## National Assessment Governing Board
### Assessment Development Committee

#### August 1-2, 2013

## AGENDA

### Thursday, August 1, 2013

**Closed Session**

<table>
<thead>
<tr>
<th>Time</th>
<th>Action</th>
<th>Secure material provided under separate cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 am – 1:45 pm</td>
<td><strong>ACTION:</strong> Review of Secure Material for NAEP Assessments:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Technology and Engineering Literacy (TEL)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Lonnie Smith, ETS</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Science Interactive Computer Tasks (ICTs)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Andrew Latham, ETS</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Science Hands-on Tasks (HOTs)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Shu-Kang Chen, ETS</em></td>
<td></td>
</tr>
</tbody>
</table>

### Friday, August 2, 2013

**Closed Session**

<table>
<thead>
<tr>
<th>Time</th>
<th>Action</th>
<th>Secure material provided under separate cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00 – 10:05 am</td>
<td><strong>Welcome, Introductions, and Agenda Overview</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Alan Friedman, ADC Chair</em></td>
<td></td>
</tr>
<tr>
<td>10:05 – 11:45 am</td>
<td><strong>ACTION:</strong> Continued Review of Secure Material for NAEP TEL and Science Assessments</td>
<td></td>
</tr>
<tr>
<td>11:45 am – 12:10 pm</td>
<td>2013 Technology and Engineering Literacy (TEL) Pilot Test: Update and Preliminary Observations</td>
<td>Attachment A</td>
</tr>
<tr>
<td></td>
<td><em>William Ward, NCES</em></td>
<td></td>
</tr>
<tr>
<td>12:10 – 12:30 pm</td>
<td>Update on Reporting Grade 4 Computer-Based Writing Information</td>
<td>Attachment B</td>
</tr>
<tr>
<td></td>
<td><em>Arnold Goldstein, NCES</em></td>
<td></td>
</tr>
</tbody>
</table>

**Information Item**

| NAEP Item Review Schedule | Attachment C |
Technology and Engineering Literacy Pilot Testing

As part of the 2013 NAEP administration, a large field trial was conducted for the Technology and Engineering Literacy (TEL) assessment. The field trial was designed to be similar to other field trials in subjects for which an entirely new framework is used. Specifically, the trial was designed to support both a detailed evaluation of the performance of items and tasks as well as how they relate to each other (e.g., through scaling and correlation-based analyses). This presentation provides preliminary findings from the field trial analyses and presents samples of student responses that exemplify some distinctions in patterns of performance. It concludes with a summary of the timeline for next steps in the TEL administration and reporting.

The analysis of item and task performance targeted two goals:

1. Individual item performance, including response time, to select the discrete items and assemble discrete item blocks for the 2014 assessment. Given the short time frame involved in preparation for the 2014 probe assessment, this analysis has been completed and was based on (observed) item responses.

2. Scaling to evaluate to what extent the relationships between the items and tasks reflect the various constructs defined and hypothesized in the TEL Framework. This involves both the core content domains (Design & Systems, Information & Communication Technology, and Technology & Society) as well as cross-cutting practices (Understanding Technological Systems, Developing Solutions & Achieving Goals, and Communicating & Collaborating). We are approaching this task in two ways: univariate scaling of the domains (completed) and bi-factor modeling of the constellation of domains and practices (for which we coined the term “competencies”).

The field trial also allowed for examination of patterns of performance within tasks, identifying characteristics of tasks that distinguish different levels of performance. This session will present several example solutions to items and tasks, illustrating the elements of performance that distinguish between low performers and high performers. Some interesting variations on the performances within tasks will be presented and discussed.

The session will close with a summary of the timeline for TEL, including the release of a second informational video on TEL, a publically available TEL assessment task, and upcoming dates for the 2014 probe administration and reporting.
Reporting on the 2011 Grade 4 Writing Pilot:

Progress Report

At the May Governing Board meeting, the Assessment Development Committee received an update about the upcoming 2011 Grade 4 Writing Pilot web report, which will disseminate lessons learned from the development and administration of the pilot assessment.

In order to summarize all of the facets of the pilot, NCES is preparing a technical memorandum that provides an overview of the study, its development, the test administration and outcomes. The technical memorandum will be the source for the information on the web and will be available to readers upon request.

The content for the website will be organized by questions of interest, such as:

1. How did we determine the W computer-based assessment (WCBA) platform for fourth graders?
2. How well did fourth graders interact with the computer-based assessment (CBA) for Writing?
3. How well did fourth graders perform on the WCBA?
4. What types of questions were administered to fourth graders on the computer?
5. How were accommodations administered on the computer platform?

The presentation to the committee will discuss some findings from the technical memorandum and will show more mock-ups of the website.
### Assessment Development Committee

**Item Review Schedule**

**May 2013 – December 2013**

(Updated 6/25/13)

<table>
<thead>
<tr>
<th>Review Package to Board</th>
<th>Board Comments to NCES</th>
<th>Survey/ Cognitive</th>
<th>Review Task</th>
<th>Approx Number Items</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>5/02/13</td>
<td>5/23/13</td>
<td>Cognitive</td>
<td>2015 Pilot SICTs (4, 8, 12)</td>
<td>8-10 alpha builds and 4 beta builds</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5/15/13</td>
<td>6/3/13</td>
<td>Survey</td>
<td>2014 TEL Probe (8)</td>
<td>46 items</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6/25/13* (via WebEx)</td>
<td>7/2/13</td>
<td>Cognitive</td>
<td>2015 Pilot Science ICTs (4, 8, 12)</td>
<td>8 alpha builds</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/18/13</td>
<td>8/8/13</td>
<td>Cognitive</td>
<td>2014 TEL Probe (8)</td>
<td>20 Tasks 100 Discretes</td>
<td>Review at August Board Meeting</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/18/13</td>
<td>8/8/13</td>
<td>Cognitive</td>
<td>2015 Science HOTs (4, 8, 12)</td>
<td>9 tasks</td>
<td>Review at August Board Meeting</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7/18/13</td>
<td>8/8/13</td>
<td>Cognitive</td>
<td>2015 Science ICTs (4, 8, 12)</td>
<td>9 beta builds</td>
<td>Review at August Board Meeting</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/13/13</td>
<td>9/3/13</td>
<td>Survey</td>
<td>2017 Pool Review Math (4, 8, 12)</td>
<td>131 items</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8/30/13* (via WebEx)</td>
<td>9/23/13</td>
<td>Cognitive</td>
<td>2015 Science ICTs (4, 8, 12)</td>
<td>7 beta builds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/27/13* (via WebEx)</td>
<td>10/18/13</td>
<td>Cognitive</td>
<td>2015 Science ICTs (4, 8, 12)</td>
<td>7 beta builds</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/21/13</td>
<td>12/13/13</td>
<td>Survey</td>
<td>2015 Operational Reading (4, 8)</td>
<td>78 items</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/21/13</td>
<td>12/13/13</td>
<td>Survey</td>
<td>2015 Operational Math (4, 8)</td>
<td>90 items</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Proposed date; not yet scheduled

**NOTE:** Alpha builds will be presented to the ADC during their in-person and virtual meetings. These will not be submitted before the review. The ADC will receive outlines and beta builds prior to the ICT review meetings. (Alpha and beta builds are the first- and second-draft versions of the rendered task, respectively.)
# AGENDA

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Chair/Presenter</th>
<th>Attachment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00 – 2:05 pm</td>
<td>Welcome and Introductions</td>
<td>Terry Holliday, Committee Chair</td>
<td>Attachment A</td>
</tr>
<tr>
<td>2:05 – 2:25 pm</td>
<td>Update on Implementation of Board Policy on Background Questions and Use of Contextual Data in Reporting</td>
<td>James Deaton, NCES Staff</td>
<td>Attachment B</td>
</tr>
<tr>
<td>2:25 – 2:55 pm</td>
<td><strong>ACTION:</strong> Revisions to NAEP Background Information Framework</td>
<td>Committee Discussion</td>
<td>Attachment C</td>
</tr>
<tr>
<td>2:55 – 3:05 pm</td>
<td><strong>ACTION:</strong> Terminology for Background Questions</td>
<td>Committee Discussion</td>
<td>Attachment D</td>
</tr>
<tr>
<td>3:05 – 3:20 pm</td>
<td><strong>ACTION:</strong> Organizing Board Consideration of Background Questions and Reporting on Contextual Data</td>
<td>Larry Feinberg, NAGB Staff</td>
<td>Attachment E</td>
</tr>
<tr>
<td>3:20 – 3:50 pm</td>
<td>Exploratory Analysis on Using NAEP Data for Key Education Indicators</td>
<td>Alan Ginsburg, Consultant</td>
<td>Attachment F</td>
</tr>
<tr>
<td>3:50 – 4:00 pm</td>
<td>Presentation to Governing Board</td>
<td>Committee Members</td>
<td></td>
</tr>
</tbody>
</table>
National Assessment Governing Board

Ad Hoc Committee on NAEP Background Information

BACKGROUND AND PURPOSE

As part of the resolution on NAEP background questions, adopted by the Governing Board in August 2012, an Ad Hoc Committee on NAEP Background Information is to be established for one year. Its purposes are as follows:


2. Review the NAEP Background Information Framework, adopted August 1, 2003. Recommend revisions, additions, or replacement, as deemed necessary or desirable.

3. Recommend a permanent arrangement for Board consideration of background questions and the reporting of contextual data in NAEP. This work is now divided between the Reporting and Dissemination and Assessment Development committees of the Board.

SPECIFIC ISSUES

As listed in the resolution, these include:

- Making greater use of contextual data in NAEP Report Cards and focused reports.
- Using background data to describe patterns and trends, including the educational experiences of different student groups.
- Detailed frameworks to support the selection of non-cognitive topics and questions, including their connection to student achievement.
- Clusters of questions on topics of continuing interest, such as technology and out-of-school learning, to be used regularly or rotated across cycles.
- Modules on issues of current policy interest.
- Elimination of duplicative, low-priority, or unproductive topics and questions.
- Use of questions from international assessments, such as TIMSS and PISA.
- Improved measures of socio-economic status (SES), including exploration of SES index.
- Spiral sampling and rotation of background questions in different years.
- Increasing the maximum time for students to answer background questions.
- Additional focused reports with the appointment of advisory committees in some cases.
- Exploratory analyses of existing data that may form the basis for subsequent reports.
- Consistency of wording to preserve trends.
- Further improvements in the NAEP Data Explorer.
COMMITTEE COMPOSITION AND TIMELINE

The Ad Hoc Committee will include Board members with a variety of perspectives and membership in different standing committees of the Board. The Committee will convene during each quarterly meeting of the Governing Board, and is expected to make its final report in August 2013.

COMMITTEE MEMBERSHIP

Terry Holliday, Chair
Doris Hicks
Andrew Ho
Brent Houston
Dale Nowlin
Joseph O’Keefe, S.J.
Susan Pimentel
Leticia Van de Putte
Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting

INTRODUCTION

By statute, the purpose of the National Assessment of Educational Progress is to provide a “fair and accurate” measure of student achievement and achievement trends. Academic or cognitive questions are its primary focus; the American public is its primary audience. However, in addition to reporting on what American students know and can do, NAEP has collected data for more than 40 years that provide a context for reporting and interpreting achievement results. According to the statute, such factors, both in and out of school, must be “directly related to the appraisal of academic achievement.”

In each assessment NAEP administers background questionnaires for students, their teachers, and schools. The questionnaires deal with educational experiences and other factors, such as teacher training or out-of-school learning activities, that are related to academic achievement. Data on several hundred background or noncognitive variables are available on the Internet through the NAEP Data Explorer. However, for more than a decade, little use has been made of this information in NAEP reports. The data have received minimal attention and had little impact despite the considerable efforts expended in developing and approving questionnaires and collecting and tabulating responses.

In October 2011 the National Assessment Governing Board convened an expert panel to recommend how to make better use of existing NAEP background questions and to propose an analytic agenda for additional topics and questions that would be useful in developing education policy and of value to the public. The panel report, entitled, *NAEP Background Questions: An Underused National Resource*, was presented to the Board in March 2012 by Marshall Smith, former U.S. Under Secretary of Education, who chaired the six-member panel.

Many of the panel recommendations build on the *Background Information Framework for the National Assessment of Educational Progress*, adopted by the Governing Board after it received final authority from Congress over non-cognitive items on the assessment. The framework was adopted in 2003, but has not been fully implemented.
The following policies are based on recommendations by the expert panel. The Board has also taken into consideration a wide range of public comment and the analysis provided by the National Center for Education Statistics.

It is important to understand that the National Assessment is not designed to show cause-and-effect relationships. Its data should not be used to “prove” what schools should do. But, as the Background Information Framework declares, NAEP’s “descriptions of the educational circumstances of students…, considered in light of research from other sources, may provide important information for public discussion and policy action.” The Board believes the National Assessment should improve upon its efforts to collect contextual information and present it clearly to the public, which will add to NAEP’s value to the nation.

POLICY PRINCIPLES

1. NAEP reporting should be enriched by greater use of contextual data derived from background or non-cognitive questions asked of students, teachers, and schools. Such data will be used both in regular Report Cards and in special focused reports.

2. Reporting of background data will describe patterns and trends, including the educational experiences of different groups of students. Care should be taken not to suggest causation.

3. Detailed frameworks will be published with the theoretical rationale and research evidence that support the selection of topics and questions in background questionnaires and their connection to student achievement. Such frameworks should be updated for each assessment cycle and provide the basis for new topics and questions.

4. An ad hoc committee of the Board will be established for one year to monitor implementation of this resolution, review the NAEP Background Information Framework, and recommend a permanent arrangement for Board consideration of background questions and the reporting of contextual data in NAEP.

IMPLEMENTATION GUIDELINES

For Questions and Questionnaires

1. Clusters of questions will be developed on important topics of continuing interest, such as student motivation and control over the environment, use of technology, and out-of-school learning, which could be used regularly or rotated across assessment cycles.
2. Modules will be prepared for special one-time studies to provide descriptive information on issues of current policy interest.

3. A thorough review will be conducted to eliminate duplicative or low-priority questions. Unproductive topics and questions will be dropped.

4. NAEP will include background questions from international assessments, such as PISA and TIMSS, to obtain direct comparisons of states and TUDA districts to educational practices in other countries.

5. Because of the value of preserving trends, consistent wording of questions should be maintained on topics of continuing interest. Changes in wording must be justified. However, as practices and circumstances change, new questions will be introduced in a timely manner to gather data on topics of current interest.

6. The development and use of improved measures of socio-economic status (SES) will be accelerated, including further exploration of an SES index for NAEP reporting.

For Data Collection

7. The maximum time for students to answer the background questionnaire will be increased from 10 to 15 minutes on new computer-based assessments. Consideration should be given to a similar increase in paper-and-pencil assessments.

8. Whenever feasible, assessment samples should be divided (spiral sampling) and background questions rotated in different years in order to cover more topics without increasing respondent burden. These practices will be initiated in the assessments of reading and mathematics, which are conducted frequently, and considered for other subject areas if the frequency of testing permits.

For Reporting

9. Special focused reports with data through the 2013 assessment will be issued on the following topics: private schools, charter schools, gender gaps, and black male students. Reports shall include significant contextual information as well as cognitive results. Advisory committees, composed of a range of knowledgeable persons, may be appointed to provide input on reporting issues.

10. Exploratory analyses will be carried out to determine if existing background questions may form the basis for additional focused reports. Such reports may be issued by the Governing Board as well as by the National Center for Education Statistics.
11. The NAEP Data Explorer should be further improved to make data more accessible to general, non-specialist users. Tables and very simple-to-construct charts will be prepared to present data on important topics of wide public interest. Additional means of disseminating information, using new technology such as simple apps that would allow parents, teachers, and others to access background and achievement data, will be explored.
NOTE TO Ad Hoc Committee on Background Information on Revisions to Background Information Framework

The revisions proposed to the Background Information Framework for NAEP are intended to make it conform to the Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting, which the Board adopted in August 2012. The changes would also update the text of the framework, which was adopted ten years ago.

Also attached is an annotated version of the August 2012 resolution, showing the pages in the framework where changes are proposed.

A new foreword is attached that explains the key changes.
Background Information

Framework

for the

National Assessment of

Educational Progress

National Assessment Governing Board
Adopted August 1, 2003
PROPOSED CHANGES July 2013
National Assessment Governing Board

Darvin M. Winick
Chair

John H. Stevens
Chair, Ad Hoc Committee on Background Questions

Charles E. Smith
Executive Director

Lawrence Feinberg
Project Officer

Background Framework Project

Mary Lyn Bourque, Writer/Editor
Mid-Atlantic Psychometric Services

David Grissmer
RAND

Consultants

Paul E. Barton
Ina V.S. Mullis
Michael Podgursky
Mark Reckase
Herbert J. Walberg
## Table of Contents

**ADD: Foreword - 2013**

Preface .................................................................................................................. 5

Executive Summary ............................................................................................... 7

**Chapter One: Introduction** .................................................................................. 12

Nature and Purpose of NAEP .............................................................................. 12
Nature and Purpose of Background Data ............................................................ 12
Development of NAEP Background Information Framework ........................... 14
Requirements of NAEP Statute ............................................................................ 15
Purpose and Rationale of Background Information Framework ..................... 16

**Chapter Two: Priorities and Criteria for Collecting and Reporting Non-Cognitive Data on NAEP** .......................................................... 17

Priorities for Non-Cognitive Information ............................................................ 17
Criteria for Selecting Topics and Questions ....................................................... 18
Criteria for Selecting Data Sources .................................................................... 20

**Chapter Three: Topics and Types of Background Data** .................................... 22

Demographic Reporting Categories ................................................................. 22
Socio-economic Status ....................................................................................... 23
Public Policy Contextual factors ....................................................................... 25
Subject-Specific Background Data .................................................................... 26

**Chapter Four: Non-cognitive Data Sources and Collection** ............................. 28

NAEP Student, Teacher, and School Samples ..................................................... 28
Other Data Sources ............................................................................................ 29
NAEP Background Information Matrix ............................................................. 31

**Chapter Five: Using Background Data to Report NAEP Results** .................. 33

Use of Descriptive Information in NAEP ............................................................ 33
Implementing No Child Left Behind ................................................................. 33
Disaggregation of NAEP Data ............................................................................ 34
Minimizing Causal Interpretations………………………………………. 34

Chapter Six: Using NAEP in Educational Research………………………………………………………….36
Based in part on papers by David Grissmer, RAND
NAEP’s Limitations and Strengths for Research……………………………………36
Previous Use of NAEP in Research………………………………………………38
Secondary Analysis Grants and District Samples…………………………39

Chapter Seven: Review and Improvement of Non-cognitive Questions……40

Independent Validity Studies…………………………………………………40
Review of Background Information Framework…………………………41

References…………………………………………………………………………42

Appendix A: Governing Board Policy on Collection and Reporting of Background Data by the National Assessment of Educational Progress (2002)……44

ADD: Appendix B: Governing Board Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting (2012)

Acknowledgments…………………………………………………………………….51
Preface

by the National Assessment Governing Board

The National Assessment of Educational Progress (NAEP) has been established by law to monitor the academic achievement of American students. In addition to its academic assessments, NAEP has collected information from hundreds of non-cognitive or background questions about students, their educational experiences in class and at home, their teachers, and their schools. Some of these questions provide data for NAEP’s reporting categories, but far more have been used to give context to NAEP results or to track factors associated with academic achievement. Some have been used by scholars in social science research.

Concerns have been raised about the selection of background variables, the quality of the information obtained, and the validity of inferences drawn from it. There is also concern about the burden that collecting background information places on respondents and on the NAEP program. After the National Assessment Governing Board was granted final authority over the background questions in early 2002, it adopted a policy to focus NAEP background data on the primary purpose of the National Assessment—to provide sound, timely information on the academic achievement of American students. The Board also initiated a process to prepare a general framework to guide the collection and reporting of background data.

