-setting methods affect the estimation of the percentages of students reaching each of achievement levels.

It is speculated that if statewide assessments and state NAEP tests measure similar content using similar methods and use similar performance standards to report results, then the results of the two tests should be similar (NAGB, 2000). This speculation needs to be empirically verified. When states have developed achievement-level descriptors similar to those adopted by NAEP and established cut-scores using NAEP-like standard-setting methods, how comparable are the two test results? Future research needs to explore how comparable the two test results are in terms of meeting standards and how state education personnel in those states perceive the use of the comparisons.

This study found that the use of NAEP by US State is primarily focused on examining the progress of its student performance over time and considering NAEP frameworks when revising the state standards. Logically, the next question to be addressed is: “What factors influence the use of NAEP data in a state context?” Specifically, this study was not designed to identify those factors. Future work is needed to identify major factors that might facilitate or impede the use of NAEP data at the state level.

Limitations of the Study

This study included a case study of a SEA’s use of NAEP to investigate the usefulness and subsequent use of NAEP by state education personnel. Seven of the state education personnel in different positions at the agency volunteered to participate in the study. Interviews with the participants allowed an in-depth examination of their perceptions of the usefulness of NAEP and the use of NAEP data in their decision making process.

However, this study is limited in that a variety of staff members were not included in the study. For example, subject specialists participating in the study were focused on mandatory NAEP subject areas under NCLB such as mathematics and reading. In addition, both curriculum and assessment specialists within a subject, who were assumed to have somewhat different views on same issues, were not interviewed for this study. Psychometricians at the SEA might provide somewhat different perspectives with expertise in the technical aspects of NAEP. Moreover, the state director of assessment in charge of directing assessments in the state did not participate in the study.

Another limitation to the study was associated with data collection. When the long interview neared the end, questions about some aspects of NAEP where improvement had been suggested by previous NAEP evaluation studies were asked, such as inclusion of students in special needs, integration of NAEP results with other indicators, reporting data on diverse elements of knowledge and skills, and using the web's multimedia features. Due to time constraints, however, each interviewee was asked only two questions on average, thus leading to collecting limited views of the participants on each of those topics. Moreover, the responses collected were found to be relatively limited probably because of their lack of knowledge about those issues or because of fatigue resulting from spending too much energy on previous questions, or both.

Also, there may be some limitation regarding the content analysis of the NAEP website. An analysis of the website was conduced from October 2005 through May 2006 and interview protocols were developed in part based on the site analysis. Therefore, the analysis might not have included the content that was uploaded after then, although the website was visited throughout the study whenever needed.

Data collection, analysis, and interpretation were done only by the researcher, who was the primary instrument of this qualitative study. It is inevitable that some aspects of her experience, beliefs, and biases might have influenced the collection, analysis, and interpretation of the data overall. For qualitative inquiry, triangulation serves to clarify meaning by identifying different ways that the case is being viewed (Stake, 2005). Therefore, to minimize any introduction of the researcher's bias to the study, she did continual triangulation of the data and carried out several rounds of analysis throughout the study. In addition, personal characteristics about the researcher presented in Chapter

III could help readers understand the experience, training, and perspective she brought to the issues under study.

Another limitation to this study was related to the transcription of interviews. Because the researcher's first language is not English, she listened very carefully to audiotaped interviews over and over again not to miss or misunderstand any word when making transcriptions. However, although she tried her best to transcribe recordings without any misunderstanding, it is possible that she might not correctly have understood some of words or phrases included in them.

Lastly, this case study was conducted focusing on state education personnel in one state. US State is not representative of the states across the nation in terms of student demographics, geography, expenditures for education, state testing programs, the status of curricular reform, the educational policy climate, etc. Although the findings of this case study reported here cannot be generalizable to other contexts, they might have broader relevance and the potential to be applied elsewhere. The researcher anticipates that readers might interpret the findings in the context of their own educational settings.

REFERENCES

Allen, N., Carlson, J., & Zelenak, C. (1999). *The 1996 NAEP technical report*. Washington, DC: National Center for Education Statistics.

Allen, N, Donoghue, J., & Schoeps, T. (2001). *The 1998 NAEP technical report.* Washington, DC: National Center for Education Statistics.

Anspach, R. (1987). Prognostic conflict in life-and-death decisions: The organization as an ecology of knowledge. *Journal of Health and Social Behavior, 28,* 215-231.

Baker, E. (1995/1996). Introduction: Policy and technical contexts of National Assessment of Educational Progress validity studies. *Educational Assessment, 3*, 1-8.

Baker, E. & Linn, R. (1997). *Emerging educational standards of performance in the United States*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Barton, P. & Coley, R. (1998). *Growth in school: Achievement gains from the fourth to the eighth grade*. Princeton, NJ: Educational Testing Service.

Beaton, A., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*, 95-109.

Berends, M., & Koretz, D. (1995/1996). Reporting minority students' test scores: How well can the National Assessment of Educational Progress account for differences in social context? *Educational Assessment, 3*, 249-285.

Blank, R., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science*. Washington, DC: Council of Chief State School Officers.

Bourque, M. (2004). A history of the National Assessment Governing Board. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 201-231). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Bullock, C. & DeStefano, L. (1998). A study of the utility of results from the 1992 trial state assessment (TSA) in reading for state-level administrators of assessment. *Educational Evaluation and Policy Analysis, 20,* 47-51.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E., & Harris, E. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3*, 9-51.

Chudowsky, N. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42,* 75-83.

Crotty, M. (1998). *The foundations of social research: meaning and perspective in the research process*. London: Sage.

Denzin, N & Lincoln, Y (2005). Introduction: The Discipline and Practice of Qualitative Research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp.). Thousand Oaks, CA: Sage Publications.

Elliott, E. & Phillips, G. (2004). A view from the NCES. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 233-249). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Epstein, J. (1998). *Power, politics and the National Assessment of Educational Progress (education policy).* Unpublished doctoral dissertation. Rutgers, the State University of New Jersey.

Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education, 10,* 145-159.

Eschenfelder, K, & Miller, C. (2005). *The openness of government websites: Toward a socio-technical government website evaluation toolkit*. Paper presented at the MacArthur Foundation Internet Credibility and the User Symposium, Seattle, WA.

Fitzharris, L. (1993). *An historical review of the National Assessment of Educational Progress from 1963 to 1991 (assessment programs, educational assessment).* Unpublished doctoral dissertation. University of South Carolina.

Fontana, A & Frey, J. (2005). The Interview: From Neutral Stance to Political Involvement. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp.). Thousand Oaks, CA: Sage Publications.

Forbes, R. (1977). NAEP: One "tool" to improve instruction. *Educational Leadership, 34*, 276-281.

Glaser, R, Linn, R., Bohrnstedt, G. (1996). *Quality and Utility: The 1994 Trial State Assessment in Reading***.** The Fourth Report of the National Academy of Education Panel on the Evaluation of the NAEP Trial State Assessment: 1994 Trial State Assessment in Reading. Stanford, CA: National Academy of Education.

Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA: National Academy of Education.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: Rand.

Hambleton., R., & Cadman, S. (1992). NAEP state reports in mathematics. *New England Journal of Public Policy, 10*, 209-222.

Hambleton, R. & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (Center for the Study of Evaluation Rep. No. 430). Los Angeles, LA: University of California, Los Angeles.

Herman, J. (1997). *Large-scale assessment in support of school reform: Lessons in the search for alternative measures.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 435631)

Hert, C., Eschenfelder, K., McClure, C., Rubin, J., Taffet, M., Abend, J., & Pimentel, D. (1999). *Evaluation of selected websites at the U.S. Department of Education: Increasing access to web-based resources*. Syracuse, NY: Syracuse University, School of Information Studies. (ERIC Document Reproduction Service No.ED460681)

Hodder, I. (2000). The interpretation of documents and material culture. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 703-715). Thousand Oaks, CA: Sage Publications.