It is important to understand the National Assessment is not designed to prove cause-and-effect relationships; it cannot prescribe what should be done. But its descriptions of the educational circumstances of students at various achievement levels—considered in light of research from other sources—may provide important information for public discussion and policy action. Used with other research, the contextual data collected by NAEP may give insights into how achievement can be improved as well report to the public on how school personnel and resources related to achievement are distributed.

This framework will define the purpose and scope of NAEP’s system of collecting background information, including background questionnaires and other sources of non-cognitive data. It will establish criteria for reporting background information as part of the National Assessment. The approach it suggests provides for asking various groups of questions to various samples of students at various times.

The framework reflects the following key principles:

- The selection of background topics and questions shall be designed to fulfill all legal requirements for the National Assessment and to carry out decisions regarding what NAEP will report and how to report it.
• Background information shall provide a context for reporting and interpreting achievement results and, as the statute provides, must be “directly related to the appraisal of academic achievement and to the fair and accurate presentation of such information.”

• The collection of background data shall be designed to obtain information that is objective, valid, reliable, and of consistently high quality.

• The system of background data collection shall be efficient and designed to minimize the burden on respondents and on the NAEP program. As much data as possible should be obtained from school records and other reliable data sources.

• These principles shall apply both to the collection of general background information and to subject-specific background questions. The frameworks for the latter must be focused and prioritized, indicating a core set of variables for regular reporting and a more comprehensive set to be collected and reported less frequently.

• The priority order for background information is as follows: (1) reporting categories, as required by law; (2) contextual factors with a well-established relationship to achievement; and (3) subject-specific information.

There is one other consideration—the new role of the National Assessment in the No Child Left Behind Act of 2001. Under this law, all states receiving federal Title I aid are required to participate every two years in NAEP’s state-level samples of reading and mathematics in grades 4 and 8. The results will provide an independent yardstick to compare trends on NAEP with performance on each state’s own set of required exams.

Because No Child Left Behind places particular emphasis on closing the persistent performance gaps between various student groups, NAEP must be able to report on changes in achievement for all groups specified by law. Through its background questions, the National Assessment might can also provide useful information about the students left behind and those who are ahead of them, including the sorts of schools that high-achieving and low-achieving students attend, the courses they take, the patterns of how they are taught, and the qualifications of their teachers. Over time, such descriptive information will allow NAEP to track changes in contextual and instructional factors related to student achievement and in the distribution of important educational resources.

In sum, the purpose of this Background Information Framework is to focus the collection and reporting of background data by the National Assessment and to establish clear priorities and limits. We hope to make it possible that with far fewer non-cognitive questions than it has had in the recent past, NAEP will serve the purposes of law and
provide the American public and decision makers with useful information. We are committed to improving the quality of data collected and the reporting of results.
Executive Summary

The National Assessment of Educational Progress (NAEP) is a federally authorized survey of student achievement at grades 4, 8, and 12 in various subject areas, such as mathematics, reading, writing, science, U.S. history, the arts, and foreign languages. The No Child Left Behind Act of 2001 (P.L. 107-110) requires the assessment to collect data on specified student groups, including race/ethnicity, gender, socio-economic status, disability, and limited English proficiency. It requires fair and accurate presentation of achievement data and permits the collection of background or descriptive information that is related to academic achievement and aids in fair reporting of results. The intent of the law is to provide representative-sample data on student achievement for the nation, the states, and subpopulations of students and to monitor progress over time.

The National Assessment Governing Board (NAGB) sets policy for NAEP and determines the content framework for each assessment. As a result of the No Child Left Behind Act, the Board is responsible for selecting and approving all of NAEP’s non-cognitive or background questions, as well as the cognitive items over which it has had final authority since 1988. This Background Information Framework will guide the development and selection of non-cognitive topics and questions, starting with the NAEP 2006 assessment. It will fulfill the purposes of law and implement Board policy.

When NAEP began in 1969-70, its background information was limited to gender, race/ethnicity, and literacy materials at home. During the 1980s the array of non-cognitive questions expanded greatly, both to provide more contextual information and in an effort—never fully realized—to use the assessment for educational research.

This background data framework will refocus the collection of non-cognitive variables on NAEP’s primary mission: providing a fair and accurate measure of student achievement and on achievement trends over time. Thus, the framework is a guide for gathering important information that will assist in reporting and understanding NAEP results. NAEP may contribute to research into improving education policy and practice, but its role in this respect is limited. Used with other research, the contextual data collected by NAEP may give insights into how achievement can be improved as well report to the public on how school personnel and resources related to achievement are distributed.

And the framework is not a comprehensive list of possible factors to explore.

Since by law NAEP may only collect information that is “directly related to the appraisal of academic achievement,” it must concentrate on non-cognitive variables that are known from other research to have such a relationship. The law also specifically prohibits NAEP from asking about personal or family beliefs and attitudes. These points
are emphasized in the Governing Board Policy Statement on the Collection and Reporting of Background Data by the National Assessment (adopted on May 18, 2002). That policy is incorporated into this framework. It is attached in the appendix.

PRIORITIES

The following priorities for collecting and reporting non-cognitive information should be followed in planning background questionnaires, the frequency with which questions are asked, and the samples from which data are collected.

(1) **Student reporting categories** that are required by law must be collected as a regular component of all NAEP assessments. These include race, ethnicity, gender, socio-economic status, disability, and limited English proficiency. A core of SES information should be collected in every assessment, such as type of community and poverty status. An expanded set of SES variables may be included periodically or administered to limited samples. **Efforts should be made to develop a composite measure or index of SES.**

(2) **Other factors that provide a context for results** should be sampled periodically, or on a rotating basis, over several NAEP cycles, although a limited set may be asked in every assessment. Contextual factors may include courses taken, student mobility, school safety and discipline, teacher-related factors such as demographics and experience, other factors related to students and schools, and educationally-relevant variables outside school. **Modules should be prepared for special studies to provide descriptive information on issues of current policy interest.** Although many non-cognitive variables may be of interest, they must be limited to meet the needs of NAEP reporting. In all cases, these **non-cognitive variables** must be clearly related to academic achievement or to the fair presentation of achievement results.

(3) **Subject-specific background information** should be gathered at the same time that achievement in a subject is assessed. This may include relevant course content and requirements, teacher preparation, and other factors related to student achievement. Questions will not be designed to determine effective practices, but to show patterns and trends of factors of interest, based on previous research. Like the contextual information, most of these variables should be sampled periodically, or on a rotating basis, over several administrations of the subject exam, although a limited core set may be repeated every time the assessment is given.

SELECTION CRITERIA

Key criteria for selecting non-cognitive topics and questions are as follows:
• **Does the current or proposed non-cognitive variable relate to the primary purpose of NAEP and how?**  The primary purpose of NAEP is to report on the academic achievement of students to the American public. It is not to report on the causes of that achievement. Other surveys with longitudinal data are far better suited to examining causality. NAEP’s choice of which non-cognitive variables to measure should be guided by how and to what extent the variables selected will support NAEP’s primary mission.

• **Do the current or proposed non-cognitive variables meet professional standards for reliability and validity?**  The NAEP legislation requires that the assessment “use widely accepted professional testing standards (P.L. 107-110, Sec. 411 (b) (5).” This requirement applies equally to non-cognitive and academic variables.

• **How stable is the non-cognitive variable from period to period?**  If a variable shows little change from year to year, it should be reviewed to determine whether it should be deleted or used on a periodic basis rather than in every assessment.

• **If new questions are added, have others been deleted in order to limit the burden and expense of NAEP’s background questionnaires?**  There will always be pressure to collect more information. Mechanisms must be developed to make sure the burden of background questionnaires does not expand over time.

• **Does a question address specific behavior rather than conclusions?**  Even for such questions, however, caution is advisable because self-reports are often unreliable.

• **Will the topic or question meet the test of broad public acceptability and not be viewed as intrusive or prying?**  NAEP’s non-cognitive questions are not kept secure, and all of them are to be posted on the Internet. Possible objections should be considered in deciding whether or not a question will be asked.

• **Does the topic or question deal with a factor in which trends over time are important?**

• **Will the information obtained be of value in understanding academic performance and taking steps to improve it?**  This is a fundamental issue to be addressed in evaluating all background questions proposed for NAEP.

Because of the value of preserving trends, consistent wording of questions should be maintained on topics of continuing interest. Changes in wording must be justified. However, as practices and circumstances change, new questions will be introduced in a timely manner to gather data on topics of current interest. NAEP should include background questions from international assessments, such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study), to obtain direct comparisons of states and TUDA districts to educational practices in other countries.
DATA COLLECTION

Whenever possible, NAEP should use information from school records and other reliable data collections in order to improve the validity of the information collected and limit the background questionnaires in NAEP itself. In exploring the utility of different data sources, the following criteria should be considered: (1) reliability, (2) universality, (3) currency, (4) respondent burden, (5) logistics, (6) efficiency and cost-effectiveness, and (7) the impact on timeliness of NAEP reporting.

Of the student reporting categories in Priority 1, information on gender, race/ethnicity, disability status, and limited English proficiency shall be collected in a uniform manner in all NAEP samples. NAEP is also required to collect information on socio-economic status. This will continue to be done in all samples, although there may be some variation in the number of factors on which data are obtained with a uniform core and more extensive data gathering in some cases.

Because socio-economic status cannot be measured simply or directly, NAEP has used “proxy” variables, such as eligibility for free or reduced-price lunch (a measure of poverty), parent education, and the number of reading materials in the home. The framework provides that NAEP explore development of a composite index for SES derived from the proxy variables currently collected from students and schools. To the extent that the index can be sharpened by additional data from readily available sources, such as zip codes and the census, this option should also be considered. Occasionally and in limited samples, more extensive SES questions may be asked. Although NAEP may never be able to produce a full composite of SES, based on family income, education, and occupation, efforts should be accelerated to develop and use improved measures of socio-economic status, including an SES index, should be made to find an approximation that is more informative than the current set of proxy variables.

For the past two decades, NAEP has collected information on a lengthy list of student, teacher, school, and beyond-school factors that may provide a context for achievement results and are of interest to policymakers, researchers, and the public. Yet, NAEP’s design as a cross-sectional survey places serious limitations on the inferences that can properly be drawn from this information. We propose a careful review of the contextual factors in NAEP to focus on the most important variables related to public policy. All such information must be clearly related to student achievement, as shown by other research. Different questions should be cycled in and out of the assessment periodically, and the use of data from non-NAEP sources should increase. Information should be collected at meaningful intervals in ways that may show significant patterns and change over time.

The collection of subject-specific background information should be focused, limited, and prioritized as part of the subject-matter frameworks adopted by the Board. For subjects tested regularly at two-year or four-year intervals, each subject there should
be a small core set of background items administered to the full sample each time a subject is assessed. An additional, more comprehensive set of questions should be administered periodically or to smaller subsamples.

Whenever feasible, student assessment samples should be divided (spiral sampling) and background questions rotated in different years in order to cover more topics without increasing respondent burden. These practices should be initiated in the assessments of reading and mathematics, which are conducted every two years, and considered for other subject areas if the frequency of testing permits.

Clusters of questions should be developed so that indexes may be prepared on important topics rather than relying on individual items alone. These clusters could be used regularly or rotated across assessment cycles.

Thorough reviews should be regularly conducted to eliminate duplicative or low-priority questions. Unproductive topics and questions should be dropped.

Detailed frameworks will be published with the theoretical rationale and research evidence that support the selection of topics and questions in background questionnaires and their connection to student achievement. Such frameworks should be updated for each assessment cycle and provide the basis for new topics and questions.

In constructing questionnaires it is important to place strict limits on the respondent burden they impose. As much data as possible should be obtained from school records and other reliable data sources. The average individual response time to answer background questionnaires for each assessment, as calculated in accordance with Office of Management and Budget (OMB) procedures, shall be limited as follows: ten minutes for each student on paper-and-pencil tests, 15 minutes per student on computer-based assessments, 20 minutes for each teacher, and 30 minutes for each school. Consideration should be given to increasing student response time on paper-and-pencil questionnaire if deemed practical and productive.

REPORTING

NAEP reporting should include contextual variables and subject-specific background information to enrich and give perspective to results. Consistent with space and operational limitations, descriptive information should be part of NAEP Report Cards and summary and highlights reports. The reports should present information on patterns and trends of non-cognitive variables known to have a relationship to academic achievement and may contain disaggregated data on school conditions and practices for various groups of students. Data on courses taken before NAEP assessments (either from transcripts or questionnaires) is of great public interest and can be related to academic results.
In addition, supplemental-special reports may should be prepared that focus on particular aspects of the background data collected-topics of public interest and importance. These reports should feature significant contextual information as well as cognitive results. In all cases, NAEP reports published by the National Center for Education Statistics must not state conclusions as to cause-and-effect relationships and avoid simplistic presentations that imply best practice.

All background questions and data collected by NAEP should be posted on the Internet so the public may be able to consider them in discussing results. Complete data files should be made available to researchers for further analysis. In all cases, NAEP reports published by the National Center for Education Statistics must not state conclusions as to cause and effect relationships and avoid simplistic presentations that imply best practice.

**RESEARCH**

As a cross-sectional survey without longitudinal data, the National Assessment is able to document school conditions and practices. It can report on achievement results. But it cannot properly be used to establish direct cause-and-effect relationships. Still, over the past three decades, NAEP has been part of two important research endeavors—exploring changes in the black-white test score gap since 1970 and seeking to establish the impact of state-level reforms during the 1990s. By monitoring achievement well, NAEP has provided sound data for researchers to use. NAEP results have been critical in identifying research hypotheses. Its contextual variables have added valuable information. Its large data sets have been combined with other information to tease out meaning and policy implications, though NAEP’s own reports have properly steered clear of these activities.

The Governing Board believes that by doing its main task of monitoring educational achievement well NAEP can make a valuable contribution to education research. The NCES program of secondary analysis grants for researchers to analyze NAEP data should continue. Educational researchers should be involved, under the auspices of NCES, in developing NAEP background questionnaires, validity studies, and other data collection efforts to carry out the provisions of this framework.

The primary purpose of NAEP is to provide fair and accurate information on student achievement. Its primary audience is the American public. The Governing Board believes that in serving its purpose and audience well, NAEP can contribute to educational research. It welcomes the interest and efforts of researchers.
Chapter One: Introduction

The National Assessment of Educational Progress is the only continuous long-term measure of student achievement in the United States in elementary and secondary schools. Its primary purpose is to report to the American public on academic achievement and its change over time.

Nature and Purpose of NAEP

The NAEP survey consists of two major components: academic assessments that measure the achievement of students on a broad range of content, and non-cognitive survey questions that collect descriptive information from students, teachers, and school administrators about demographic characteristics and the educational process. Since 1969 NAEP has measured achievement in most areas of the school curriculum, including mathematics, reading, writing, science, U.S. history, world geography, civics, economics, foreign language, computer science, and the arts. The content of NAEP assessments is determined through a framework development process that articulates the content parameters for each area and recommends subject-specific non-cognitive areas for data collection and reporting.

NAEP’s purpose is to report to the public on the status of academic achievement in America. The assessment does not report results for individual students, but only for groups of test-takers having large, representative samples, e.g., students from rural schools, from various ethnic groups, or from participating states, and, on a trial basis, large urban school districts. It must be able to provide data for fair and accurate comparisons between the states and subgroups on which it reports. The background data play a crucial role in ensuring the fair comparisons—over time and between student groups—that are at the heart of NAEP’s mission and value.

Nature and Purpose of Background Data

The most recent NAEP reauthorization (P.L. 107-110) gives the National Assessment Governing Board “final authority” to approve “all cognitive and non-cognitive assessment items.” This framework deals with the non-cognitive side of the Board’s responsibility, including the items that identify students in NAEP’s required reporting categories and the other information that provides a context for results and tracks factors associated with academic achievement.
The term “non-cognitive,” as used in the law, seems more inclusive than the phrase “background questions” by which the collection of non-academic information has been termed by NAEP in the past. However, non-cognitive is also less readily understandable than background information, and so the two terms will be used interchangeably in this document. Both will refer to all of the information beyond the academic assessment that NAEP uses to make its academic results more meaningful to the public.

When NAEP began, the collection of non-cognitive data was limited to the demographic categories of gender and race/ethnicity, and to two measures of home environment or socio-economic status—level of parents’ education and literacy materials in the home. In addition, an index was constructed, based on data from the U.S. Census and a brief school questionnaire, to report achievement results for schools in three types of communities—disadvantaged urban, advantaged urban, and rural.

During the 1980s the use of non-cognitive questions was greatly expanded to accommodate several functions within NAEP (Reckase, 2002). First, they were used to define a more extensive array of subgroups of the student population for reporting purposes. For example, NAEP results are now reported by gender, race/ethnicity, parents’ highest level of education, type of school, participation in Title I, and eligibility for free/reduced-price lunch.

A second reason for collecting non-cognitive information is to inform educational policy by describing the contexts for learning, sometimes called opportunities to learn (Mullis, 2002). Broadly, this involves the content specified in the curriculum, whether and how that content actually is taught, students’ propensity to learn, as well as home and school factors that can enhance learning.

In conjunction with the descriptions of students, background information about educational settings and experiences can reveal striking differences in how important aspects of education and educational resources are distributed among different groups. For example, do disadvantaged minority students have less access to science laboratory equipment than more advantaged groups? Do girls take less rigorous mathematics courses than boys? The data on course taking has been used widely to discuss the patterns and trends in mathematics achievement. Having this information as part of NAEP has added to the public impact of assessment results.

A third function of the non-cognitive questions has been to support research into factors that may be related to student achievement. The background questions serving this function have sought information not only on curriculum, teaching methods, and discipline in the school, but also on educational activities at home. For example, The 1998 NAEP Reading Report Card (Donahue, Voelkl, Campbell, & Mazzeo, 1999) reports on television viewing, daily reading habits, classroom reading and writing assignments, and discussion of schoolwork at home. While secondary researchers have used NAEP to investigate relationships to student achievement, the basic design of the assessment as a cross-sectional survey without longitudinal data limits its usefulness. Research has been
most productive when NAEP is combined with other data sources and in descriptive studies that track changes over time.

Non-cognitive data are also necessary to support certain technical functions of NAEP. For example, some non-cognitive information is used to evaluate the potential for bias resulting from non-participation. That is, did the students absent or refusing to participate in the assessment differ in such significant ways from those who did take part that results were changed? Non-cognitive variables also play an important role in NAEP’s sampling and weighting procedures, and sometimes in checking the validity of results. Many of these variables are taken from other data sources, such as the Common Core of Data (CCD), but some come from the administration roster collected from schools prior to testing, the records kept by test administrators, and student questionnaires.

Finally, NAEP non-cognitive questions have been used in the technical process for preparing estimates of student proficiency distributions on the cognitive component of the assessment. But their role in this process is limited to facilitating data analysis. Only the student responses to cognitive questions are used to determine achievement results. Background variables are used to define the groups for which cognitive data are reported.

Once test results for a group are determined, the NAEP analytic process makes use of background data available to prepare a second data set—identical in its group scores to the first—that can be handled by much simpler computer programs to prepare other analyses and reports. However, only the background factors to be reported on are needed for this analytical work, called conditioning. The precision of NAEP results is not reduced if background items not used for reporting are eliminated.

This background information framework will focus the collection of non-cognitive information on NAEP’s primary mission: providing, as the law stipulates, “a fair and accurate measurement of student academic achievement and reporting trends in such achievement” over time. Thus, the framework is a guide for gathering important information that will assist in reporting and understanding NAEP results.

**Development of NAEP Background Information Framework**

In the Policy Statement on Redesigning the National Assessment of Educational Progress (adopted in August 1996), the Governing Board sought to improve the validity of background information on NAEP, increase the efficiency with which it is collected, and reduce the number of background questions in the assessment itself. The statement was based on the report of a Design/Feasibility Team (Forsyth et al, 1996), headed by Robert Forsyth, which recommended a design that would rotate the collection of non-cognitive data into distinct modules administered over several assessment cycles. NAGB endorsed implementing that recommendation through a system of comprehensive and standard NAEP assessments that would be administered on a cyclical basis (NAGB, 1996).
Standard assessments would ask a short, essential core of background questions associated with a content area. Periodically, a comprehensive assessment would employ a much fuller complement of such questions to probe that area more extensively. Although some efforts have been made to reduce the background questionnaires and streamline data collection, the full impact of the NAGB policy has not yet been realized.