Hombo, C. (2003). NAEP and No Child Left Behind: Technical challenges and practical solutions. *Theory into Practice, 42*, 59-65.

Jaeger, R. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics.

Jones, L. (1996). A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher, 25*, 15-22.

Johnson, E. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics, 14*, 303-334.

Johnson, E. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 2*, 95-110.

Johnson, E., Cohen, J., Chen, W., Jiang, T., & Zhang, Y. (2005). *2000 NAEP -- 1999 TIMSS linking report*. Washington, DC: National Center for Education Statistics.

Kifer, E. (2001). *Large-scale assessment: Dimensions, dilemmas, and policy*. Thousand Oaks, CA: Corwin Press.

Koretz, D. (1988). Arriving in Lake Woebegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator: The Professional Journal of the American Federation of Teacher, 12,* p8-15.

Koretz, D. (1991). State comparisons using NAEP: Large costs, disappointing benefits. *Educational Researcher, 20,* 19-21.

Koretz, D. (1992). NAEP and national testing: Issues and implications for educators. *NASSP Bulletin, 76,* 30-40.

Koretz, D., & Deibert, E. (1995/1996). Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educational Assessment, 3,* 53-81.

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED378220)

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED462410)

Krathwohl, D. (1998). *Methods of educational & social science research* (2nd ed.). New York, New York: Longman.

Lazer, S. (2004). Innovations in instrumentation and dissemination. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 469-487). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Lehmann, I. (2004). The genesis of NAEP. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 25-92). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Linn, R. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Linn, R. (1993). *Assessment and accountability*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R., Baker, E., & Betebenner, D. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31,* 3-16.

Linn, R., & Dunbar, S. (1992). Issues in the design and reporting of the National Assessment of Educational progress. *Journal of Educational Measurement, 29*, 177-194.

Linn, R., & Kiplinger, V. (1995). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education, 8,* 135-155.

Linn, R., Koretz, D., & Baker, E. (1996). *Assessing the validity of the National Assessment of Educational Progress: NAEP technical review panel white paper*. (Center for the Study of Evaluation Rep. No. 416). Los Angeles, LA: University of California, Los Angeles.

Madaus, G. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan, 66,* 611-617.

Marlaire, C. & Maynard, D. (1990). Standardized testing as an interactional phenomenon. *Sociology of Education, 63,* 83-101.

Martin, W. (2004). NAEP from three different perspectives. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 309-328). Bloomington, IN: Phi Delta Kappa Educational Foundation.

McDonnell, L. (1994). *Policymakers' views of student assessment*. RAND Institute on Education and Training, prepared for Office of Educational Research and Improvement, U.S. Department of Education. Santa Monica, CA: RAND.

McLaughlin, D. (2000). *NAEP-state reading assessment linkage study: Report to US State*. Washington, DC: American Institutes for Research.

McLaughlin, D. (2001). *Study of the linkages of 1996 NAEP and state mathematics assessments in four states*. Washington, DC: National Center for Education Statistics.

McLaughlin, M., & Shepard, L. (1995). *Improving education through standards-based reform*. Stanford, CA: National Academy of Education.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1990). *Validity of test interpretation and use*. Princeton, NJ: Education Testing Service, Policy Information Center.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Miller, G. (1997) Contextualizing texts: Studying organizational texts. In G. Miller & R. Dingwall (Eds.), *Context and method in qualitative research* (pp. 77-91). London: Sage Publications.

Mislevy, R. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Education Testing Service, Policy Information Center.

Mislevy, R. (2002). *Psychometric principles in student assessment*. (Center for the Study of Evaluation Rep. No. 583). Los Angeles, LA: University of California, Los Angeles.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131-154.

Mosher, F. (2004). What NAEP really could do. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 329-340). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Mullis, I. (1992). Developing the NAEP content-area frameworks and innovative assessment methods in the 1992 assessments of mathematics, reading, and writing. *Journal of Educational Measurement, 29,* 111-131.