In early 2002, the No Child Left Behind Act transferred final authority over the non-cognitive questions from the National Center for Education Statistics to the National Assessment Governing Board. The Board adopted a new policy governing the development and selection of non-cognitive questions in May 2002, and initiated a process to prepare a general framework for non-cognitive data (NAGB, 2002). This framework would define the scope of NAEP background questionnaires, the priorities for collecting non-cognitive information, and the criteria for reporting non-cognitive data in NAEP. (See Appendix for full text of the policy.)

The Board created an Ad Hoc Committee on Background Questions and conducted an all-day workshop on the NAEP non-cognitive questions on September 24, 2002. Six consultants prepared and presented papers at the meeting that was attended by Board members, academic researchers, representatives of the national teacher organizations and other education groups, and NAEP contractors and staff. The six consultants are identified on the title page as contributors to this document.

In the months after the workshop, a draft framework was prepared. It was refined at several meetings of the Ad Hoc Committee, posted for public comment on the Internet, and was the subject of a public forum in Washington, D.C., on May 1, 2003. Altogether, oral comment and written testimony were received from 22 persons and organizations, many with differing perspectives and views. The Ad Hoc Committee and the Board carefully considered these comments, and the draft framework was revised at a Committee meeting on June 25. The Committee heard additional comment and made final revisions on July 31. The background information framework was reviewed by the full Governing Board several times during the course of its development. The Board adopted it unanimously on August 1, 2003.

While this framework is not a consensus document, it does encompass the thinking of a wide range of researchers, policy analysts, and users of NAEP data. It is the product of discussion and deliberation by the Governing Board, and incorporates Board decisions on the nature and focus of the background information to be included in NAEP. The framework will become operative in the 2006 National Assessment.

Requirements of NAEP Statute

The No Child Left Behind Act of 2001 (P.L. 107-110) requires NAEP to collect information on gender, race/ethnicity, socio-economic status, disability, and limited English proficiency. It must report test data on these groups, whenever feasible, that is cross-tabulated, compared, and reported according to the categories required.
The law also requires NAEP to collect only information that is directly related to academic achievement and to the presentation of such information in a fair and accurate manner. This means that NAEP needs to concentrate on variables that are known to be related to achievement rather than on theoretical constructs. The statute requires the Governing Board to ensure that all NAEP questions are “free from racial, cultural, gender, or regional bias”—a provision from previous law. But it adds new language that questions must be “secular, neutral, and non-ideological” and must not “evaluate or assess personal or family beliefs and attitudes.”

In their report on the bill, the House-Senate conference committee that negotiated its final form says the law “does not preclude the use of non-intrusive, non-cognitive questions, approved by the National Assessment Governing Board, whose direct relationship to academic achievement has been demonstrated and is being studied as part of [NAEP] for the purposes of improving such achievement.” The report language is not binding, but is intended to guide implementation of the law. This framework emphasizes that the legal prohibitions must be followed in preparing background questions and collecting any other non-cognitive data for NAEP.

In addition, the law makes it clear that NAEP may not disclose any personally identifiable information or maintain any system of records that contains such data. These restrictions are not new. They have dictated careful procedures in the past, which must be continued.

Purpose and Rationale of Background Information Framework

The purpose of the framework on background information is similar to that of NAEP’s content area frameworks: to guide the development of the assessment. The content frameworks have described the topics to be tested by NAEP and provided an outline of the assessment for each subject area. Purposefully, the frameworks attempt to be independent of a particular pedagogy. They do not specify what educational resources or processes should be used, but rather describe important achievement results. They provide states, schools, policymakers, and the public with a logical outline of the approach used in constructing the assessment.

The framework for NAEP background data will specify the parameters of the assessment from a reporting perspective. The background information that NAEP uses in its reports helps to give context and meaning to the cognitive results. It must be collected in a systematic way from the NAEP testing samples either through questionnaires or from other reliable sources, such as school records and other federal surveys. Collecting descriptive information from a variety of sources can improve the quality of the data obtained and increase efficiency while reducing the burden on respondents.

The Governing Board adopted a Policy Statement on the Collection of Reporting of Background Data on May 18, 2002 (NAGB, 2002). The statement is incorporated into this framework and attached in the Appendix.
A further statement, entitled Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting, was adopted by the Board on August 4, 2012. It has been used in revising the framework text and has been included in the Appendix.
Chapter Two: Priorities and Criteria For Collecting and Reporting Non-cognitive Data on NAEP

This chapter presents priorities for collecting and reporting non-cognitive information on NAEP. It also includes the criteria for selecting particular topics and questions, and for determining the frequency with which various data elements are reported. A final section presents criteria for identifying and selecting background data sources.

Priorities for Non-Cognitive Information

The following priorities for collecting and reporting non-cognitive information are based on legal requirements, the purposes of NAEP, and the strengths and limitations of the assessment. They should be followed in planning background questionnaires, the frequency with which questions are asked, and the samples from which data are collected.

(1) **Student reporting categories that are required by law must be collected as a regular component of all NAEP assessments.** These include race, ethnicity, gender, socio-economic status, disability, and limited English proficiency. A core of SES information should be collected in every assessment, such as type of community and poverty status. An expanded set of SES variables may be included periodically or administered to limited samples. **Efforts should be made to develop a composite measure or index of SES.**

(2) **Other factors that provide a context for results should be sampled periodically, or on a rotating basis, over several NAEP cycles, although a limited set may be asked in every assessment.** Contextual factors may include courses taken and course requirements, student mobility, school safety and discipline, teacher-related factors such as teacher demographics, preparation, credentials, and experience, and other factors related to students, schools, and educationally-relevant variables beyond the school. **Modules should be prepared for special studies to provide descriptive information on issues of current policy interest.** Although these types of non-cognitive variables are of interest, they must be limited so that they meet the needs of NAEP reporting. In all cases, they—non-cognitive variables—must be clearly related to
academic achievement or to the fair presentation of achievement results.

Subject-specific information may be gathered at the same time that academic achievement in a particular area is assessed. This may include relevant course content and requirements, teacher preparation, and other factors related to achievement in the subject assessed. Questions will not be designed to determine effective practices, but to show the patterns and trends of factors of interest, based on previous research. Like other contextual information, most of these variables should be sampled periodically, or on a rotating basis, over several administrations of the subject exam, although a limited core set may be repeated every time the assessment is given.

With regard to the points above, Walberg (2002) makes a suggestion that might be a workable solution to consider. Just as students in the NAEP samples do not respond to all the questions, say, in reading, but only to a portion of those for any one grade-level, so too, the non-cognitive questions could be rotated through different (smaller) NAEP samples. These non-cognitive “testlets” could be rotated through the NAEP samples by class or school, with students receiving different, expanded “testlets” in addition to a core set of background questions.

Criteria for Selecting Non-cognitive Topics and Questions

The Advisory Council on Education Statistics (ACES), a technical panel that used to advise the National Center for Education Statistics, spent a considerable amount of effort on the issue of NAEP non-cognitive questions. Its guidelines, adopted in May 1997, include a set of key questions that should be utilized in selecting topics and questions for NAEP background data collection. The questions with commentary are summarized below:

- Does the current or proposed non-cognitive variable relate to the primary purpose of NAEP and how? The primary purpose of NAEP is to report on the academic achievement of students to the American public. It is not to report on the causes of that achievement. Other surveys with longitudinal data are far better suited to examining causality. NAEP’s choice of which non-cognitive variables to measure should be guided by how and to what extent the variables selected will support NAEP’s primary mission.

- Do the current or proposed non-cognitive variables meet professional standards for reliability and validity? The NAEP legislation requires that the assessment “use widely accepted professional testing standards (P.L.107-110, Sec. 411 (b) (5)).” This requirement applies equally to non-cognitive and academic variables. It is already known that some non-cognitive variables in NAEP have weak reliability (e.g., data from 4th graders on their parents’ highest level of education and the self-reports of teachers on classroom
practice. If more reliable sources of such data cannot be found, these variables should be deleted from the assessment.

- **How stable is the non-cognitive variable from period to period?** If a variable shows little change from year to year, it should be reviewed to determine whether it should be deleted or used on a periodic basis rather than in every assessment.

- **Is the proposed or current non-cognitive variable of timely interest?** The educational environment changes from time to time, and consequently public interest in particular variables will change as well. It would serve NAEP well to review the set of non-cognitive variables periodically with this criterion in mind, deleting those that do not meet the test of timeliness and substituting others of current interest.

- **If new questions are added, have others been deleted in order to limit the burden and expense of NAEP’s background questionnaires?** There will always be pressure to collect more information. Mechanisms must be developed to make sure the burden of background questionnaires does not expand over time.

- **Does a question address specific behavior rather than conclusions?** For example, a question that asks teachers whether they adhere to national standards in mathematics or another subject is conclusionary and hard to interpret, since many teachers are apt to say yes, regardless of what they do. It would be better to ask about specific behaviors, such as homework assignments or computer use. Caution is advisable in this area too because self-reports are often unreliable.

The Board believes three other important criteria must also be considered:

- **Will the topic or question meet the test of broad public acceptability and not be viewed as intrusive or prying?** NAEP’s non-cognitive questions are not kept secure and must readily be available to anyone requesting a copy. Under Board policy, all questions asked are to be posted on the Internet. Possible objections should be considered in deciding whether or not to ask them.

- **Does the topic or question deal with a factor in which trends over time are of importance?** If trends are deemed important and the factor is related to achievement, the topic or question should be included periodically on a four-year or eight-year cycle, rather than being part of the background questionnaire each year. For example, measuring television watching in every NAEP assessment is not necessary. But it can be valuable to measure TV-watching every four or eight years to find out whether or not it is increasing.
• **Will the information obtained be of value in understanding academic performance and taking steps to improve it?** This is a fundamental issue to be addressed in evaluating all background questions proposed for NAEP.

Because of the value of preserving trends, consistent wording of questions should be maintained on topics of continuing interest. Changes in wording must be justified. However, as practices and circumstances change, new questions will be introduced in a timely manner to gather data on topics of current interest. NAEP should include background questions from international assessments, such as PISA (Program for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study), to obtain direct comparisons of states and TUDA districts to educational practices in other countries.

**Criteria for Selecting Data Sources**

NAEP has collected non-cognitive information from students, teachers, and schools, using NAEP background questionnaires. There are also administration rosters, completed by test administrators at the school level in advance of testing to determine characteristics of the testing samples. The Common Core of Data (CCD) is used to identify characteristics of schools (e.g., Title I funding), and schools also complete a questionnaire on special needs students (e.g., students with disabilities and limited English proficiency).

However, the collection of non-cognitive data may be shifted among these sources or to new sources in order to improve reliability, increase efficiency, or reduce burden. State management information systems and data collected for school report cards, as required by the No Child Left Behind Act, may have become very increasingly useful for NAEP. *Whenever possible, NAEP should use information from school records and other reliable data collections about students and schools in order to improve the validity of the information collected and limit the background questionnaires in NAEP itself.*

In exploring the utility of different data sources, the following criteria should be considered:

- **Validity** – Is the data obtained from the new source a valid indicator of what it purports to measure?

- **Reliability** – Is the data from the new source at least as reliable and consistent as that from the source previously used?

- **Universality** – Can the required data be collected by this method for all (or almost all) of the students and schools participating in NAEP and will it support valid comparisons over time?
• **Currency** – Will data obtained from a new data source be current enough to relate clearly to the assessment being conducted? If data from the census or some other source is several years old it may not accurately describe school or neighborhood conditions at the time of testing.

• **Respondent Burden** – Will the new source(s) reduce the burden on students, teachers, and schools in filling out NAEP questionnaires? Will the total amount of respondent burden be decreased?

• **Logistics** – Will the alternative source(s) be logistically possible, or will there be more logistical problems than with the previous data source? Logistics includes such considerations as cost, time, administrative personnel resources, and steps needed to ensure accurate coding and data analysis.

• **Efficiency and cost-effectiveness** – How efficient will the new data source be in comparison to the previous one? For example, it may be more efficient to collect data from a state management information system about the state’s schools, teachers, or students, rather than obtaining it from the test samples directly, but efficiency and cost-effectiveness should be determined before a change is made.

• **Timeliness of NAEP reporting** – How will a change in data sources affect the speed with which NAEP can be reported? Some changes will speed operations, but those that slow down NAEP reporting are not desirable.
Chapter Three: Topics and Types of Background Data

This chapter will cover the non-cognitive topics that are required for reporting under the No Child Left Behind Act of 2001 (P.L. 107-110), as well as those that should be considered for inclusion in NAEP on a cyclical basis. It discusses socioeconomic status (SES), contextual factors of interest to public policy, and subject-specific variables.

Demographic Reporting Categories

The demographic variables currently collected by NAEP come from two sources. Information is obtained from school records on gender, age, race/ethnicity, and two elements of socio-economic status (SES)—participation in Title I and eligibility for free or reduced-price lunch, which is based on family income. The school records are also used to indicate whether a student is classified as disabled. In addition, information is obtained on disability status and on students who are classified as limited English proficient. All of this information is collected on an administration roster, completed from school records in advance of testing. In addition, data on race/ethnicity is also collected on the NAEP student questionnaire, and students are asked to report on the highest level of each parent’s education and on several aspects of home environment, including number of books, internet access, and whether they have their own bedroom.

A more extensive questionnaire is completed by school staff on each student selected for NAEP who is classified as either disabled or limited English proficient (LEP). For students with disabilities (SD), the questionnaire collects data on the specific disability and its severity, the student’s Individualized Education Plan (IEP), type of curriculum, whether the student participates in standardized testing (with or without accommodations), and the accommodations allowed on state and district standardized tests in presentation, response, setting, and timing. For LEP students, the questionnaire covers native language, number of years of academic instruction in English, percent of instruction in English and/or native language, and the testing accommodations provided under district or state policy. In the future, NAEP might also identify students who recently exited from LEP programs and track their achievement.

NAEP is required to collect information on all of these categories (except age), but has some discretion in determining definitions and aggregating responses. These data will continue to be collected in a uniform manner in every NAEP assessment, although, for socio-economic status, as explained in the section below, there may be some variation, with a uniform core and more extensive data-gathering in some cases.
Under current law, NAEP is required to collect information on socio-economic status. SES also is clearly a factor that has been shown to be related to academic achievement in many research studies, beginning with the Equality of Educational Opportunity Commission Report (Coleman et al., 1966). The research community’s consensus over the past four decades has been to deal with the influence of SES on other achievement-related variables by holding SES constant while examining the other effects, for example, adjusting for SES while looking at effects of class size or teacher training. NAEP does not adjust for SES, but it does report on the relationship between student achievement and SES proxy variables like parents’ education or Title I participation.

NAEP has not been able to measure SES directly, using its present set of questions and data sources, i.e., the student, teacher, and school questionnaires. The assessment has used “proxy variables” for SES, including students’ eligibility for the National School Lunch program, participation in Title I, parents’ education, and the number of reading materials in the home (newspapers, magazines, books, etc.)—information on the latter two factors being reported by students in the assessment samples. In addition, NAEP uses census data to classify schools into different types of location, based on Census Bureau definitions, such as central city, suburban/large town, and rural/small town. The questions on newspapers and magazines were dropped in the mid-2000s as circulation dwindled, and were replaced by an item on internet access.

Strictly speaking, these are individual proxy variables and are not combined into a composite variable. However, both the questions on parent education and home environment have been coded in a pseudo-composite manner. For example, the parent education related to the student is the higher of either the mother’s or father’s education level. On the four home environment questions used—until the mid-2000s student responses were coded differently for a “yes” answer to two questions or fewer, “yes” to three questions, and “yes” to four questions, as well as omitted responses (Allen, Carlson, & Zelenak, 1999).

At the lower grade levels, students’ reports of their parents’ education are questionable at best, while the National School Lunch program sorts students only into three categories (Yes, No, and Unknown) and Title I into two categories (Yes or No). For many years, NAEP used a reporting category of disadvantaged urban schools, constructed from information provided by school principals. This was discontinued in the mid-1990s because the category lacked a consistent definition from year to year and between different state samples. There also were serious doubts about the reliability of the information on which it was based. The data on eligibility for the National School Lunch Program have also become increasingly problematic because of expansion of the program and administrative changes allowing whole-school or whole-district eligibility in high-poverty areas. In short, there has been considerable concern over many years about the quality of the SES measures in NAEP, both for reporting to the public and for analysis by researchers.
Barton (2002) suggests two alternative approaches for improvement: (1) a composite index for SES, or (2) a parent questionnaire. A composite index is viable using the same information that is currently collected in NAEP, or perhaps augmented with a few targeted questions or census data, possibly the zip code of student home addresses. The necessary analytical work should be initiated through small research studies using extant NAEP data sets in order to check systematically the validity of a composite index as a better measure of SES in NAEP samples. The results could vary by grade level, in which case, adjustments might be needed in the way the data are collected, augmented, and/or confirmed. NAEP may never be able to produce a full composite of income, education, and occupation, but efforts should be accelerated to develop and use improved measures of socio-economic status, including an SES index. 

In November 2012, an expert panel convened by the National Center for Education Statistics recommended prompt development of an SES composite measure. The argument in favor of this approach is that it advances the goals of the current law without impacting data collection in unforeseen ways. Barton suggests that such an index would enable NAEP to report results in terms of SES quartiles (much the same way that the National Educational Longitudinal Survey, NELS, does). Further, it would allow the assessment to report cross-tabulations on distributions of students in the NAEP achievement level categories by SES. A good measure of SES would improve the monitoring of achievement gaps among various racial/ethnic groups, although sample sizes may not be large enough within all ethnic groups or types of schools. Finally, a composite SES index may be beneficial to states and districts in the Trial District Assessment (TUDA), enabling NAEP to compare the performance of groups of students with the same socio-economic status, which is a factor of high public and policy interest.

The argument against such an approach is that SES would continue to be measured indirectly, i.e., by using proxy variables, albeit through a composite index. There would also be disagreements about precisely which variables to include in the index and how to weight different factors. For example, Armor (D. J. Armor, personal communication, December 18, 2002) has suggested that two variables recently deleted from the NAEP student questionnaire in 2000 be reinstated, namely, the number of siblings in the home and family status (student lives with both parents, mother or father, neither). These variables were dropped because of concerns about intrusiveness, but they may be of considerable importance in constructing an SES index. The item on number of parents in the home was restored in 2013. The Board will have to weigh the considerations involved, and may decide there is value in using them periodically or in limited samples.

A parent questionnaire has been proposed as a more reliable means of collecting SES data than relying on student reports, school records, or census data. Other National Center for Education Statistics surveys, for example, NELS and the Early Childhood
Longitudinal Study, have employed parent questionnaires that ask direct questions regarding occupation and income.

However, the National Assessment of Educational Progress involves far more students than any of these research surveys. Accordingly, a parent questionnaire on NAEP would entail far more respondent burden and might arouse more controversy, making it more difficult to accomplish the primary mission of the assessment to measure student achievement. A parent questionnaire has been considered by NAGB in the past, but rejected as too burdensome and intrusive. Because these considerations are still persuasive, particularly as the scope of NAEP has expanded, no work should be undertaken on developing a parent questionnaire.

In sum, because of its importance and the requirements of law, information on socio-economic status must be collected in all NAEP samples, although there may be some variation in the number of factors on which data are obtained. Research efforts should be conducted—made to develop into creating—a composite measure or index of SES based on school records and the student questionnaire. To the extent that an index can be sharpened by additional information from readily available sources, such as zip codes and/or census data, this option should be considered as well.

A core of SES information should be collected in every assessment, such as type of community (e.g., central city, rural, etc.), poverty status (e.g., eligibility for free or reduced-price lunch and Title I participation), reading materials in the home, and level of parent education—though steps must be taken to ensure that such data are reliable. An expanded set of additional SES variables may be included also be periodically administered to limited samples, including such as number of siblings and parents at home, possession of computers, and parent occupation. Periodically, an expanded set may be administered.

NAEP should explore the use of an SES index derived from proxy variables currently on the administration roster or student questionnaire. To the extent that an index can be sharpened by additional information from readily available sources, such as zip codes and/or census data, this option should be considered as well.

Public Policy Contextual Factors

For the past two decades NAEP has collected information on student, teacher, school, and beyond-school factors that are of interest to policymakers and the public. For students, some of these factors have included course-taking patterns, TV-watching, homework, and use of computers. For teachers, the contextual factors have included educational background, credentials, years of experience, and participation in professional organizations, to name a few.
The lists of factors have been long. They have become burdensome both to respondents and to the efficient scoring, analysis, and reporting of the NAEP survey. The way they have been reported—through simple one-way tabulations—has encouraged unwarranted conclusions about cause-and-effect relationships.

We propose a careful review of the contextual factors on which information is collected by NAEP to focus on the most important variables related to public policy. All such information must be clearly related to student achievement, as shown by other research. Data should be obtained periodically, on a rotating basis, over several NAEP cycles, although a limited set of factors may be included in every assessment. Modules should be prepared for special studies to provide descriptive information on issues of current policy interest. Information on data should be collected at meaningful intervals in ways that may show significant patterns and change over time.

Two documents are helpful in surveying the research base and presenting alternatives for NAGB to consider. The first is Monitoring School Quality: An Indicators Report (Mayer, Mullens, & Moore, 2001), prepared by Mathematica Policy Research, Inc. for NCES. This report presents a research synthesis, indicating factors for which there is a research base showing a strong relationship to academic achievement. The synthesis, involving a review panel as well as statistical analyses, identifies the following as factors related to student results: the academic skills of teachers, teacher assignments (such as out-of-field teaching), course content, student discipline and school safety, class size, and focus on academic achievement. Other sources of information are available on all of these factors, but only through NAEP can they be related to the achievement of broad groups of students over time.

The second document, Making Connections (Greenberg, Stancavage, Farr, & Bohrnstedt, 2001), was prepared for NCES by the American Institutes for Research and presents an elaborate typology of non-cognitive variables that could be measured by NAEP. It is organized into seven broad categories of non-cognitive information related to students, instructional content and practice, teachers, schools, school community factors, beyond school factors, and federal, state, and district policy. The listing goes beyond what NAEP can and should handle, but its discussion is thoughtful and the document is useful for planning.

Subject-Specific Background Data

For each subject assessed by NAEP, additional subject-specific background information has been collected from students, teachers, and schools. These data fall into the broad category of instructional content and practice. Under that umbrella come such topics as the curriculum taught, course offerings, class management and style, ability grouping, and modes of instruction. Subject-specific data collection has expanded enormously over the past two decades, and in recent years has included five to ten minutes of questions for students, about 30 minutes of questions for teachers, and 30 to 45 minutes for school administrators.
Now is the time for these questions to be focused, limited, and prioritized. Future subject-matter frameworks adopted by the Governing Board should spell out clearly what these priorities will be.

Whenever feasible, student assessment samples should be divided (spiral sampling) and background questions rotated in different years in order to cover more topics without increasing respondent burden. These practices should be initiated in the assessments of reading and mathematics, which are conducted every two years, and considered for other subject areas if the frequency of testing permits.

The design for doing this was presented to the Board in the 1996 report of a Design/Feasibility Team of prominent researchers (Forsyth, R., et al, 1996). The group recommended that a core set of non-cognitive questions should be administered to students each time a subject is assessed by NAEP. In addition, a more comprehensive questionnaire would be given whenever a new framework is introduced and repeated every eight to ten years. For example, an extensive set of background questions in reading and mathematics (grades 4 and 8) was administered in 2003, the baseline year for the No Child Left Behind legislation. Another complete set should be administered in mathematics in 2005 and in reading in 2007, the years in which revised frameworks are first used, and then should be repeated at an interval of eight years. In the intervening years, only the more limited core modules will be administered. Similar patterns should be established for the school and teacher questionnaires.

In The NAEP assessments in other subjects given at intervals of four years or more, such as writing, science, history, geography, and civics, should have a core set of non-cognitive questions administered to the full sample, with different sets of longer, more extensive questionnaires being administered to smaller sub samples.

With states now required to participate in NAEP every two years, the total number of students tested has expanded substantially from what it was in the program’s first decades. This makes even more compelling the case for limiting the NAEP background questionnaires and rotating the background questions.

Clusters of questions should be developed so that indexes may be prepared on important topics rather than relying on stand-alone items only.

NCES should prepare for Board review and approval a plan indicating the frequency, sample size, and schedule of rotation for all background variables and questions on which information is to be collected by NAEP. This should include both questionnaires and alternate data sources to obtain core reporting data, subject-specific information, and data on achievement-related contextual variables from a variety of NAEP samples—national only, national and state, and a subset of the national sample. The plan should indicate the frequency and schedule of rotation for each of the questions proposed. It should also indicate any questions needed for
quality control purposes. The recommendations should be prepared with input from researchers and state policy analysts, as appropriate, and updated on a regular basis.

Table 1 presents a model schedule for comprehensive and core sets of subject-related variables through 2013. It is based on the schedule of assessments approved by the Board in May 2003.

Table 1

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Data Collection Year for Comprehensive Set of Variables</th>
<th>Data Collection Year for Core Variables Only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign Language (12)</td>
<td>2004, 2012</td>
<td>TBD</td>
</tr>
<tr>
<td>World History (12)</td>
<td>2010</td>
<td>2006</td>
</tr>
<tr>
<td>Economics (12)</td>
<td>2006</td>
<td>TBD</td>
</tr>
<tr>
<td>Arts (8)</td>
<td>1997, 2008</td>
<td>2005</td>
</tr>
<tr>
<td>Science</td>
<td>2000, 2009</td>
<td></td>
</tr>
<tr>
<td>US History</td>
<td>2001, 2006</td>
<td></td>
</tr>
<tr>
<td>Geography</td>
<td>2001, 2010</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Based on schedule approved by NAGB on May 17, 2003.
Chapter Four: Non-cognitive Data Sources and Collection

This chapter discusses the sources of non-cognitive information for NAEP and the reporting categories that the information describes. It includes a NAEP Background Information Matrix, organized by priorities, which summarizes the types of descriptive information NAEP collects, reporting units, and data sources.

NAEP Student, Teacher, and School Samples

The NAEP student samples vary in size and purpose. Their overall total has become very large. Starting in 2003, national NAEP samples are specified at the state and jurisdictional levels, with approximately 3,000 students per subject and grade (4 and 8 only) for each of the 50 states, plus the District of Columbia, and Department of Defense domestic and overseas schools. Puerto Rico (in mathematics only) has a sample of about 3,000. In addition, the ten Trial Urban District Assessment (TUDA) districts have sample sizes of the order of 3,000 to 5,000 each. There also are a nationally-representative sample of charter schools, totaling about 3,000 students, and national private school samples totaling about 12,000 in each grade.

At grade four, therefore, the total NAEP sample approximates 436,000 students. The grade eight sample is about the same at 432,000 (excepting charter schools). The grade 12 sample is for a pilot test and includes only about 6,000 students (Rust, 2002). In most future years the twelfth grade samples are expected to have about 30,000-40,000 students assessed in national samples only for three subjects.

In addition to the nearly one million students tested, about 80,000 teachers of those students complete teacher questionnaires and some 13,000 schools complete school questionnaires. Several thousand school districts also supply data for the assessment. The sampling and weighting procedures in NAEP use data from the CCD files as well as census data and school-level achievement data from the states for improving NAEP stratification procedures. The NAEP non-cognitive data collection effort is enormous and challenging.
Other Data Sources

The Governing Board is strongly committed to improving the quality of background information while reducing respondent burden and the complexity of data collection and analysis. The self-report questionnaires given to students, teachers, and schools are sometimes burdensome to fill out, labor-intensive to collate and analyze, and subject to concerns about reliability. All questionnaires should be scrutinized to replace as many items as possible with data from centralized records, gathered by test administrators, or, ideally, from computerized data files.

The data available from federal, state, district, and school records should be carefully explored. With implementation of the school report card requirements of the No Child Left Behind law, In recent years much more information should become available in standardized computer formats. Barton (2002) has suggested some specific sources of data collected outside of NAEP that should be considered to improve NAEP reporting. These include the U.S. Census, Quality Education Data, Inc. (QED), and the Common Core of Data (CCD) and School and Staffing Survey (SASS), both compiled by the National Center for Education Statistics.

This approach of utilizing more data from outside specific NAEP data collections has been elaborated on extensively in the most recent evaluation of NAEP by the National Academy of Sciences (Pellegrino, J.W., Jones, L.R., & Mitchell, K.J., 1999). The panel proposed “a coordinated system of indicators for assessing educational progress, housed within NCES and including NAEP and other currently discrete, large-scale data collections (p. 34).” Figure 1 is reprinted from the NAS report to show the extent of these data collections on students, teachers, and schools, and to indicate what might be obtained from these other sources. To use them for NAEP would greatly lessen the burden on the assessment itself. Merged data sets could be made available, some to the general public, and more to researchers in restricted data files.

For many years state-level NAEP reports have included appropriate collateral data that provide a context for interpreting NAEP results; see for example the NAEP 1996 Mathematics: Report Card for the Nation and the States (Reese et al., 1997). These state contextual variables have included enrollment in elementary and secondary schools, poverty status of children from 5 to 17 years old, number of children receiving disability services, per-pupil expenditures, pupil-teacher ratios, and average teacher salaries. To the extent that these data are readily available and are helpful in setting a context for interpretation of NAEP results the practice ought to be continued. However, more effort should be made to ensure that such data are as up-to-date as possible and easily-accessible as part of NAEP reporting on the Internet.
### Figure 1

#### Overview of Current NCES Data Collections

<table>
<thead>
<tr>
<th>Data and Design Elements</th>
<th>NAEP</th>
<th>NELS</th>
<th>ELS</th>
<th>ECLS</th>
<th>TIMSS</th>
<th>CCD</th>
<th>PSUS</th>
<th>SASS</th>
<th>NHES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Elements</strong></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Student achievement</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student background</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support for learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standards and curricula</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructional practices</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>and learning resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School organization/gov-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ernance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher education and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>professional development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Financial resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School climate</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Design Elements**      | CS,L | L   | L   | L   | CS   | L   | L   | CS,L | CS   |
| Type of design (CS=cross-sectional; L=longitudinal) |     | 2-4 | 2-6 | CS   | L   | L   | CS   | L   | L   |
| Periodicity (TBD=to be determined)                  |     | 2-4 | 2-6 | CS   | L   | L   | CS   | L   | L   |
| Unit of observation (S=student; T=teacher; A=administrar; P=parent; SC=schools; D=district; ST=states; H=households) |     |     |     |     |     |     |     |     |     |
| Data collection method (S=survey; R=record analysis; I= interview; V=video; C=case study; O=other) |     |     |     |     |     |     |     |     |     |
| Population of inference (N=national; S=state; G= demographic group) |     |     |     |     |     |     |     |     |     |

NELS: National Education Longitudinal Study of 1988
ELS: Educational Longitudinal Study of 2002
ECLS: Early Childhood Longitudinal Study
TIMSS: Third International Mathematics and Science Study

NAEP Background Information Matrix

The types of descriptive information NAEP collects, reporting units, and data sources are summarized in the NAEP Background Information Matrix, which is displayed as Figure 2. The matrix is intended to assist in conceptualizing NAEP background information collections. It is organized by priorities—both for types of information and for how data should be obtained. Note that in each case information is to be obtained from reliable official records before it is sought through questionnaires.

The entries in the cells are illustrative, showing the kinds of information that are currently collected by NAEP and the various data sources (records and questionnaires) that are used. As the principles of this framework are implemented, more information will come from records, less from questionnaires. The sources with higher reliability and less respondent burden should be utilized in priority order.

The Ad Hoc Committee on NAEP Background Questions considered a proposal by Paul Barton (2002) to permit states or groups of states to add customized sets of questions to the background questionnaires. Although these might track progress on topics of particular interest and increase support for NAEP, the Committee felt strongly that the proposal should not be pursued because any customization of NAEP questionnaires would create serious logistical and quality control problems.

In constructing questionnaires it is important to place strict limits on the respondent burden they impose. The average individual response time to answer background questionnaires for each assessment, as calculated in accordance with Office of Management and Budget (OMB) procedures, shall be limited as follows: ten minutes for each student on paper-and-pencil tests, 15-minutes per student on computer-based assessments, 20 minutes for each teacher, and 30 minutes for each school. Consideration should be given to increasing student response time on paper-and-pencil questionnaires if deemed practical and productive.
### Figure 2

**NAEP Background Information Framework**

<table>
<thead>
<tr>
<th>Reporting Unit and Data Sources</th>
<th>Type of Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STUDENT</strong> School Records Questionnaire</td>
<td>Student Reporting Categories</td>
</tr>
<tr>
<td>Gender Race/ethnicity SD/LEP Race/ethnicity</td>
<td>Socio-Economic Status Core Expanded</td>
</tr>
<tr>
<td>% Free/RP lunch participation Title I</td>
<td>Other Contextual Information</td>
</tr>
<tr>
<td>Parent education Reading materials and Internet access in home</td>
<td>New enrollee Type/degree of disability</td>
</tr>
<tr>
<td>Own bedroom Parent occupation</td>
<td>Daily reading Discuss school work</td>
</tr>
<tr>
<td></td>
<td>TV-watching Absenteeism</td>
</tr>
<tr>
<td></td>
<td>Language in home After-school learning activities</td>
</tr>
<tr>
<td><strong>SCHOOL</strong> Dist/State Recds School Records CCD/Census Questionnaire</td>
<td>Subject-Specific Information</td>
</tr>
<tr>
<td>School type (public, private, charter, etc.) School ach. data Community type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Course taking in mathematics</td>
</tr>
<tr>
<td></td>
<td>Time spent on math homework</td>
</tr>
<tr>
<td></td>
<td>Good in math?</td>
</tr>
<tr>
<td></td>
<td>Graduation requirements in math and science</td>
</tr>
<tr>
<td></td>
<td>Higher level math courses</td>
</tr>
<tr>
<td></td>
<td>Graduation testing</td>
</tr>
<tr>
<td></td>
<td>Extracurricular options in math and English</td>
</tr>
<tr>
<td></td>
<td>Availability of computers for writing</td>
</tr>
<tr>
<td></td>
<td>Graduation rates</td>
</tr>
<tr>
<td></td>
<td>Post-secondary ed rates</td>
</tr>
<tr>
<td></td>
<td>Frequency of science lab work</td>
</tr>
<tr>
<td></td>
<td>Correct for spelling and grammar?</td>
</tr>
<tr>
<td></td>
<td>Frequency of science lab work</td>
</tr>
<tr>
<td><strong>TEACHER</strong> School Records Dist/State Recds Questionnaire</td>
<td></td>
</tr>
<tr>
<td>Race and Gender Experience Credentials Undergrad/Grad content training Professional Devel</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>STATE</strong> CCD/Census State Records Questionnaire</td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Non-NAEP contextual variables</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DISTRICT</strong> CCD/Census State Records District Records Questionnaire</td>
<td></td>
</tr>
<tr>
<td>Community type (urban, rural, etc.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** Information type and data sources are arranged in priority order.
Chapter Five: Using Background Data to Report NAEP Results

This chapter discusses the descriptive information that NAEP should provide, the levels of disaggregation now possible with merged national and state samples, and the importance of minimizing causal interpretations.

Use of Descriptive Information in NAEP

NAEP reporting should include contextual variables and subject-specific background information to enrich and give perspective to results. Consistent with space and operational limitations, descriptive information should be part of NAEP Report Cards and summary and highlights reports. The reports should present information on the patterns and trends of non-cognitive variables known to have a relationship to academic achievement.

In addition, special supplemental reports may be prepared that focus on particular aspects of the background data collected, topics of public interest and importance. Such reports should feature significant contextual information as well as cognitive results. Advisory committees, including a range of knowledgeable persons, may be appointed to provide input on reporting issues. In all cases, NAEP reports published by the National Center for Education Statistics must not state conclusions as to cause and effect relationships and avoid simplistic presentations unsupported by research that may imply best practice.

All background questions and data collected by NAEP should be made available on the Internet at the time of the initial release of the principal academic results or soon afterwards so the public may be able to consider them in discussing results. Complete data files should be available to researchers for further analysis.
Implementing No Child Left Behind

The intent of the No Child Left Behind Act of 2001 (P.L. 107-110) is to hold public schools accountable for closing achievement gaps between different groups of students. NAEP is asked to contribute to this end by providing an accurate measure of current levels of student achievement and to monitoring change over time.

Descriptive information about all students, but particularly on low-performing groups, would contribute powerfully to the dialogue on the challenges before American education. For example, the NAEP achievement levels focus on the segments of the performance distribution that are at or above Basic, Proficient, and Advanced. Information should also be provided about those Below Basic, who clearly have been “left behind.” e.g. the proportion having qualified teachers, receiving free or reduced-price lunch, or moving to different schools frequently, as measured by attending the same school for less than two years.

Such profiles of low-performing or high-performing students should not attempt to ascribe causation, but they would provide important information on the distribution of practices and resources that are of concern to the public and policymakers. Periodic collections of such background data could be used to track change in the distribution of these factors over time. Do the trends seem favorable or adverse to educational progress?

Disaggregation of NAEP Data

For more than three decades since it was established NAEP has provided data disaggregated by race/ethnicity, gender, school type (e.g., public/private), and community type (e.g., urban/rural). The No Child Left Behind Current law calls for disaggregation by major subgroups (when feasible) of race, ethnicity, and gender, and also by socio-economic status, disability, and limited English proficiency.

Because of the large size of the recently combined national and state NAEP samples, NAEP reports should be able to provide information disaggregated at a much greater level of detail than was possible in the program’s first decades. Pooling the data from all states, which now are required to provide NAEP samples in fourth and eighth grade reading and mathematics will produce a much enlarged national sample that will sharply reduce the number of empty cells in any cross-tabulations. Such disaggregation adds to the richness of NAEP reporting even without only a limited set of non-cognitive questions. Disaggregation is also very important for reporting on the distribution of student characteristics within the different achievement levels, as described above.
Minimizing Causal Interpretations

NAEP has often reported on the average performance of students by particular non-cognitive variables. One example, presented in many NAEP reports until the early 2000s, was the average scale score of students who watch different amounts of television each day, cf. *The Nation’s Report Card: Reading, 2000* (Donahue et al., 2001). Another example has been the average scale scores for 12th graders who report different amounts of time working at a part-time job, cf. *The Nation’s Report Card: Mathematics, 2000* (Braswell et al., 2001).

While there may be a correlation between TV-watching and reading performance, or between hours working outside school and math results, NAEP is not designed to prove cause-and-effect relationships. As a cross-sectional survey, nearly all of its data is on current activities and practices—not on the complex chain of experience in school and outside, of prior learning and achievement that all contribute heavily to current academic performance. *While the correlations may be of interest, they cannot be conclusive. But they may be cited to stimulate discussion or encourage further research. Yet, NAEP has encouraged simple causal inferences by reporting average scores for varying amounts of time spent on current activities.*

There is one important exception to the absence of data on learning-related activity over time. This is the information NAEP collects on the transcripts of high school seniors and its questionnaires on courses that students have taken and schools provide. These do show prior instruction before current exams. The trends in course taking have been of great public interest and it is reasonable to relate them to student achievement.

*NAEP reports should present information on the patterns and trends of non-cognitive variables known from other sound research to have a relationship to academic achievement. These presentations should be straightforward and impartial, and care must be taken to avoid stating conclusions as to cause and effect relationships. Further analysis of any relationships should be left to researchers.*

NAEP Data Explorer and Other Online Means of Data Dissemination

The NAEP Data Explorer should be further improved to make data more accessible to general, non-specialist users. Tables and very simple to construct charts should be prepared to present data on important topics of wide public interest. Additional means of disseminating information through new technology should be explored. These may include simple apps that would allow parents, teachers, and others to access pertinent contextual data as well as NAEP achievement results.
Chapter Six: Using NAEP in Educational Research

As a cross-sectional survey without longitudinal data, the National Assessment of Educational Progress is able to document school conditions and practices. It can report on achievement results. But it cannot properly be used to establish direct cause-and-effect relationships. Still, over the past three-four decades, NAEP has been part of two-three important research endeavors—exploring changes in the black-white test score gap since 1970, and seeking to establish the impact of state-level reforms during the 1990s, and evaluating the stringency of state standards enacted under No Child Left Behind.