Mullis, I. (2003). *Optimizing state NAEP: Issues and possible improvements: NAEP validity studies*. Washington, DC: National Center for Education Statistics. (ERIC Document Reproduction Service No. ED478427)

National Assessment Governing Board (1995). *Science framework for the 1996 and 2000 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board (2000). *Student performance standards on the National Assessment of Educational Progress: Affirmation and Improvements*. Washington, DC: Author.

National Assessment Governing Board (2002a). *Mathematics framework for the 2003 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board (2002b). *Using the National Assessment of Educational Progress to confirm state test results*. The Ad Hoc Committee on Confirming Test Results. Washington, DC: Author.

National Assessment Governing Board (2003). *Background information framework for the National Assessment of Educational Progress*. Washington, DC: Author.

National Center for Education Statistics (1994). *Overview of NAEP assessment frameworks*. Washington, DC: Author.

National Center for Education Statistics (1999a). *Directory of NAEP publications.* Washington, DC: Author.

National Center for Education Statistics (1999b). *The NAEP guide: A description of the content and methods of the 1999 and 2000 assessments.* Washington, DC: Author.

National Center for Education Statistics (2003). *NAEP validity studies: An agenda for NAEP validity research*. Washington, DC: Author.

National Center for Education Statistics (2005). *Programs and plans of the National Center for Education Statistics.* Washington, DC: Author.

National Center for Education Statistics (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*. Washington, DC: Author.

National Research Council (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Committee on equivalence and linkage of educational tests. M. Feuer, P. Holland, & B. Green, eds. Commission on Behavioral and Social Science and Education. Washingotn, DC: National Academy Press.

National Research Council (2000). *Reporting district-level NAEP data*. Committee on NAEP reporting practices: Investigating district-level and market-basket reporting. P. DeVito & J. Koenig, eds. Commission on Behavioral and Social Science and Education. Washingotn, DC: National Academy Press.

Olson, L. (2003). NAEP board worries states excluding too many from tests. *Education Week*, 22, 7-7.

Patton, M. (2002). *Qualitative research & evaluation methods*, (3rd ed.). Thousand Oaks, CA: Sage Publications.

Pearson, D., & DeStefano, L. (1993). *Content validity of the 1992 NAEP in reading: Classifying items according to the reading framework.* Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Pellegrino, J., Jones, L., & Mitchell, K. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. (Report of the Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council). Washington, DC: National Academy Press.

Perakyla, A. (2005). Analyzing Talk and Text. In N. Denzin & Y. Lincoln (Eds.), *The SAGE handbook of qualitative research* (pp. 869-886). Thousand Oaks, CA: Sage Publications.

Phillips, G., Mullis, I., Bourque, M., Williams, P., Hambleton, R., Owen, E., & Barton, P. (1993). *Interpreting NAEP scales*. Washington, DC: National Center for Education Statistics.

Podgursky, M. (2002). *NAEP background questions: What can we learn from NAEP about the effect of schools and teachers on student achievement?* Columbia, MO: University of Missouri.

Raudenbush, S., Fotiu, R., & Cheong, Y. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis, 20*, 253-267

Reckase, M. (1998). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 science NAEP process. *Applied Measurement in Education, 11*, 9-21.

Reckase, M. (2002). *Contributions of background questions to improving the precision of NAEP results*. Lansing, MI: Michigan State University.

Resnick, L. (1999). *Reflections on the future of NAEP: Instrument for monitoring or foe accountability?* (Center for the Study of Evaluation Rep. No. 499). Los Angeles, LA: University of California, Los Angeles.

Rothsteln, R. (2006). *Reforms that could help narrow the achievement gap. Policy perspectives*. San Francisco, CA: WestEd. (ERIC Document Reproduction Service No. ED493060)

Rust, K. (2004). Sampling and field operations at Westat, 1983to 2001. In L. Jones & I. Olkin (Eds.), *The nation's report card: Evolution and perspectives* (pp. 427-448). Bloomington, IN: Phi Delta Kappa Educational Foundation.

Schwandt, T. (1994). Constructivist, interpretivist approaches to human inquiry. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (p. 118-137). Thousand Oaks: Sage.

Sears, J. (1992). Researching the other/searching for self: Qualitative research on [homo] sexuality in education. *Theory Into Practice, 31*, 147-156.

Sebring, P. & Boruch, R. (1983). How is National Assessment of Educational Progrss used? Results of an exploratory study. *Educational Measurement: Issues and Practice, 2*,16-20.

Silver, E., Alacaci, C., & Stylianou, D. (2000). Students' performance on extended-constructed-response tasks. In E. Silver & P. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 307-343). Reston, VA: National Council of Teachers of Mathematics.

Silver, E., & Kenney, P. (1993a). *The content and curricular validity of the 1992 NAEP TSA in mathematics*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Silver, E., & Kenney, P. (1993b). An examination of relationships between the 1990 NAEP mathematics items for grade 8 and selected themes from the NCTM standards. *Journal of Research in Mathematics Education, 24*, 159-167.

Smith, M. & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*, 334-344.

Stake, R. (2005). Qualitative Case Studies. In Denzin & Y. Lincoln (Eds.), *The SAGE handbook of qualitative research* (pp. 443-466). Thousand Oaks, CA: Sage Publications.

Swanson, C., & Stevenson, D. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis, 24*, 1-27.

Ungerleider, C. (2003). Large-scale student assessment: Guidelines for policymakers. *International Journal of Testing, 3*, 119-128.

U.S. Department of Education. (1991). *America 2000: An education strategy*. Washington, DC: Author.

US State Department of Education (2004). *NAEP-language and literacy issues: Factors related to student performance*. US State: US State Department of Education, Office of Educational Improvement and Innovation.

US State Department of Education (2005). *Academic Content Standards* (Vol. 1, No. 2). US State: US State Department of Education, Office of Educational Improvement and Innovation.

Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

Whitney, T. (1993). *Voters and school finance: The impact of public opinion*. Denver, CO: National Conference of State Legislatures. (ERIC Document Reproduction Service No. ED355666)

Williams, V., Rosa, K., McLeod, L., Thissen, D., & Sanford, E. (1998). Projecting to the NAEP scale: Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement, 35*, 2777-296.

Wilson, L. & Blank, R. (1999). *Improving mathematics education using results from NAEP and TIMSS*. Washington, DC: Council of Chief State School Officers.

Wise, L. (2003). *The National Assessment of Educational Progress: What it tells educators*. Washington, DC: U.S. Department of Education.

Yin, R. (2003). *Applications of case study research* (2nd ed.). Thousand Oaks, CA: Sage Publications.

APPENDICES

APPENDIX A

Informed Consent Form I

Project Title: The Use of NAEP Data in a State Context
Principal Investigator: Dr. Larry Flick, Department of Science and Mathematics Education

Dear _____:

You are invited to take part in a dissertation research study designed to investigate the nature of the NAEP website and how NAEP data provided on the website are/could be used to inform educational decisions in a state context. This empirical study of NAEP data use at the state level is expected to shed light on an important issue such as the usefulness of NAEP. This consent form gives you the information you will need to help you decide whether to be in the study or not. Please read the form carefully.

You are invited to participate in this study because you are believed to be actively involved in educational issues in the state as a key NAEP user. If you deicide to participate in this study, I will ask you to complete an interview. Phone interviews will be conducted to elicit and probe your perceptions and perspectives on NAEP information and its utility. The questions to be asked include: "How do you currently use NAEP data? and "What factors influence your NAEP data use?" The interview will take approximately 50 minutes to complete and might be audiotaped to help me focus on the actual details of your responses.

There are no foreseeable risks and/or discomforts associated with the procedures described in this study. You will not benefit from being in this study, other than making a valuable contribution to education. You will not be paid for being in this research study. However, we hope that, in the future, other people might benefit from this study because this study will contribute to helping improve the utility of NAEP data and reporting practices for NAEP assessments.