By doing its main task of monitoring achievement well, NAEP has provided sound data for researchers to use. NAEP results have been critical in identifying hypotheses for other research to pursue. Its large data sets, including contextual variables, have been combined with other information to tease out meaning and policy implications, though NAEP’s own reports have properly steered clear of these activities.

The Governing Board believes the National Assessment can be of value to educational research and the interest of researchers in the assessment should be encouraged. The NCES program of secondary analysis grants for researchers to use NAEP data should continue. Educational researchers should be involved, under the auspices of NCES and its contractors, in developing NAEP background questionnaires and other data collection efforts to carry out the provisions of this framework.

This chapter considers the limitations and strengths of NAEP for educational research and discusses research that has made use of NAEP data. The chapter draws on papers by David Grissmer, senior research scientist at RAND, who has used NAEP extensively in analyzing educational factors and trends.

NAEP’s Limitations and Strengths for Research

The primary purpose of NAEP is to accurately and fairly monitor achievement over time and accurately and fairly compare achievement across states and important sub-groups of students. Beyond providing such data, any research with NAEP, particularly into the causes of academic achievement, is severely limited by its design.
As a representative sample survey, in which no individual student takes more than a small part of the full exam, NAEP has shortcomings in most of the elements commonly used to evaluate academic achievement (Podgursky, 2002):

- It provides no prior data on student achievement, and can’t be made longitudinal to do so.
- It can only collect contemporaneous information on school practices and resources, and has no way of ascertaining how students were taught or what school experiences they may have had in previous years.
- There is considerable measurement error in survey responses obtained from teachers and schools because they may well give the expected “right” answers rather than report accurately what they do.
- The current classroom practices that teachers report may be a response to student achievement levels, not the cause of such achievement, and it is difficult to disentangle causation.
- It is difficult for NAEP to get good information on socio-economic status or family background factors, but these are powerfully correlated with academic achievement, and must be controlled for in any analysis of school effects.

On the other hand, NAEP does have unique strengths and comparative advantages (Grissmer, 2003), and thus has the potential to address some important research and public policy questions with its cognitive data and background information:

- NAEP is the only data set on student achievement that has collected data from nationally representative samples of students continuously from 1969-70 to the present.
- It is the only data set that has collected academic achievement data simultaneously, repeatedly, and consistently from three separate age groups.
- It is the only data set that collects from statistically reliable samples at the state level, and within states for different types of communities (central city, suburban and rural) and for racial/ethnic groups within most states.
- NAEP has far larger sample sizes than any other nationally representative survey of student achievement, such as the National Education Longitudinal Study (NELS) and the Early Childhood Longitudinal Study (ECLS). These surveys are only approximately 10 to 20 percent as large as NAEP in any single application, and 1 to 5 percent as large as NAEP for any repeated data collection.
- NAEP is the only survey that tests a wide range of academic subjects.
• NAEP achievement measures at fourth and eighth grade fill an important void in measuring the well-being of children during this developmental period.

• NAEP generally incorporates a higher quality and unique design of test instruments, administrative procedures, and scoring methodology, compared to other data sets.

**Previous Use of NAEP in Research**

As a result of its strengths, NAEP has been used in important educational research by authors such as David Grissmer, Alan Krueger, David Armor, and Christopher Jencks. These studies point to an important comparative advantage of NAEP, namely, that it is the only representative sample data in existence on student achievement in the United States from 1969 to 2002. Thus, research into important historical questions about the effects of changing families, communities, and schools on achievement almost require NAEP data. Without NAEP, it is unlikely that the significant narrowing of the black-white score gap would be known and its possible causes the subject of research.

Similarly, NAEP data have been used to help analyze the effects of differences in resources, systemic reform initiatives, differential opportunity for learning, and other educational policies on state-level academic achievement. Such research has concluded that the rates of improvement in achievement varied markedly across states in the 1990s, and that changing resources or demographics cannot account for the gains in the states with most rapid improvement. This research points to another strong comparative advantage of NAEP. State NAEP is the only survey that includes representative samples of students in many different states, and thus plays a central role in monitoring and explaining the differences in academic achievement and achievement trends across the states. NAEP can identify where positive trends are occurring so researchers can puzzle out causation.

A review of research studies using NAEP (Grissmer, 2003) suggests that only a small proportion of the non-cognitive items collected by the assessment have been utilized in productive research. Also, such research has often supplemented NAEP with data from other sources, such as the U.S. Census and the Common Core of Data (CCD) and Schools and Staffing Survey (SASS), both conducted by the National Center for Education Statistics. However, the National Assessment played such a crucial role in these studies that they could not have been conducted without NAEP data, including some of its non-cognitive variables, principally those on socio-economic status, family structure, and school resources.

On the other hand, NAEP data have also been misused for simplistic and weak research. Many background items on school practice and student behavior have been used in a simplistic way to imply a direct, causal relationship to achievement while ignoring the complex mix of other, more fundamental factors that may well have a stronger impact. NAEP has encouraged such associations by presenting one-way
tabulations in its reports, e.g. average scale score by hours of television watched, type of reading instruction, or books read per week, and these have been disseminated widely to support particular beliefs or public policy positions. Simple, single variable linkages can often be misleading because of the strong correlations between many background variables, particularly with socio-economic status, prior academic achievement, or family background. They should only be included in NAEP reports when there is strong justification based on previous research.

Also, most of the hundreds of background questions in NAEP have never been used for either public reporting or research. Many come from the early 1980s, and would be difficult to justify in a sound research design today.

Secondary Analysis Grants and District Samples

For many years, NCES has been making awards to education researchers for secondary analyses of NAEP data. These explored a range of topics, often in combination with other data sets. Many of the studies have focused on state-to-state differences in student achievement and the impact of state-level policies, relying on NAEP academic data, a few background questions for SES controls, and much additional information from other sources. The program has been valuable as a means of encouraging the use of NAEP for research, and, in a few cases, notably the Grissmer studies, has had considerable impact. As in any grant program, all findings are the responsibility of the individual researchers, not of the agency making the grant.

The program should continue, and now that NCES has become part of the Institute for Education Sciences in 2003, the leadership of the new agency should ensure that separate NAEP analysis grants were absorbed in a more general research program, are aligned with the research priorities of the Institute. We believe this program should increase awards that make use of NAEP data. Efforts should be made through training and other small-scale grants to expand capabilities for using NAEP in productive education research.

In addition, data from the school district NAEP samples in the Trial Urban District Assessment, which started in 2002, will provide important new opportunities for research. NAEP results for school districts can readily be combined with Census data, which include pertinent information on family background and socio-economic status. The school district samples can also be tied to important education policy variables, such as per pupil spending, for which information is available at this level but not for schools.

The primary purpose of NAEP is to provide fair and accurate information on student achievement. Its primary audience is the American public. The Governing Board believes that in serving its purpose and audience well, NAEP can contribute to educational research. It welcomes the interest and efforts of researchers.
Chapter Seven: Review and Improvement of Non-cognitive Questions

This chapter discusses several mechanisms for the review and improvement of NAEP’s non-cognitive questions and for implementation of the NAEP Background Information Framework.

Independent Validity Studies

Since the early 1990s NAEP has had the benefit of independent outside advice on topics of urgency or interest. These studies have been very helpful to the Governing Board and NCES as they made decisions about the future of the NAEP program. For example, several years ago some research was conducted to examine the possibility of combining the NAEP national and state samples to achieve more efficiency and cost-savings. Starting in 2003 NAEP moved in that direction. The decisions surrounding the change, however, were only as good as the research that bolsters it. The work of the current NAEP Validity Panel, in conjunction with the current NAEP operations contractors, has contributed significantly to making the change possible.

The value of this kind of applied research cannot be overestimated. Neither can the value of the independent nature of this work. The NAEP program is very large and complex and demands a commitment of many resources from the NAEP contractors. NAEP contractors should not be burdened with conducting simultaneous research studies while carrying out the requirements of the operations contracts. There is a precedent for this approach in the current separation of responsibilities for operations and research in separate NAEP contracts. There are two reasons why independent validity studies on topics associated with the non-cognitive framework are recommended. First, there are some non-cognitive variables that will need validation, particularly if those variables are new or are new composite indexes of existing variables. Second, following the approach already established for the NAEP cognitive components, recommendations from such research studies must be truly independent and free from any conflict of interest.
Review of the Background Information Framework

The background information framework should be reviewed on a periodic basis. The NAEP cognitive frameworks are reviewed every ten years. This policy was adopted at the time of the NAEP redesign in 1996. Reviewing a NAEP framework can result in major revision, minor revision, or even no revision and re-adoption. The framework may be updated as needed. A thorough review of the Background Information Framework should be undertaken since the background framework is a new undertaking; a required review after five years is appropriate with additional reviews every ten years thereafter.
References


Armor, D.J. (December 18, 2002). *Comments on NAEP non-cognitive questions*. Available from the National Assessment Governing Board, 800 N. Capitol Street, NW, Washington, DC 20002.


Appendix

Adopted May 18, 2002

National Assessment Governing Board

Policy Statement on Collection and Reporting of Background Data by the National Assessment of Educational Progress

INTRODUCTION

As the Nation’s Report Card, the National Assessment of Educational Progress (NAEP) is an ongoing, Congressionally-authorized program to collect data through surveys on the academic knowledge and skills of American students. Its primary goal is to report fair and accurate information on student achievement in reading, mathematics, and other subjects taught in elementary and secondary schools. This information is to be made available in a clear and timely manner to members of the public, policymakers, and educators throughout the country.

Since it began in 1969-70, NAEP has administered, in addition to cognitive questions, background questionnaires that provide information for reporting categories and collect non-cognitive data on students, their family background, teachers, and schools. These have enriched reporting of the National Assessment and increased the precision of NAEP results. The background data have also been used in secondary analyses. However, because NAEP tests a cross-section of students at a particular time with no follow-up of the students tested, the assessment can only show correlations or associations rather than causal relationships between background factors and achievement.

By statute (P.L. 107-110), the National Assessment shall include, “whenever feasible, information collected, cross-tabulated, compared, and reported by race, ethnicity, socioeconomic status, gender, disability, and limited English proficiency.” The statute provides that NAEP may “not evaluate or assess personal or family beliefs and attitudes” and may “only collect information that is directly related to the appraisal of academic achievement and to the fair and accurate presentation of such information.” These provisions are intended to prevent intrusive, inappropriate, or unnecessary questions being asked about students and their families.
The law requires that the Governing Board take steps to ensure that all NAEP questions are “free from racial, cultural, gender, or regional bias, and are secular, neutral, and non-ideological.” However, a House-Senate Conference report, accompanying the legislation, says the law does not preclude the use of “non-intrusive, non-cognitive questions,” with a direct relationship to academic achievement.

The National Assessment is conducted by the Commissioner of Education Statistics under the policy guidance of the National Assessment Governing Board. The Board’s specific areas of responsibility include: (1) assessment objectives and test specifications; (2) the methodology of the assessment; (3) guidelines for reporting and disseminating results; and (4) “appropriate actions needed to improve the form, content, use, and reporting” of the National Assessment. Under the statute, the Board has “final authority” on the appropriateness of all NAEP items—both cognitive and non-cognitive.

To carry out these responsibilities, the National Assessment Governing Board hereby adopts guiding principles, policies, and procedures for the collection and reporting of background data by the National Assessment of Educational Progress.

**GUIDING PRINCIPLES**

1. Background data on students, teachers, and schools is needed to fulfill the statutory requirement that NAEP include information, whenever feasible, disaggregated by race or ethnicity, socioeconomic status, gender, disability, and limited English proficiency. In addition, background data is collected to enrich the reporting of NAEP results by examining factors related to academic achievement. However, the collection of such data should be limited, and the burden on respondents kept to a minimum. It must always be considered in light of NAEP’s primary purpose: providing sound, timely information on the academic achievement of American students.

2. All background questions must be directly related to academic achievement or to the fair and accurate presentation of achievement results.

3. Issues of cost, benefit, appropriateness, and burden shall be carefully considered in determining the background questions to be asked and the samples to which they shall be administered.

4. In accordance with law, questions shall be non-intrusive and free from bias, and must be secular, neutral, and non-ideological.

5. No personally identifiable information shall be included in NAEP reports or data releases.

6. Decisions on the retention or addition of background items shall draw on technical studies on the reliability and validity of current and proposed
questions and on the contribution such items make to the precision of NAEP results.

7. Consideration should be given to obtaining background information from non-NAEP sources and to avoiding duplication with other federal surveys.

8. Questionnaires should be revised to keep background questions timely and related to academic achievement. Those questions showing little change over time and/or a stable relationship to achievement should be deleted or asked less frequently and to limited samples, unless required to assure the precision of NAEP results.

9. Questions should not address personal feelings and attitudes.

10. Since security considerations do not apply, background questionnaires shall be readily available to the public.

11. Interpretation of results shall be limited in official NAEP reports and must be strongly supported by NAEP data. Because of the survey nature of the assessment, reports may show correlations and generate hypotheses, but may not state conclusions as to cause and effect relationships.

12. Background questions for NAEP assessments shall be prepared in accordance with frameworks and specifications adopted by the Governing Board.

13. The Governing Board shall review and approve all background items before they are administered in NAEP surveys or pilot and field tests.

POLICIES AND PROCEDURES

1. Framework and Specifications

The Governing Board will adopt a general framework for background questionnaires and specifications for the questions on selected topics and in specific subject areas.

Since this is a new area of responsibility for the Board, the process of developing a framework for background questions and specifications will begin with commissioned papers on relevant issues, such as the reliability and validity of current background questions, their contribution to improving the precision of NAEP results, their value and limitations for educational research, and changes that may be needed in response to the No Child Left Behind legislation. Following consideration of these issues, the Board will define the scope of
background questionnaires and adopt a process for preparing a framework and specifications. This work will include the active participation of teachers, education researchers, state and local school administrators, assessment specialists, parents of children in elementary and secondary schools, and interested members of the public.

2. Background Question Development

In preparing background questions, the National Center for Education Statistics shall follow adopted frameworks and specifications, and consider the review criteria adopted by the Governing Board. NCES may use cognitive laboratories of students, teachers, and school officials to help determine the clarity and burden of proposed questions. Ad hoc advisory committees may also be established, comprised of teachers, parents, technical experts, and others interested in NAEP. Steps shall be taken to determine the reliability of questions used.

3. Governing Board Review and Approval of Background Questions

Background questions for all NAEP pilot tests, field tests, and operational use shall be reviewed and approved by the Governing Board. The category of respondents, e.g. students, schools, and grade level, shall clearly be designated, as will the NAEP samples, e.g. national, state, or district, in which the questions will be asked.

For each questionnaire there shall be an explanation of its intended use in NAEP reporting and analysis and of the hypothesized relationships between the background items and student achievement that demonstrates the need to know such information. Technical data shall be presented on the reliability and validity of questions and, if applicable, on their contribution to improving the precision of NAEP results. The Board will use the explanations and data presented along with the review criteria in this policy statement in determining the appropriateness of background questions.

The Reporting and Dissemination Committee will have primary responsibility for the review and approval of background questions. The Assessment Development Committee will participate in the approval of questions relating to specific subject-matter assessments. Ad hoc committees of Board members may be established by the Board Chairman for background question review. Questions may also be reviewed by external advisors, including teachers, parents, and technical experts. Recommendations on background questionnaires shall be subject to final approval by the full Governing Board.
4. Criteria for Governing Board Review

The following criteria for review and approval of background questions are based on the most recent revision of the authorizing statute of the National Assessment of Educational Progress (P.L. 107-110) and the Guiding Principles of this policy statement:

A. Background information is needed to fulfill the statutory requirement that NAEP report and analyze achievement data, whenever feasible, disaggregated by race or ethnicity, gender, socio-economic status, disability, and limited English proficiency. Non-cognitive data may enrich the reporting and analysis of academic results, but the collection of such data should be limited and the burden on respondents kept to a minimum.

B. All background questions must be related to the primary purpose of NAEP: the fair and accurate presentation of academic achievement results.

C. Any questions on conditions beyond the school must be non-intrusive and focused on academic achievement and related factors.

D. Questions shall be free from racial, cultural, gender, or regional bias.

E. All questions must be secular, neutral, and non-ideological. Definitions of these terms, accompanied by clarifying examples, are presented in Appendix A, as adopted in the Governing Board Policy on NAEP Item Development and Review.

F. NAEP must not evaluate or assess personal feelings or family beliefs and attitudes unless such questions are non-intrusive and have a demonstrated relationship to academic achievement.

G. Issues of cost, benefit, appropriateness, and burden shall be carefully considered in determining which questions to include in background questionnaires. These factors must also be considered in determining the frequency with which various questions shall be administered and whether they shall be included in both national and state samples.

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

5. Public Access to Background Questions

Since security considerations do not apply, all background questionnaires shall be readily available to parents, teachers, state and local officials, and interested members of the public. Such questionnaires shall be available before
field tests and operational assessments or at any other time members of the public wish to obtain them. Background questions in operational use shall be posted on the Internet prior to each assessment, accompanied by explanations and rationales.

6. Reporting of Background Information

The presentation of background data in official NAEP reports shall be straightforward and impartial. Because of the survey nature of the assessment, reports may show correlations and generate hypotheses, but may not state conclusions as to cause and effect relationships. Any composite indices including demographic and socioeconomic factors shall be presented to the Board for approval before use as reporting categories in NAEP data releases and reports.

Background data should be available for extensive secondary analyses by scholars and researchers, who are responsible for conclusions reached. Responses to background questions shall be presented and tabulated on the Internet, although, if necessary, posting may be delayed for a brief period after release of the principal NAEP results.
APPENDIX A: Definitions of Secular, Neutral, and Non-ideological Item Review Criteria

From Governing Board Policy on NAEP Item Development and Review—5/18/02

Items shall be secular, neutral, and non-ideological. Neither NAEP nor its questions shall advocate a particular religious belief or political stance. Where appropriate, NAEP questions may deal with religious and political issues in a fair and objective way. The following definitions shall apply to the review of all NAEP test questions, reading passages, and supplementary materials used in the assessment:

**Secular** — NAEP questions will not contain language that advocates or opposes any particular religious views or beliefs, nor will items compare one religion unfavorably to another. However, items may contain references to religions, religious symbolism, or members of religious groups where appropriate.

Examples: The following phrases would be acceptable: “shaped like a Christmas tree,” “religious tolerance is one of the key aspects of a free society,” “Dr. Martin Luther King, Jr. was a Baptist minister,” or “Hinduism is the predominant religion in India.”

**Neutral and Non-ideological** — Items will not advocate for a particular political party or partisan issue, for any specific legislative or electoral result, or for a single perspective on a controversial issue. An item may ask students to explain both sides of a debate, or it may ask them to analyze an issue, or to explain the arguments of proponents or opponents, without requiring students to endorse personally the position they are describing. Item writers should have the flexibility to develop questions that measure important knowledge and skills without requiring both pro and con responses to every item.

Examples: Students may be asked to compare and contrast positions on states rights, based on excerpts from speeches by X and Y; to analyze the themes of Franklin D. Roosevelt’s first and second inaugural addresses; to identify the purpose of the Monroe Doctrine; or to select a position on the issue of suburban growth and cite evidence to support this position. Or, students may be asked to provide arguments either for or against Woodrow Wilson’s decision to enter World War I. A NAEP question could ask students to summarize the dissenting opinion in a landmark Supreme Court case.

The criteria of neutral and non-ideological also pertain to decisions about the pool of test questions in a subject area, taken as a whole. The Board shall review the entire item pool for a subject area to ensure that it is balanced in terms of the perspectives and issues presented.
Acknowledgements

The Ad Hoc Committee on NAEP Background Questions of the National Assessment Governing Board was chaired by Board member John H. Stevens. He also serves as chairman of the Board’s standing Committee on Reporting and Dissemination, which has responsibility for reviewing all core NAEP background questionnaires and making recommendations on them to the full Board.

The Ad Hoc Committee also included members of the Board’s two other standing Committees—Assessment Development and Standards, Design, and Methodology—with a wide range of backgrounds and perspectives: Amanda Avallone, Dwight Evans, Thomas Fisher, Sheila Ford, Jo Ann Pottorff, and Sister Lourdes Sheehan. The Board Chairman, Darvin Winick, participated in many of the lively discussions that marked the Committee’s deliberations.

Among the many discussants and presenters at the workshop and public forum, we wish to recognize the care and thoughtfulness of Robert Mislevy, of the University of Maryland, and Harold Wenglinsky, of Baruch College of the City University of New York. The comments submitted by Sandra Feldman, president of the American Federation of Teachers, were particularly perceptive and helpful. This project also benefited greatly from the continuing advice and insight of Paul Barton and David Grissmer, both of whom have used NAEP data for many years to understand and explain American education to its public.
Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting

INTRODUCTION

By statute, the purpose of the National Assessment of Educational Progress is to provide a “fair and accurate” measure of student achievement and achievement trends. Academic or cognitive questions are its primary focus; the American public is its primary audience. However, in addition to reporting on what American students know and can do, NAEP has collected data for more than 40 years that provide a context for reporting and interpreting achievement results. According to the statute, such factors, both in and out of school, must be “directly related to the appraisal of academic achievement.”

In each assessment NAEP administers background questionnaires for students, their teachers, and schools. The questionnaires deal with educational experiences and other factors, such as teacher training or out-of-school learning activities, that are related to academic achievement. Data on several hundred background or noncognitive variables are available on the Internet through the NAEP Data Explorer. However, for more than a decade, little use has been made of this information in NAEP reports. The data have received minimal attention and had little impact despite the considerable efforts expended in developing and approving questionnaires and collecting and tabulating responses.

In October 2011 the National Assessment Governing Board convened an expert panel to recommend how to make better use of existing NAEP background questions and to propose an analytic agenda for additional topics and questions that would be useful in developing education policy and of value to the public. The panel report, entitled, NAEP Background Questions: An Underused National Resource, was presented to the Board in March 2012 by Marshall Smith, former U.S. Under Secretary of Education, who chaired the six-member panel.

Many of the panel recommendations build on the Background Information Framework for the National Assessment of Educational Progress, adopted by the Governing Board after it received final authority from Congress over non-cognitive items on the assessment. The framework was adopted in 2003, but has not been fully implemented.
The following policies are based on recommendations by the expert panel. The Board has also taken into consideration a wide range of public comment and the analysis provided by the National Center for Education Statistics.

It is important to understand that the National Assessment is not designed to show cause-and-effect relationships. Its data should not be used to “prove” what schools should do. But, as the Background Information Framework declares, NAEP’s “descriptions of the educational circumstances of students…, considered in light of research from other sources, may provide important information for public discussion and policy action.” The Board believes the National Assessment should improve upon its efforts to collect contextual information and present it clearly to the public, which will add to NAEP’s value to the nation.

POLICY PRINCIPLES

1. NAEP reporting should be enriched by greater use of contextual data derived from background or non-cognitive questions asked of students, teachers, and schools. Such data will be used both in regular Report Cards and in special focused reports. [New Foreword, pp. 5 and 8]

2. Reporting of background data will describe patterns and trends, including the educational experiences of different groups of students. Care should be taken not to suggest causation. [Chapter 5, pp. 37-39; also pp. 5 and 8]

3. Detailed frameworks will be published with the theoretical rationale and research evidence that support the selection of topics and questions in background questionnaires and their connection to student achievement. Such frameworks should be updated for each assessment cycle and provide the basis for new topics and questions. [p. 12]

4. An ad hoc committee of the Board will be established for one year to monitor implementation of this resolution, review the NAEP Background Information Framework, and recommend a permanent arrangement for Board consideration of background questions and the reporting of contextual data in NAEP.

IMPLEMENTATION GUIDELINES

For Questions and Questionnaires

1. Clusters of questions will be developed on important topics of continuing interest, such as student motivation and control over the environment, use of technology, and out-of-school learning, which could be used regularly or rotated across assessment cycles. [pp. 12 and 30]

2. Modules will be prepared for special one-time studies to provide descriptive information on issues of current policy interest. [p.29]
3. A thorough review will be conducted to eliminate duplicative or low-priority questions. Unproductive topics and questions will be dropped. [p. 12]

4. NAEP will include background questions from international assessments, such as PISA and TIMSS, to obtain direct comparisons of states and TUDA districts to educational practices in other countries. [pp. 10 and 23]

5. Because of the value of preserving trends, consistent wording of questions should be maintained on topics of continuing interest. Changes in wording must be justified. However, as practices and circumstances change, new questions will be introduced in a timely manner to gather data on topics of current interest. [pp. 10 and 23]

6. The development and use of improved measures of socio-economic status (SES) will be accelerated, including further exploration of an SES index for NAEP reporting. [pp. 11 and 27]

For Data Collection

7. The maximum time for students to answer the background questionnaire will be increased from 10 to 15 minutes on new computer-based assessments. Consideration should be given to a similar increase in paper-and-pencil assessments. [pp. 12 and 35]

8. Whenever feasible, assessment samples should be divided (spiral sampling) and background questions rotated in different years in order to cover more topics without increasing respondent burden. These practices will be initiated in the assessments of reading and mathematics, which are conducted frequently, and considered for other subject areas if the frequency of testing permits. [pp. 12 and 30]

For Reporting

9. Special focused reports with data through the 2013 assessment will be issued on the following topics: private schools, charter schools, gender gaps, and black male students. Reports shall include significant contextual information as well as cognitive results. Advisory committees, composed of a range of knowledgeable persons, may be appointed to provide input on reporting issues. [p. 37]

10. Exploratory analyses will be carried out to determine if existing background questions may form the basis for additional focused reports. Such reports may be issued by the Governing Board as well as by the National Center for Education Statistics.

11. The NAEP Data Explorer should be further improved to make data more accessible to general, non-specialist users. Tables and very simple-to-construct charts will be prepared to present data on important topics of wide public interest. Additional means of disseminating information, using new technology such as simple apps that would allow parents, teachers, and others to access background and achievement data, will be explored. [p. 39]
In October 2011, eight years after adoption of the NAEP Background Information Framework, the National Assessment Governing Board convened an expert panel to study the NAEP background questions and recommend possible changes. The six-member group was chaired by Marshall S. Smith, former dean of the Graduate School of Education at Stanford University and a former U.S. Under Secretary of Education. The panel’s report, presented to the Board in March 2012, called the background questions “a potentially important but largely underused national resource.” (Smith, et al., NAEP Background Questions: An Underused National Resource. A Report to the National Assessment Governing Board by the Expert Panel on Strengthening the NAEP Background Questions)

The report described the information gathered through background questionnaires as “a rich collection of student, teacher and school responses…that can help in understanding the context for NAEP achievement results and give insights into how to improve them.” But it said over the past decade the questionnaires had been cut back and little used in NAEP reports. It urged NAEP to “restore and improve upon” its practice of the early 1990s by “making much greater use of background data, but do so in a more sound and research-supported way.”

With “proper attention,” the expert panel declared, NAEP’s contextual data “could provide rich insights into a wide range of important issues about the nature and quality of American primary and secondary education.”

After gathering public comment, the Governing Board adopted a Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting. The policy, approved in August 2012, was based on recommendations by the expert panel and provided for an important change in emphasis:

- NAEP reporting should make greater use of contextual data in both regular Report Cards and special focused reports.

- The reporting of background data will describe patterns and trends, including the educational experiences of different groups of students. Such information will enrich NAEP reporting, but care should be taken not to suggest causation.

- Detailed frameworks will be published with the theoretical rationale and research evidence that support the selection of topics and questions and their connection to student achievement.

- Modules will be prepared for special studies to provide descriptive information on issues of current policy interest.
NAEP will include contextual questions from international assessments to obtain direct comparisons of states and TUDA districts with educational practices in other countries.

The Board resolution included a set of implementation guidelines. It also established an ad hoc committee, which reviewed the framework.

The committee believes the approach adopted in the 2003 framework remains sound, as do many of its detailed provisions, but some updating is needed to reflect changes over the past decade and the August 2012 policy statement. The revisions are based largely on the 2012 resolution and are incorporated in the text that follows.

As NAEP makes the transition from paper-and-pencil to a computer-delivered assessment, the Board hopes the new technology will help make possible the range of topics and flexibility in sampling envisioned a decade ago while limiting the burden on students, teachers, and schools.
NOTE TO Ad Hoc Committee on Background Information on Terminology for Background Questions

The Committee may wish to discuss the issue of whether to continue to employ the term “background questions” in referring to NAEP’s non-cognitive or contextual data questionnaires. NAEP has used “background questions” without any significant problems for about 40 years but the term may be seen by some to denote the intrusive personal questions in a “background investigation.” With this concern in mind NCES has avoided using the word “background” in recent materials for students, teachers, and schools. Instead it simply refers to these non-cognitive surveys as “student questionnaires” or “teacher questionnaires” or “school questionnaires.” However, articles on the NAEP web site use the terms “background data” and “background items” as well as “non-cognitive” items and questionnaires.

By law, NAEP is authorized only to collect background information that is “directly related to the appraisal of academic achievement, and to the fair and accurate presentation” of assessment results. NAEP must not evaluate or assess “personal or family beliefs and attitudes.” The assessment may not disclose “personally identifiable information” and cannot report data on individual students or schools. Under Board policy, adopted in 2002 and retained in the 2003 framework and new update, any questions on student attitudes toward school or various academic subjects, such as reading or science, must be “non-intrusive and have a demonstrated relationship to academic achievement.”

In the major international academic surveys—TIMSS, PIRLS, and PISA—the non-cognitive questionnaires are referred to in different documents as either background, context, or contextual questionnaires or simply as questionnaires. PISA, for example, publishes what it calls a Questionnaire Framework, which explains in its introduction that the testing program collects “background information on families, students, and schools that might be linked to [academic] achievement.” TIMSS publishes a Contextual Framework, which “provides the foundation for the information that will be collected via the TIMSS background questionnaires.”

The Ad Hoc Committee may wish to make a recommendation to the full Governing Board about which terminology NAEP should use.
National Assessment Governing Board

Staff Recommendation on How to Organize Board Consideration of Background Questions and the Use of Contextual Data in NAEP Reports

July 8, 2013

Since 2002 the law authorizing the National Assessment of Educational Progress has granted “final authority” on the appropriateness of all assessment items to the National Assessment Governing Board. The Board has delegated responsibility for reviewing and approving cognitive questions to its Assessment Development Committee, which also is in charge of developing test frameworks. However, consideration of non-cognitive or background questions has been divided.

The Assessment Development Committee (ADC) reviews the largest number, which are related to the specific subjects being assessed and based on recommendations in the assessment frameworks. A second committee, Reporting and Dissemination, reviews the core background questions asked of students, teachers, and schools in all NAEP subjects. These deal with the reporting categories of race, socio-economic status, limited English proficiency, and disability. They also concern a range of factors that provide a context for results and may affect academic achievement, such as student mobility, absenteeism, school safety, and course offerings and governance in regular and charter schools. The Reporting and Dissemination Committee has also overseen the general planning and procedures for developing background questions by the National Center for Education Statistics (NCES) and its contractors.

The Expert Panel on Strengthening NAEP Background Questions, in its report to the Board in February 2012, said this divided arrangement was unsatisfactory. It said coordination was poor and the background questions were of “secondary interest” to both committees. The panel suggested establishing a “unified standing committee” to monitor all aspects of background question development.

Board staff acknowledges the concerns that led to the expert panel recommendation but believes establishing a fifth standing committee of Board members would be unwieldy and unproductive. Instead, the staff believes there should be much greater coordination of the work on background questions by the ADC and Reporting committees. The committees should hold a joint meeting at least once a year on important policy issues and also to ensure that background questionnaires are of reasonable length and that the data obtained are used productively. Joint meetings on specific issues might be held more frequently and members of each committee should participate occasionally in item reviews conducted by the other. Also, members of the Committee on Standards, Design and Methodology (COSDAM) may participate in the joint meetings when their technical expertise is relevant to the issues discussed.

The Board policy, adopted in August 2012, states clearly that greater use should be made of contextual data as a means of enriching National Assessment reports and adding to the value of NAEP for public. This should enhance the importance of background or context questionnaires in the work of the two Board committees. The greater coordination between them should ensure against overlap and lead to more effective monitoring of questionnaire development and reporting.
NOTE TO Ad Hoc Committee on Background Information
on Using NAEP Data for Key Education Indicators

As authorized by the Policy Statement on Background Data last year, two Governing Board consultants are preparing an exploratory analysis on using NAEP data for key education indicators. The purpose of this project is to illustrate the usefulness of NAEP in developing a limited number of indicators to represent crucial components of the education system and their interrelationships. The key idea is that instead of starting with background variables and looking for education issues they might address, there should first be a framework of important education policy issues and objectives that can be used to identify relevant background variables.

This work is being undertaken by Alan Ginsburg, former Director of Policy and Program Studies at the U.S. Department of Education, who has completed several other exploratory analyses for the Board, and Marshall S. Smith, former Dean of the Stanford University Graduate School of Education, who chaired the Board's Expert Panel on Strengthening NAEP Background Questions, which presented its report in February 2012.

As explained in the statement of work for this project, an education indicator is an individual or composite statistic that measures progress toward an educational objective and is useful in a policy context. Such objectives are concerned not only with student performance but with the quality, equity, and efficiency of the education system in supporting academic achievement. One possible indicator might be the percentage of 8th grade science students with a teacher who majored or minored in science in college. Others might be the extent of severe absenteeism or the use of technology in science instruction.

At this meeting Alan Ginsburg will discuss the indicator project and present preliminary results. Additional information will be distributed under separate cover before the meeting.
<table>
<thead>
<tr>
<th>Time</th>
<th>Agenda Item</th>
</tr>
</thead>
</table>
| 4:30 pm | Welcome and Agenda Overview  
*David Driscoll, Chair* |
| 4:35 pm | Committee Issues and Challenges  
Ad Hoc Committee on Background Information: Overview of Committee Recommendations  
*Terry Holliday, Ad Hoc Committee Chair*  
ADC: Grade 4 Writing Information  
*Alan Friedman, ADC Chair*  
COSDAM: Achievement Level-Setting for TEL  
*Lou Fabrizio, COSDAM Chair*  
Reporting & Dissemination: Review of Board Policy on Reporting, Release, and Dissemination of NAEP Results  
*Andrés Alonso, R & D Vice Chair*  
Nominations: New Web Outreach Strategies  
*Tonya Miles, Nominations Committee Chair* |
| 4:55 pm | **ACTION ITEM**  
Nomination of Vice Chair for Term Beginning October 1, 2013  
*Alan Friedman*  
*Attachment A* |
| 5:00 pm | **ACTION ITEM**  
Draft Policy Statement on the Conduct and Reporting of NAEP  
*Alan Friedman and Shannon Garrison, Executive Committee*  
*Attachment B* |
| 5:05 pm | Interpreting NAEP Results Using Preparedness Research Findings  
*Lou Fabrizio*  
*Attachment C* |
| 5:10 pm | **ACTION ITEM**  
NAEP Schedule of Assessments  
*Ray Fields, Assistant Director for Policy and Research*  
*Attachment D* |
| **CLOSED SESSION 5:15 – 6:30 p.m.** | |
| 5:15 pm | NAEP Contracts, Budget, and Schedule for 2013 and Beyond  
*Cornelia Orr, Executive Director*  
*Peggy Carr, Associate Commissioner, NCES* |
Nomination of Governing Board Vice Chair for the Term October 1, 2013 through September 30, 2014

ACTION ITEM

While the Governing Board Chair is appointed by the U.S. Secretary of Education, the Vice Chair is elected annually by the Governing Board from among its current members. The practice of the Board electing its Vice Chair has been in effect since the Board’s inception in 1988 and is incorporated in the Board’s By-laws.

Because Governing Board terms begin each October 1, the Governing Board elects the Vice Chair annually at the quarterly Board meeting conducted the preceding August. The Executive Committee is responsible for nominating a candidate for consideration by the full Board. The nomination process is set in motion each year around the May Board meeting.

Following the May 16, 2013 Executive Committee meeting, Chairman Driscoll asked Alan Friedman to lead the process for nominating the Vice Chair for the term beginning October 1, 2013.

At the August 1, 2013 Executive Committee meeting, Mr. Friedman will report on the results of the nominations process. He will present a candidate for Vice Chair for consideration by the Executive Committee.

The Executive Committee will act on that recommendation and decide on the individual to be nominated for election by the Governing Board. The election will occur before the conclusion of the Board meeting on August 3, 2013.
General Policy: Conducting and Reporting
The National Assessment of Educational Progress

At its November 29, 2012 meeting, the Executive Committee reviewed a staff-prepared “track-changes” version of the Governing Board policy entitled “Redesigning the National Assessment of Educational Progress” (adopted in 1996), having decided earlier that the policy should be updated. The “track-changes” version highlighted elements of the 1996 policy that would benefit from updating or revision.

The Executive Committee directed staff to present recommendations for changes to the policy for discussion by the full Board. The Governing Board discussed a staff-prepared policy analysis at the February 28-March 2, 2013 quarterly meeting and directed staff to prepare a draft policy statement for discussion at the May 16-18, 2013 meeting. Following the February-March 2013 meeting, chairman Driscoll appointed a subcommittee of Executive Committee members to guide the drafting of the policy statement. Lou Fabrizio, Alan Friedman, and Shannon Garrison are the members of the subcommittee.

The subcommittee presented a draft policy statement for full Board discussion at the May 2013 meeting. The draft policy statement has been revised with the intent of faithfully addressing the Board members’ comments and suggestions. It also includes suggestions NCES has made at each step along the way. The revised draft appears on the following pages, with revisions indicated in “track changes” format.

The full Board will discuss the revised draft on August 2, 2013, with action by the Board to adopt a final policy statement expected on August 3.
General Policy: Conducting and Reporting
The National Assessment of Educational Progress

Foreword
This policy is a guide for those responsible for the National Assessment of Educational Progress (NAEP)—the Nation’s Report Card. These are the members and staff of the National Assessment Governing Board (NAGB), which oversees NAEP; the Commissioner and staff of the National Center for Education Statistics (NCES) responsible for NAEP operations; and the staff of the contractors that carry out NAEP.

NAEP performs an exceptional public service. It provides trusted information on the performance and progress of the nation’s elementary and secondary schools and school children. Over the course of its history, a set of essential, enduring principles and values have become embodied in NAEP. These principles and values are set forth below.

Introduction
Thomas Jefferson said “If a nation expects to be ignorant and free in a state of civilization, it expects what never was and never will be.” Horace Mann, the advocate for the Common School, said “Education…beyond all other devices of human origin, is the great equalizer of the conditions of men—the balance-wheel of the social machinery.” John F. Kennedy, paraphrasing H.G. Wells, said “…the course of civilization is a race between catastrophe and education. In a democracy such as ours, we must make sure that education wins the race.”

The nation’s leaders have long recognized education as a foundation for democracy. Education fosters capable civic participation; supports individual human development; promotes national, state, and individual economic well-being; and advances national security. Providing for the education of its citizens and monitoring their levels of achievement are key functions of states and the nation. NAEP was established for the latter function—to monitor student achievement.

History and Evolution of the National Assessment of Educational Progress
The first U.S. Department of Education was created by Congress in 1867. It was the early predecessor of NCES, established to “[collect] such statistics…as shall show the condition and progress of education in the several States and Territories, and [diffuse] such information…as shall aid the people of the United States in the establishment and maintenance of efficient school systems….” For more than 100 years, this Department and its successors provided information on the number of schools, school districts, student enrollment, revenues, expenditures, and the like, but collected no information on student achievement.