Confidentiality will be maintained through the use of identifying codes rather than names in order to ensure protection of the privacy of participants. The information you provide during this research study will be kept confidential to the extent permitted by law. All data sources will be kept in a locked location at all the times. Publications based on this study will be done in such a way that participants cannot be directly identified. It must be important to note that your participation is voluntary and that you are free to skip any questions you would prefer not to answer and can stop at any time during the study.

If you have any questions about the study, please contact Dr. Larry Flick at Larry.Flick@oregonstate.edu or (541) 737-3664. If you have questions about your rights as a participant, please contact the Oregon State University Institutional Review Board (IRB) Human Protections Administrator, at (541) 737-3437 or by email at IRB@oregonstate.edu.

Thank you for your cooperation.

……

Your signature indicates that this research study has been explained to you, that your questions have been answered, and that you agree to take part in this study. You will receive a copy of this form.

Participant's Name (printed): _____

_____ _____

(Signature of Participant) (Date)

APPENDIX B

Informed Consent Form II

Project Title: The Use of NAEP Data in a State Context
Principal Investigator: Dr. Larry Flick, Department of Science and Mathematics Education

Dear _____:

You are invited to take part in a dissertation research study designed to investigate the nature of the NAEP website and how NAEP data provided on the website are/could be used to inform educational decisions in a state context. This empirical study of NAEP data use at the state level is expected to shed light on an important issue such as the usefulness of NAEP. This consent form gives you the information you will need to help you decide whether to be in the study or not. Please read the form carefully.

You are invited to participate in this study because you are in a position to provide important information on the site content and/or management. If you deicide to participate in this study, I will ask you to complete an interview. Phone interviews will be conducted to gain general background information on information dissemination activities via the website and/or the site management. The questions to be asked include: "What is the process for creating content and getting it on the website?" and "How does NCES identify information needs and information preferences of the site's users?" The interview will take approximately 50 minutes to complete and might be audiotaped to help me focus on the actual details of your responses.

There are no foreseeable risks and/or discomforts associated with the procedures described in this study. You will not benefit from being in this study, other than making a valuable contribution to education. You will not be paid for being in this research study. However, we hope that, in the future, other people might benefit from this study because this study will contribute to helping improve the utility of NAEP data and reporting practices for NAEP assessments.

Confidentiality will be maintained through the use of identifying codes rather than names in order to ensure protection of the privacy of participants. The information you provide during this research study will be kept confidential to the extent permitted by law. All data sources will be kept in a locked location at all the times. Publications based on this study will be done in such a way that participants cannot be directly identified. It must be important to note that your participation is voluntary and that you are free to skip any questions you would prefer not to answer and can stop at any time during the study.

If you have any questions about the study, please contact Dr. Larry Flick at Larry.Flick@oregonstate.edu or (541) 737-3664. If you have questions about your rights as a

participant, please contact the Oregon State University Institutional Review Board (IRB) Human Protections Administrator, at (541) 737-3437 or by email at IRB@oregonstate.edu.

Thank you for your cooperation.

……

Your signature indicates that this research study has been explained to you, that your questions have been answered, and that you agree to take part in this study. You will receive a copy of this form.

Participant's Name (printed): _____

_____ _____

(Signature of Participant) (Date)

APPENDIX C

Interview Protocol (state education personnel)

A. Background Information

1. What is your job title?
2. Please tell me about your primary responsibilities.
3. How many years have you been working at the SEA?
4. How long have you been working in this position?

B. Assessment Purpose

1. What purposes do you expect assessments to serve?
2. What aspects of assessments are important to you in doing your job?

C. NAEP Data Use

1. How familiar are you with NAEP and its results?
2. If your job needs NAEP information, please describe what part of your job needs it?
3. How do you obtain NAEP information?
4. The literature indicates that NAEP (1) provides descriptive information; (2) serves an evaluative function; and (3) provides interpretive information for policy implications.

   (a) How have you used NAEP data? Please specify.

   * Explain that "NAEP data" here include NAEP methods and materials as well

* Check if interviewee has described the below uses:

- to persuade or confirm his/her own beliefs, to understand issues, or to recognize problems:
- to provide information to the general public, teachers, legislators, administrators, or colleagues:

- to inform policy decisions at the state level:
- to inform educational programs at the state level:
- any other uses?