This began to change in 1963. U.S. Commissioner of Education Francis Keppel was testifying before the House Appropriations Committee on the FY 1964 budget for education. A committee member asked Keppel a simple question—“How well are U.S. students achieving?” Keppel was not able to answer the question because there was no source of information to answer it.
The question—important at any time—was raised in the context of the Cold War and concern about national security. What was then the Soviet Union had launched Sputnik by missile on October 4, 1957. This feat caused the nation’s leaders to fear that the United States lacked sufficient scientific and engineering capability to compete and keep the country safe.

Keppel recognized the threat inherent in failing to know the levels of U.S. student achievement. It set him on a path that led to the creation of the National Assessment of Educational Progress and the conduct of the first assessment—in science—in 1969. Keppel began by forming a committee to design a national assessment. The committee was established late in 1963. It was funded in large part by the Carnegie Corporation and led by Ralph Tyler, the preeminent education researcher of his day.

Some leaders in school administration, curriculum, and the teaching force opposed the idea of a federal assessment of student achievement. They were concerned that a federal test would lead to federal intrusion in school curriculum and accountability, responsibilities of state and local education officials. This is a recurring theme in the evolution of NAEP: finding the right way to serve the national interest as a monitor of student achievement while honoring state and local authority over schools.

The proposed design addressed the opponents’ concerns. The Education Commission of the States (ECS) would carry out the assessment with funding from the U.S. Office of Education. This put authority in a state-based organization and placed the federal role at arms-length from the assessment. Decisions about content and subjects to test would be made by ECS. There would be no student, school, district, or state-level results. Data would be reported for the nation and for regions of the country. Student samples would be age-based rather than grade-based. Together, these addressed concerns that the National Assessment would lead to a national curriculum and federal entanglement in school governance.

The education landscape has changed since the initial assessment in 1969. Accordingly, the National Assessment has evolved. Where there was some opposition in the beginning, NAEP has earned trust, is recognized for its quality, and is highly valued. Little known in the early years except by interested researchers, NAEP results have become widely used by policymakers and education leaders and are featured by the news media, with ever increasing awareness by the public, teachers, school administrators and others. The original design was an innovation responsive to the times. Since then, many responsive innovations have been made in NAEP’s governance, the subjects assessed, item types, test procedures, and the use of information and communication technology. State level and grade-based reporting are now a regular part of NAEP.

Change in the education environment continues. Change and innovation in response to the needs of the time are hallmarks of NAEP. These are balanced against the imperative to maintain NAEP’s independence as a stable measure for reporting achievement trends. Balancing competing goals is a continual challenge to NAEP, a tension that is the source of its continual creative evolution to better serve the American public.
Policy Statement

Purpose and Characteristics of the National Assessment of Educational Progress

The National Assessment of Educational Progress is an independent monitor of student academic achievement in the United States at the elementary and secondary levels. It is a trusted, valid source of data on student achievement in public and private schools. It reports on achievement at specific points in time and trends in achievement over time.

Congressionally authorized and funded, NAEP is uniquely positioned to serve as an independent monitor of student achievement. As the Nation’s Report Card, NAEP is uniquely obliged to maintain the public trust. This is achieved through a governance structure and assessment procedures that are transparent, involve stakeholders, and are subject to scrutiny by technical experts, policymakers, and the public. These mechanisms ensure the accuracy, timeliness, integrity and credibility of NAEP results. They provide for the validity of inferences made about the results. They keep NAEP free of ideology, inappropriate influences and special interests. They ensure the privacy and confidentiality of each individual respondent.

Each NAEP assessment is a complex project, with a five-to-six-year life cycle for new assessments. This includes about 18 months for developing a new framework, about one year for test development, one year for pilot testing, one year for the conduct of the assessment and scoring and analysis, and one year for achievement-level setting and reporting. Each step is conducted in a thoughtful, deliberate manner with input from hundreds of stakeholders and experts, requiring careful coordination among NAGB, NCES, and the many NAEP contractors and participants.

NAEP is a representative sample survey, using statistically sound means for drawing its samples. NAEP's approaches for gaining the voluntary participation of public and private schools are thoughtfully designed to accommodate the schools' needs and schedules, and customer oriented approaches for gaining the voluntary participation of public and private schools. NAEP results are presented in a manner that assures fairness in comparisons of achievement and trends over time for all subgroups reported; for geographic units, such as the nation, states, and school districts; and for public and private schools.

NAEP covers a wide range of important subjects or topics. This includes reading, mathematics, science, writing, U.S. history, civics, geography, economics, foreign language, the arts, and technology and engineering literacy. NAEP uses matrix sampling to ensure breadth and depth of subject coverage while minimizing testing time for students.

Assessments are conducted at grades 4, 8, and 12. The 4th grade was selected as the point at which the foundations for further learning are expected to be in place (e.g., when “learning to read” becomes “reading to learn”). The 8th grade was selected because it is the typical transition point to high school. The 12th grade was selected because it is the end of the K-12 experience, the transition point for most students to postsecondary education, training, the military, the workforce, and other adult pursuits. NAEP is unique as the only source of 12th grade results at the national and state levels. Assessments are also administered at ages 9, 13, and 17, in connection with the reading and mathematics assessments conducted at NAEP’s beginning-founding (referred to as the long-term trend assessments), and when appropriate for comparisons with international assessments.
NAEP reports results by gender, race/ethnicity, and income level, socioeconomic status, and for students with disabilities and for students who are English language learners. NAEP was a pioneer in reporting data on education achievement disaggregated by student demographic subgroups. The Nation’s Report Card brings public attention to gaps in achievement between subgroups, where they exist, and to trends over time in the size of these gaps.

**Limitations: What the National Assessment of Educational Progress Is Not**

NAEP only provides group results; it does not produce results for individual students. Although NAEP collects information on student demographics and other characteristics, it does not collect information that is intrusive to individual students or families nor does it collect personally identifiable information on any respondent.

To enrich the reporting of NAEP results for each assessment, a limited amount of background information is collected from students, teachers, and administrators about the context of teaching and learning. NAEP’s background information includes factors that may affect student achievement, such as educational policy, instructional activities and teacher preparation. Reporting on these factors can help stimulate policy discussions by national, state, and local leaders, as well as the formation of hypotheses for further research. However, data from the background information NAEP collects do not, by themselves, support conclusions about the effectiveness of these factors on student outcomes nor about ways to improve education practice because those data are correlational and cannot establish causation.

Each NAEP assessment is developed through a national consensus process. This process takes into account education practices, the results of education research, and changes in curricula. However, NAEP is independent of any particular curriculum and does not promote specific ideas, ideologies, or teaching techniques.

**The Audiences for the National Assessment of Educational Progress**

The primary audience for NAEP results is the American public and their congressional representatives, including especially those in states and districts that receive their own NAEP results. With this audience as the target, NAEP reports are written to be understandable, free of jargon, easy to use, and easy to access. Assessment questions and samples of student work are included in NAEP reports when appropriate to illustrate the meaning of NAEP achievement levels and NAEP scores. Although written for a lay audience, NAEP reports do not trade accuracy for simplicity.

Another audience is made up of those who use—employ, utilize—NAEP resources—the national and state policymakers, and educators, business community and parent leaders concerned with student achievement, curricula, testing, and standards. NAEP data and related information (e.g., assessment frameworks and items) are available online to all users in forms designed to support their efforts to interpret NAEP results, to improve education performance, and to perform secondary analyses.

To be relevant and useful to these audiences, NAEP results must be timely. Therefore, NAGB has set the goal of releasing initial NAEP student achievement results within six months of testing the completion of data collection for each assessment.
Objective Goals and Activities for Conducting and Reporting the National Assessment of Educational Progress

The National Assessment Governing Board hereby sets forth six goals for conducting and reporting the National Assessment of Educational Progress. The six goals and associated activities are described below.

Objective Goal 1: To serve as a consistent external, independent measure of student achievement by which results across education systems can be compared at points in time and over time.

National, state, and local education leaders and policymakers—public and private—rely on NAEP data as an independent monitor of student achievement and as a way to compare performance across education systems. For NAEP to serve in this role, NAGB, in consultation with NCES and stakeholders, periodically establishes a dependable, publicly announced assessment schedule of at least ten years in scope. The schedule specifies the subject or topic (e.g., High School Transcript Study), grades, ages, assessment year, and sampling levels (e.g., national, state) for each assessment.

The NAEP schedule of assessments is the foundation for states’ planning for participation in the assessments. It is the basis for NCES operational planning, annual budget requests, and contract statements of work. In making decisions about the NAEP schedule of assessments, NAGB includes the wide range of important subjects and topics to which students are exposed. NAGB also considers opportunities to conduct studies linking NAEP with international assessments.

As the NAEP authorizing legislation provides, assessments are conducted in reading and mathematics, and, as time and resources allow, in subjects such as science, writing, history, civics, geography, the arts, foreign language, economics, technology and engineering literacy and other areas, as determined by NAGB. The goal for the frequency of each subject area assessment is at least twice in ten years, to provide for reporting achievement trends.

In order to compare results across geographic jurisdictions, the samples drawn must be representative. For each assessment, the National Assessment program takes affirmative steps to achieve statistically sound levels of school and student participation and optimal levels of student engagement in the assessment, including steps to maximize the participation of students with disabilities and students who are English language learners.

NCES employs safeguards to protect the integrity of the National Assessment program, prevent misuse of data, and ensure the privacy of individual test takers. NAEP results are accompanied by clear statements about school and student participation rates; student engagement in the assessment, when feasible; and cautions, where appropriate, about interpreting achievement results.
OBJECTIVE 2: To develop technically sound, relevant assessments designed to measure what students know and can do.

NAEP assessment frameworks spell out how each subject area assessment will be put together. The frameworks are the foundation for what NAEP will assess and report. Assessment frameworks describe the knowledge and skills most important for NAEP to assess at each grade. They provide for the item types and appropriate mix that best represent such knowledge and skills (e.g., multiple-choice, constructed response, hands-on task, information and communication technology-based task or simulation, etc.). Test specifications provide detailed instructions to the test writers about the specific content to be tested at each grade, the item type for each test question, and how items will be scored.

The National Assessment Governing Board is responsible for developing assessment frameworks and specifications for NAEP. NAGB does this through a comprehensive, broadly inclusive process lasting about 18 months. It involves hundreds of teachers, curriculum experts, state and local testing officials, administrators, policymakers, practitioners in the content area (e.g., chemists for science, demographers for geography, etc.) and members of the public.

The framework development process helps determine what is important for NAEP to assess and how it should be measured. The frameworks also include preliminary achievement level descriptions (see Objective 3). The framework development process considers both current classroom teaching practices and important advances in each subject area. Where applicable, the curricula, performance standards, and student achievement in other nations are also considered.

NCES is responsible for developing items for each assessment that comprehensively measure the subject domain as defined by the assessment framework and specifications. NAGB is responsible for approving all items, including those for background information, before use in an assessment.

NCES regularly evaluates the extent to which the set of items for each assessment meets the framework requirements, assessment specifications, and achievement level descriptions.

To ensure that NAEP data fairly represent what students know and can do, the frameworks and specifications are subjected to wide public review before adoption, and the items developed are reviewed for relevance and quality by representatives from participating states.

For NAEP to measure trends in achievement accurately, the frameworks (and hence the assessments) must remain sufficiently stable. However, as new knowledge is gained in subject areas, the information and communication technology for testing advances, and curricula and teaching practices evolve, it is appropriate for NAGB to consider changing the assessment frameworks and items to ensure that they support valid inferences about student achievement. But if frameworks, specifications, and items change too abruptly or frequently, the ability to continue trend lines may be lost prematurely, costs go up, and reporting time may increase. For these reasons, NAGB generally maintains the stability of NAEP assessment frameworks and specifications for at least ten years. NCES assures that the pool of items developed for each subject provides a stable measure of achievement for at least the same ten year period. In
deciding to develop new assessment frameworks and specifications, or to make major alterations to approved frameworks and specifications, NAGB considers the impact on reporting trends.

Whenever feasible, technically defensible steps are taken to avoid breaking trend lines. In rare circumstances, such as where significant changes in curricula have occurred, NAGB may consider making changes to assessment frameworks and specifications before ten years have elapsed.

In developing new assessment frameworks and specifications, or in making major alterations to approved frameworks and specifications, NAGB, in consultation with NCES, estimates the cost of the resulting assessment. NAGB considers the effect of that cost on the overall priorities for the NAEP schedule of assessments.

**OBJECTIVE**

3. To **continue to** set and report achievement levels for NAEP results.

In the 1988 re-authorization of NAEP, Congress made three major innovations. It provided for the first ever state-level assessments, created NAGB to oversee and set policy for NAEP, and authorized NAGB to set explicit performance standards, called achievement levels, for reporting NAEP results.

Previously, NAEP reporting focused primarily on average scores and whether they had changed since prior assessments. The average mathematics score of 4th graders may have gone up (or down) four points on a five-hundred-point scale. But there was no way of knowing whether the current and previous scores represented strong or weak performance and whether the amount of change should give cause for concern or celebration.

There had been attempts to give meaning to the NAEP scales through what were referred to as “performance levels.” Starting at 250—the midpoint of the 0-500 scale—points were selected for reporting at 50-point intervals above and below. The cluster of skills that differentiated each major level were identified by the items that students were more likely to answer correctly at one level than students at lower levels. Descriptions of what students know and can do at each performance level were developed from the content of the respective item clusters. However, the performance levels still did not address whether achievement was “good enough.”

NAGB approved the first policy statement on the use of achievement levels in May 1990. The policy called for the NAEP achievement levels to be denoted as “Basic,” “Proficient,” and “Advanced.” Proficient, the central level, represents “competency over challenging subject matter,” as demonstrated by how well students perform on NAEP. Basic denotes partial mastery and Advanced signifies superior performance on NAEP. Using achievement levels to report results and track change over time adds meaning to the score scale. Reporting by achievement levels helps readers judge whether performance is adequate and progress over time sufficient.

The NAEP achievement levels are developed through a thorough procedure with comprehensive technical documentation, involving expert judgment. For each achievement level-setting project, an explicit design document is developed. The design document describes the qualifications for the
individuals who will serve on the achievement level-setting panels and the specific process that will be conducted, including evaluation procedures and validity research. The panels’ recommendations are subject to technical and public comment. Ultimately, while considering the panels’ recommendations, the achievement levels are set by NAGB.

NAEP achievement levels are widely used by national, state, and local education leaders and policymakers. They contribute to NAEP’s role as an independent external monitor of student achievement. The achievement levels provide a common reference by which state and local performance standards and results can be compared.

The NAEP achievement levels have been the subject of several independent evaluations. NAGB uses information from these evaluations, as well as from other experts, to improve and refine the procedures by which achievement levels are set. Although NAGB’s standard-setting procedures may be among the most comprehensive and sophisticated used in education, NAGB continually improves the achievement level-setting and reporting process.

NAGB conducts continuing research to support the validity of inferences made in relation to NAEP achievement levels. Where the research indicates that there are limitations on the inferences that can be made in relation to NAEP achievement levels, these limitations are included in NAEP reports. Average scores, percentiles, and other relevant statistics are reported along with NAEP achievement levels to provide context and avoid misinterpretations.

**OBJECTIVE GOAL 4. To bring attention to achievement gaps where they exist among demographic subgroups and the urgency of closing those achievement gaps.**

Because education is the cornerstone of a nation’s strength, the existence of persistent achievement gaps between demographic subgroups in the U.S. is a threat ignored at our peril. The nation’s founding documents and Constitution provide for equal opportunity and equal justice under law for all. Supreme Court decisions and federal legislation undergird these civil protections against discrimination, especially in the arena of public education.

For these reasons, NAEP monitors student achievement by gender, race/ethnicity, and income level and socioeconomic status, and for students with disabilities and English language learners. In order to address achievement gaps, it is necessary first to identify them.

NAEP reports highlight achievement gaps among the student demographic subgroups so that the public is made aware and officials with responsibility have information on which to take action. The members of NAGB, individually and collectively, carry out initiatives to convey the urgency of closing achievement gaps to the public. These initiatives include preparation of special NAEP reports focused on achievement gaps, presentations, symposia, and public statements made in connection with the release of NAEP results.
OBJECTIVE GOAL 5. To disseminate timely NAEP reports and to make NAEP data and information useful and easily accessible to various audiences, including educators, policymakers, parent leaders and the public.

Given the importance of NAEP results, their timely release is critical to their impact. The goal is to release initial NAEP assessment student achievement results within six months of the completion of testing data collection for each assessment.

The information available from the National Assessment program is rich and varied. It includes:
- NAEP reports;
- assessment frameworks and specifications for the broad array of subjects included in NAEP;
- hundreds of released assessment items, including student data, exemplar student responses, and scoring guides;
- assessment results for the nation, public and private schools, states, and urban districts;
- achievement level results and descriptions; and
- background information collected from students, teachers, and school administrators.

This information is available on-line at no charge. Providing electronic versions of these materials makes them easily accessible and minimizes the need for printed copies.

NAGB and NCES continually evaluate audience needs and employ innovations in information and communication technology to improve access, usability, and usefulness of NAEP data and related resources. The aim is to optimize the potential of NAEP information to help states and others improve education achievement and close achievement gaps.

This includes procedures developed by NCES to facilitate the ability of states to link performance on NAEP with data in state longitudinal data bases. It also includes the option for states to use NAEP assessments planned for administration at the national level only. States can do this by assuming the costs and adhering to requirements that protect the integrity of the NAEP program. NAGB and NCES ensure that state decision makers receive timely notice of this option and that the cost to states is minimized.

OBJECTIVE GOAL 6. To continue to innovate in NAEP framework development, item development, test administration, data collection, test security, scoring, analysis and reporting.

Innovation is at the heart of NAEP and has been since its inception. NAEP is recognized for its advances in large-scale assessment administration, item formats, data collection, test security, scoring, analysis, quality assurance, and reporting. NAEP has embraced information and communication technology as subject matter (e.g., in the Technology and Engineering Literacy Assessment), as a tool for conducting assessments (e.g., the Writing Assessment and the Science Assessment interactive computer tasks), and as a channel to disseminate NAEP information (e.g., the on-line data tools). NAEP continually seeks innovations to improve the timeliness of NAEP reporting; enhance the precision of NAEP data; expand the ways that NAEP measures
types of students' knowledge and skills. NAEP measures in national and state sampling procedures to reduce burden on schools and students; increase the efficiency of national and state sampling, and test administration procedures; and minimize costs. Innovation is built into the NAEP modus operandi and this will continue into the future.

Prepared by:

Executive Committee Subcommittee on Updating NAEP Policy: Louis M. Fabrizio, Alan Friedman, and Shannon Garrison.

Staff: Ray Fields.

July 5, 2013
Interpreting NAEP Results Using Preparedness Research Findings

At the August 2013 meeting, the Governing Board will discuss the way forward on reporting the NAEP 12th grade results from the 2013 reading and mathematics assessments. As background for the discussion, included in this tab are:

- a draft of a prototype chapter for the report (Attachment C1; new document)
- the independent technical reviews of the preparedness validity argument by Gregory Cizek and Mark Reckase (Attachments C2 and C3; new documents)
- the preparedness validity argument (Attachment C4; included in the May 2013 COSDAM briefing materials, but changes were made to the proposed inferences as described below and indicated in highlighting on pages 8 and 41)

The draft prototype chapter was prepared as an example of what NAEP reporting on academic preparedness for college would look like in the report of the 2013 12th grade assessment results.

As previously reported to the Governing Board, the Board staff and NCES staff have been working collaboratively since March 2013 to develop options for reporting NAEP 12th grade results based upon the preparedness research findings. The options ranged from merely providing information about the 12th grade preparedness research and findings to reporting 12th grade results using statements (inferences) about 12th grade students’ academic preparedness.

After the May 2013 Board meeting, at which the Board reviewed the draft validity argument, the two staffs met and agreed that the next step should be to use the guidance from the Board discussion on the validity argument and prepare a prototype chapter for the report. This would provide something specific and concrete as a basis for further Board discussion.