(b) Have you used NAEP data for accountability purposes? Please describe these uses.

(c) Have NAEP data ever been used to promote or support specific legislative actions? If so, please describe these uses.

5. NAEP presents disaggregated results by gender, race/ethnicity, SES, and a small number of additional variables to provide an interpretive context for NAEP data.

   (a) How policy-relevant do you find these data?

   * Interviewer might ask:

   How do you use these disaggregated NAEP data?

   (b) How would you think NAEP could provide better policy implications of NAEP results to better find out possible sources of disappointing or promising performance?

6. State-to-state comparisons are provided for State NAEP.

   (a) How useful do you find state-to-state comparisons?

   (b) Do you have any concern about these comparisons? If so, what concern?

   * Interviewer might ask:

   - Do you think NAEP can be used to confirm trends in state performance on the state assessment?
   - What do you think about ranking order states based on state NAEP results

7. Would it be useful to you to have NAEP data for comparisons at the district or school level?

8. What would you think about the relationship between NAEP and US State assessment programs?

   * Interviewer might ask:

What if there is a discrepancy between NAEP and state test results?

9. NCLB encourages states to verify state gains with trends on NAEP. If NAEP might be used to audit state measures of yearly educational progress under NCLB, how would you react?

9. What are your suggestions for maximizing utility of NAEP information from a state perspective?

* Interviewer might ask:

What factors, if any, make you resistant to using NAEP information?

D. Standards-based Reporting (try this section, if time permits)

NAEP provides information about what students *should* know and be able to do in an effort to make NAEP results more useful, in addition to about what students know and can do. There are three achievement levels for each grade assessed by NAEP (4, 8, and 12): *Basic*, *Proficient*, and *Advanced*.

1. How do you feel about NAEP performance standards?

2. How easy to understand do you find the achievement-level descriptions provided with NAEP results to aid in interpretation? (show an example of NAEP achievement-level descriptions, if needed)

3. How do you use these descriptors?

4. (a) Are the NAEP achievement-level descriptors consistent with the achievement levels defined by US State?

* NAEP has both grade-specific and subject-specific definitions of performance levels. Definition of performance levels & standard-setting methods

(b) Is the percentage of students reaching the proficient level on NAEP consistent with that on the state assessment? If not, why not?

E. The NAEP Website

1. Have you ever used the NAEP website? (if not, ask why not and then skip to #6)

2. (a) Do you find the site engaging and stimulating?

* Interviewer might ask about: the quality of data on the site and interpretation of what you get from the site

(b) How easy to use do you find the site?

3. (a) What NAEP product(s) provided on the site have you used?

   * If not mentioned, then ask:

   - Have you used NAEP Questions Tool? If so, how?
   - NAEP Data Explorer? If so, how?
   - State Profiles? If so, how?
   - Item Maps? If so, how?
   - NAEP frameworks?

   (b) How useful do you find them?

4. (a) What areas of the website do you most often select?

   (b) Please explain why?

5. I'd like to ask you about what questions you expect to be answered through navigation of the website from the state perspective. Please describe as many as you can. (If no on #1, what do you want to see on the site?).

* If not mentioned, then interviewer might ask about:

- measurement and reporting of the various facets of achievement

  * Do you think NAEP needs to provide data on diverse aspects of achievement (e.g., problem solving)?

- integration of NAEP results with education inputs at the student, classroom, school, and state level from non-NAEP sources
- inclusion of students with disabilities and English-language learners
- information that takes advantage of the web's interactive and multimedia features
- use of technology in NAEP

F. Information Dissemination (w/ NAEP coordinator)

1. What is the protocol in your state for developing NAEP reports?

2. What is the protocol in your state for communicating results?
3. What is the protocol in your state for briefing the press?

4. How does NCES assist in producing your state's own reports?

5. Any suggestions regarding NAEP information dissemination?

G. Demands for Information

1. (a) Have you made informal requests for information through the NAEP site? If yes, how? If not, why not?

   (b) What was the outcome of those requests?

   (c) If no, whom would you contact if you had questions regarding NAEP information?

H. Other Issues

Please make any final comments on NAEP data use at the state level, if anything?

APPENDIX D

Interview Protocol (NAEP webmaster)

A. Background Information

1. What is your job title?
2. Please briefly tell me about your job.
3. How many years have you been working at NCES?
4. How many years have you been working in this position?

B. Website Purpose & Audience

1. What is the purpose(s) of the NAEP website?
2. What are the audiences intended for the site?

C. The Site Content & Management

1. What is the process for creating content?
2. What is the process for getting it on the website?
3. Tell me about the basic organizational structure associated with the website content and management:

   (a) Who is involved in content development?

   (b) Who is involved in the site management?

   (c) How do people involved in the site content & management communicate with one another?

   (d) Please describe an organizational diagram of the hierarchy of each job title to others in website development and management?

4. (a) Who is responsible for reviewing the website?

   (b) How is the site reviewed?

   (c) Who is responsible for implementing recommended changes?

5. How have the processes surrounding providing content in print changed since the creation of the website?

D. Log & transaction analysis

1. How does NCES regularly analyze data in the web server to better understand the ways in which users use the website?

2. What are most and least requested pages?

E. NAEP Data Use

1. What kinds of expectations does NCES perceive users have about the site?

2. (a) Has NCES assessed the usefulness of the site from the perspective of users?

(b) If so, what aspects of the website? (e.g., content, ease of use, navigability, availability & quality of help)

(c) How often does NCES review the usefulness of the site?

3. How does NCES identify information needs and information preferences of the site users?

4. How does NCES actively solicit the input of states to better support them in terms of NAEP data use?

F. User Feedback & Requests

1. (a) What mechanisms are in place to receive feedback or requests from users of the site?

(b) How does NCES facilitate user feedback?

2. Please describe the process of dealing with feedback.

(a) Who receives the feedback?

(b) How do they evaluate it?

(c) How do good ideas feed into the content provision process?

G. Evaluation Processes

1. What processes are in place to evaluate the success of the website on a regular basis?

2. What do you feel are critical factors leading to the success of the site?

H. Other

Are there other aspects of website planning, management, and evaluation that we should know about?

APPENDIX E

Interview Protocol
(project officer for state NAEP coordinators)

A. Background Information

1. What is your job title?
2. Please briefly tell me about your primary responsibilities.
3. How many years have you worked at NCES?
4. How long have you been working in this position?

B. NAEP State Coordinators

1. Please explain why the position for NAEP state coordinators has been established?
2. What is the job of a NAEP state coordinator?
3. How does NCES aid NAEP state coordinators in serving as the liaison between the State Education Agency and NAEP?
4. How does NCES provide ongoing support for NAEP State Coordinators?
5. (a) What is the purpose(s) of the NAEP State Service Center?

   (b) How does the Center function?
6. How would you think state participation and states' use of NAEP data have changed since the establishment of the position of NAEP state coordinators?

C. NAEP Data Use

1. What kinds of expectations does NCES perceive states have about NAEP?
2. How does NCES assist states in using NAEP data and resources?
3. How does NCES identify NAEP information needs and information preferences of the states?
4. How does NCES actively solicit the input of NAEP state coordinators to better

   support states in terms of NAEP data use for policy and/or education decisions?

5. How does NCES respond to ongoing state demands for information?

6. (a) Has NCES conducted research focusing on the usefulness of NAEP data provided on the NAEP website from a perspective of states?

   (b) If yes, what was the outcome? (ask for a copy)

D. Information Dissemination

1. In what ways does NCES involve state departments of education in facilitating accessibility and dissemination of NAEP results within their states to make NAEP data more accessible to intended audiences?

2. How does NCES support states in generating their own reports of NAEP findings customized to their own situations and needs?

E. Other Issues

Please make comments if there are other aspects of state participation, NAEP information dissemination, NAEP data use, etc. that we should be aware of?