The Board staff drew two main conclusions from the Board discussion in May about the validity argument:

- While finding the validity argument supportive, the Board wanted to consider the independent technical reviews that were to be presented at the August 2013 meeting to inform its decision making.
- The Board found the inferences that were being proposed to be “not quite there yet.”

The inference proposed in May was of the form “12th grade students scoring at or above Proficient are likely to be academically prepared….” Because “likely” was not quantitatively defined, the Board found this formulation ambiguous and potentially confusing to the public. During the discussion, Board member Andrew Ho said he was proposing a solution that he would share with staff. Mr. Ho proposed an inference of the general form as follows:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness, the percentage of students scoring at or above Proficient on Grade 12 NAEP is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities that would make them academically prepared for college.
Mr. Ho’s formulation for the preparedness inference was shared with Michael Kane, who is advising Board staff on the validity argument. Mr. Kane supported using this formulation in place of the one originally proposed and suggested adding “or reasonable” after “plausible.” Board staff revised the validity argument accordingly and it was this formulation that was considered by the independent technical reviewers of the validity argument.

**Question for Board Consideration:**

With the understanding that additional work will be required in collaboration with NCES, along with additional guidance from the Board, is the general approach exemplified in the prototype chapter (Attachment C1) an acceptable basis for moving forward with reporting on academic preparedness for college as a part of the reporting of the NAEP 12th grade reading and mathematics assessment results for 2013?
Towards NAEP as an Indicator of Academic Preparedness for College and Job Training
Ray Fields July 18, 2013

For over a decade, the National Assessment Governing Board has been conducting research to enable 12th grade NAEP to serve as an indicator of academic preparedness for college and job training. This chapter provides the rationale for pursuing this goal; the research results from studies conducted in connection with the 2009 administration of 12th grade NAEP; and the implications for NAEP 12th grade reporting.

INTRODUCTION

Indicators of many kinds are used to monitor critical aspects of national life and inform public policy. These include economic indicators (e.g., gross domestic product), health indicators (e.g., cancer rates), and demographic indicators (e.g., population trends by race/ethnicity and gender).

NAEP serves the public as a national and state indicator of education achievement at the elementary and secondary levels. NAEP monitors student achievement at key points in the elementary/secondary progression: grades 4, 8, and 12.

According to the National Assessment Governing Board, the 4th grade is the point at which the foundations for further learning are expected to be in place (e.g., when “learning to read” becomes “reading to learn”)

The 8th grade is the typical transition point to high school.

The 12th grade is the end of the K-12 education experience, the transition point for most students to postsecondary education, training, the military, and the work force. (Draft Policy Statement on NAEP).

NAEP is the only source of nationally representative 12th grade student achievement results. State tests of academic achievement are usually administered before 12th grade and are quite different across the country. Likewise, college admission tests like the ACT and SAT are generally taken before 12th grade by a self-selected sample and therefore, are not representative of all 12th graders.

Consequently, NAEP is uniquely positioned to serve as an indicator of academic preparedness for college and job training at grade 12—the point that represents the end of mandatory schooling for most students and the start of postsecondary education and training for adult pursuits.

A wide array of state and national leaders has embraced the goal that 12th grade students graduate “college and career ready.” These include the leadership and members of the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), the Business Roundtable (BRT), the U.S. Chamber of Commerce (the Chamber), a task force on education reform of the Council on Foreign Relations, and state and national political leaders. (Fields and Parsad).

NAEP and ACADEMIC PREPAREDNESS

The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would afford an invaluable public service: providing an indicator of the human capital potential of today’s and future generations of the nation’s population.
The Board began this initiative in 2004, after receiving recommendations from a distinguished blue-ribbon panel that had examined whether NAEP should continue assessing at the 12th grade.

The panel stated that “America needs to know how well prepared its high school seniors are... [only NAEP] can provide this information...and it is necessary for our nation’s well-being that it be provided.” The panel recommended that NAEP continue to assess at grade 12 and that the 12th grade assessment be transformed to measure preparedness for college, job training, and the military. (National Commission on NAEP 12th Grade Assessment and Reporting; p. 2.)

To transform 12th grade NAEP into an indicator of academic preparedness, the Governing Board took several significant steps.

1. The Board determined that measuring academic preparedness for college and job training should be an intended purpose of 12th grade NAEP.

2. The Board contracted with Achieve, Inc., in 2005 to review the NAEP 12th grade reading and mathematics assessment frameworks and identify where changes, if any, would be needed. Modest changes were recommended.

3. Accordingly, the Board made changes to the frameworks to be used for the administrations of the 12th grade assessments, scheduled for 2009 and 2013.

4. In 2006, the Governing Board assembled a team of noted psychometricians, industrial/organizational psychologists, and K-12 and postsecondary researchers to serve as a technical panel, advising on validity research to conduct.

5. In 2008, the technical panel recommended a comprehensive program of research. The validity of statements about academic preparedness in NAEP reports would be affected by the degree to which the results were mutually confirming.

Figure 1. presents a model of the research program, with five types of research displayed, the interrelationships that would be examined, and the potential meaning of the research results in terms of the NAEP score scale.

6. The Governing Board began contracting for the research studies in 2008, in connection with the 2009 administration of the 12th grade reading and mathematics assessments. More than 30 research studies were completed during the period 2009-2012.

The Research Findings
The research findings were consistent across studies and across years. For example, the content of the 12th grade NAEP reading and mathematics assessments was found to be similar to widely recognized tests used for college admission and placement (see http://www.nagb.org/what-we-do/preparedness-research/types-of-research/content-alignment.html).

Performance by the same students on NAEP and the SAT mathematics and reading tests was correlated at 0.91 and 0.74, respectively.
Statistical linking studies examining performance on NAEP and the college admission tests found that the college readiness benchmarks set for the ACT and SAT reading and mathematics were in a range around the Proficient achievement levels on the 12th grade NAEP reading and mathematics assessments. For example, the average NAEP reading score of students scoring at the SAT benchmark was 301, not significantly different from the cut-score for Proficient of 302 (see Fig. 2 and 3).

A longitudinal study followed a representative sample of Florida 12th grade NAEP test-takers into the state’s public colleges (see Fig. 2 and 3). The longitudinal study permitted an analysis of performance on NAEP and actual student outcomes. In the first year of this study, an analysis was conducted of performance on NAEP and (1) enrollment in regular versus remedial courses, and (2) first year overall college grade point average (GPA). As with the other statistical studies, the average NAEP score of the students who were not placed into remedial courses or who had a first year college GPA of B- or better was in a range around the 12th grade reading and mathematics Proficient achievement levels.

Results from the more than 30 studies were used to develop a validity argument to support proposed inferences (claims) about academic preparedness for college in relation to student performance on 12th grade NAEP. The validity argument was reviewed by two independent technical reviewers. The technical reviewers concluded that the validity argument supports the proposed inferences.

The complete research reports and the validity argument, along with the two independent technical reviews, can be found at http://www.nagb.org/what-we-do/preparedness-research.html.

Although the research results support inferences about NAEP performance and academic preparedness for college, the research results to date do not support inferences about NAEP performance and academic preparedness for job training.

A second phase of NAEP preparedness research began in 2013 and is expected to be completed in time for reporting 12th grade results in 2015. The second phase of research results will be examined to determine the degree to which they confirm existing results.

**A TRANSITION TO REPORTING ON ACADEMIC PREPAREDNESS**

The reporting of the 12th grade results for 2013 represents a transition point for NAEP.

The interpretations of the 2013 NAEP 12th grade reading and mathematics results related to academic preparedness for college set forth in this report are considered foundational and subject to adjustment in the future.

These interpretations are included in this report because the independent technical reviewers found them to be technically defensible, but more importantly, to promote public discussion about their meaningfulness and utility.

**The Context for Academic Preparedness for College**

In the United States in 2013, there is no single, agreed upon definition of “academic preparedness for college” used by colleges for admission and placement (Fields and Parsad). Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.
In this complex mix are 2-year institutions, 4-year public and private institutions with a wide range of selectivity, and proprietary schools. Institutions range from highly selective (i.e., with admission criteria including very high grade point averages, successful completion of rigorous high school coursework and very high SAT and/or ACT scores) to open admission (i.e., all applicants are admitted).

Even within institutions, requirements may vary across majors or programs of study. For example, the mathematics and science high school coursework and academic achievement needed for acceptance into an engineering program in a postsecondary institution may be more rigorous than the general requirements for admission to the institution or for a degree in elementary education in that institution.

**Defining Academic Preparedness for College**

Given the diversity of postsecondary education institutions, it is essential to provide a reasonable definition of academic preparedness for NAEP reporting. The definition should be relevant to NAEP’s purpose of providing group estimates of achievement. (It is important to note that NAEP does not provide individual student results.) The definition should be meaningful to NAEP’s primary audiences: the general public and national and state policymakers.

The definition proposed in this report is intended to apply to the typical degree-seeking entry-level student at the typical college. For NAEP reporting, “academically prepared for college” refers to the reading and mathematics knowledge and skills needed for placement into entry-level, credit bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

It is important to note the focus on “placement” rather than “admission.” This distinction is made because students who need remedial courses in reading, mathematics or writing may be admitted to college, but not placed into regular, credit-bearing courses. The criterion of importance is qualifying for regular credit-bearing courses, not admission.

The definition is not intended to reflect
- academic requirements for highly selective postsecondary institutions;
- the additional academic requirements for specific majors or pre-professional programs, such as mathematics, engineering, or medicine; or
- academic requirements applicable to entry into certificate or diploma programs for job training or professional development in postsecondary institutions.

The definition is focused on the first year of college; it does not address college persistence beyond the first year or completion of a degree. The definition will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.

**Proposed Inferences for NAEP Reporting**

The NAEP preparedness research does not affect the NAEP results in any way. The distribution of student achievement is unchanged. That is, the average scores, the percentiles, and the achievement level results are not impacted by the NAEP preparedness research.
The independent technical reviewers confirmed that the research findings support inferences about performance on NAEP 12th grade results in reading and mathematics in relation to academic preparedness for college.

**Proposed Inferences**

In the NAEP/SAT linking study for reading (Figure 2), the average NAEP score for 12th grade students scoring at the SAT college readiness benchmark for critical reading is 301, not significantly different from the Proficient cut-score of 302. The results from the Florida longitudinal study are confirmatory.

These data, together with the content analyses that found NAEP reading content to be similar to college admission and placement tests, support the inference for reading that

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

- **the percentage of students scoring at or above a score of 302 (Proficient) on Grade 12 NAEP in reading is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.**

In 2013, **XX%** of 12th graders nationally scored at or above 302 (Proficient) in reading.

The study results support these inferences. However, there will be students scoring at or above Proficient who are not academically prepared and students scoring below Proficient who are academically prepared (i.e., there will be false positives and false negatives). This will be true for any assessment program that sets cut-scores for a similar purpose.

In the NAEP/SAT linking study for mathematics (Figure 3), the average NAEP score for 12th grade students scoring at the SAT college readiness benchmark for mathematics is 163, lower than and significantly different from the Proficient cut-score of 176. The results from the High School Transcript Study and the Florida longitudinal study are confirmatory.

These data, together with the content analyses that found NAEP mathematics content to be similar to college admission and placement tests, support the inference for reading that

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships...
between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

In 2013, XX% of 12th graders nationally scored at or above 163 in mathematics.

To consider the plausibility of these estimates, comparisons can be made with the percentages of students who met the ACT or SAT college readiness benchmarks.

Information is available about students who were seniors in 2009 (ACT) and in 2010 (SAT). Thus, the ACT data are for the same student cohort as the NAEP data, but the SAT data are for a cohort that followed one year later.

It also must be noted that, unlike the NAEP results, neither the ACT nor the SAT results represent all 12th graders. Further, there is overlap among ACT and SAT test-takers, with about 20% estimated to take both tests.

Assuming that a substantial portion of students who do not take either test are not academically prepared for college, it is not inconsistent that the NAEP percentages are lower than those for the respective college readiness benchmarks.

| Percentages* Scoring at/above ACT and SAT College Readiness Benchmarks and at/above Proficient in Reading on NAEP and at/above 163 in Mathematics on NAEP |
|-------------------------------------------------|------------------|------------------|
|                                                 | Reading         | Mathematics      |
| ACT (2009)                                      | 53              | 42              |
| SAT (2010)                                      | 50              | 54              |
| NAEP (2009)                                     | 38              | 40              |

* About 48% of 12th graders took the ACT or SAT. NAEP represents 100% of 12th graders.

Limitations on Interpretation and Other Caveats

False Negatives and False Positives
Some proportion of 12th grade students scoring below Proficient on the 12th grade NAEP Reading or below a score of 163 on the Mathematics Assessment are

- likely to be academically prepared for college
- not likely to need remedial/developmental courses in reading or mathematics in college,

but with a lower probability than those at or above Proficient in reading or 163 in mathematics.

In addition, some proportion of 12th grade students scoring at or above Proficient on the 12th grade NAEP Reading or 163 on the Mathematics Assessment may not

- be academically prepared for college
- need remedial/developmental courses in reading or mathematics in college.

Not a Preparedness Standard
The proposed inferences are not intended to represent or be used as standards for minimal academic preparedness for college. The proposed inferences are intended solely to add meaning to interpretations of the 12th
grade NAEP reading and mathematics results in NAEP reports.

**GPA of B- or Better**
The variable “first-year GPA of B- or better” was selected because of its use as a research-based criterion in defining college readiness benchmarks developed for the SAT by the College Board. The College Board had agreed to partner with the Governing Board in a study linking performance on 12th grade NAEP with the SAT. Another leader in college testing programs, ACT, Inc. has developed similar benchmarks for its college admission assessments using a similar criterion and similar methodology. Because they are based on credible research related to college outcomes, and because performance on the respective tests could be linked to performance on NAEP, the college readiness benchmarks used by these testing programs were relevant, useful points of reference for the NAEP preparedness research.

The College Board has set a score of 500 on the SAT Mathematics and Critical Reading tests as its college readiness benchmarks in those areas. Based on its research, the College Board has determined that the score of 500 predicts, with a probability of .65, attainment of a first-year overall GPA of B- or higher. Similarly, the ACT college readiness benchmarks are based on research indicating a .50 probability of attaining first-year grades in relevant courses (e.g., college algebra and courses requiring college level reading) of B or better and .75 probability of C or better.

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

**Data Limitations**
The NAEP preparedness research studies are comprehensive and the results consistent and mutually confirming, but, for reading the statistical studies are limited to one year for data at the national level and to one state-based longitudinal study. For mathematics, there are two separate years of data at the national level and one state-based longitudinal study. Therefore, more evidence exists to support the plausibility of inferences related to mathematics than to reading.

**Preparedness for Job Training**
The completed research with respect to academic preparedness for job training does not support conclusions relative to the NAEP scale. Plans for future research will be reviewed by the Governing Board.

**Conclusion**
The independent technical reviewers found the Governing Board’s preparedness research to be methodical, rigorous, and comprehensive. They concluded that the research findings support the use of the proposed inferences in NAEP reports about 12th graders’ academic preparedness for college.

The interpretations of NAEP results in relation to academic preparedness for college are being reported on a preliminary basis. They are provided to help foster public
understanding and policy discussions about defining, measuring, validating and reporting on academic preparedness for college by NAEP and more broadly.

Including these inferences in NAEP 12th grade reports is intended to add meaning to the interpretation of the NAEP 12th grade results. However, the potential for misinterpretation exists. For these reasons, the section above on limitations on interpretation and other caveats is included in this chapter.

The Governing Board will monitor the use of these inferences as well as unintended consequences arising from their use as a part of the next phase of the preparedness research.

The next phase of the preparedness research is being conducted in connection with the NAEP reading and mathematics assessments administered in 2013. The research results will be used as additional validity evidence in relation to NAEP reporting on 12th grade academic preparedness.

Figure 1. Model of the Preparedness Research Program
Figure 2.

NAEP 12th-Grade Preparedness Research: Reading

Average Scores and Inter-quartile Ranges for Selected Variables, SAT and ACT College Readiness Benchmarks From the 2009 NAEP/SAT Linking Study and 2009 Florida Longitudinal Study

Figure 3.

NAEP 12th-Grade Preparedness Research: Mathematics

Average Scores and Inter-quartile Ranges for Selected Variables, SAT and ACT College Readiness Benchmarks From the 2009 NAEP/SAT Linking Study, 2005 High School Transcript Study, 2009 High School Transcript Study, and 2009 Florida Longitudinal Study
References to be added.
Review and Comment on

Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College

Prepared for:
National Assessment Governing Board

Prepared by:
Gregory J. Cizek, PhD
Professor of Educational Measurement and Evaluation
University of North Carolina at Chapel Hill
cizek@unc.edu

June 30, 2013
Review and Comment on

*Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College*

Introduction

The National Assessment Governing Board (NAGB) sought input on the constellation of logical and empirical evidence it has amassed in support of certain claims centering on how scores on the 12th Grade National Assessment of Educational Progress (NAEP) might be interpreted with respect to college preparedness. The logic underlying those claims and the logical and empirical support for the claims can be referred to as the *validity argument*.

According to Kane (2013):

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (p. 1)

The remainder of this paper presents the preparedness score interpretation claims proposed for the 12th grade NAEP scores and an overall an evaluation of the plausibility of those claims.

To produce this evaluation, I relied primarily on two documents that presented the NAEP preparedness validity argument and evidence (Fields, 2013a, 2013b). A draft response to *Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College* (Fields,
2013a) was submitted to the National Assessment Governing Board on May 29, 2013 (Cizek, 2013). This paper is a response to a revision of Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College (Fields, 2013b)

The Proposed Interpretations and Claims

The proposed score interpretations related to college preparedness for NAEP Reading and Mathematics are the following:

READING – "The percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college."

MATHEMATICS – "The percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college." (Fields, 2013b, p. 8)

The proposed interpretations are grounded in four claims (taken from Fields, 2013b):

1. The 12th grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

2. Performance on 12th grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

**Evaluation of Validity Evidence in Support of the Proposed Interpretations**

Overall, my review and analysis leads me to conclude that the logical and empirical evidence amassed provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear and coherent. The proposed interpretations are warranted in two ways: 1) by the accumulation of confirming evidence that is uniformly in the direction that would be hypothesized by the proposed interpretations; and 2) by the paucity of disconfirming evidence. On this point, it is noteworthy that the present validation effort appeared to be searching, objective, and contemplated the potential for disconfirming evidence.

It is my opinion, based on the evidence provided, that future NAEP reporting can provide reasonably confident and accurate indications of college preparedness in Reading and Mathematics.

It should be recognized, of course, that validation efforts typically should not be considered final or complete at any given juncture (see Cizek, 2012). Additional data can be gathered; additional experience with the test is gained; theory related to (in this case) college preparedness evolves; and new relationships among variables can be explored. The following three recommendations suggest additional validation strategies or evidential sources that may have the potential to strengthen warrants for the intended preparedness score interpretations

1) To enhance the clarity of the proposed interpretations, I offer the following recommendation: *NAGB should consider making the score interpretations parallel by specifying the NAEP scale score associated with preparedness in Reading.*
As currently worded, a defensible and specific scale score associated with preparedness is offered for NAEP Mathematics score interpretations; however, the interpretation for Reading is phrased as an achievement level: “The percentage of students in the 12th grade NAEP distribution at or above (Proficient for reading and a score of 163 for mathematics) is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in (reading or mathematics) that would make them academically prepared for college.”

The lack of parallelism in construction seems awkward, unnecessary, and potentially confusing to readers and users of this information. I recommend expressing both the Reading and Mathematics interpretations as NAEP scale scores, with elaboration as achievement levels if desired. An example of a slightly reworded interpretation along these lines would be:

“The percentage of students in the 12th grade NAEP distribution at or above a scaled score of XXX (Proficient) in Reading and a score of 163 in Mathematics is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in those subjects that would make them academically prepared for college.”

2) To enhance the coherence of the proposed interpretations, I offer the following recommendation: *NAGB should consider conducting additional research into the content coverage of the NAEP and the alignment of NAEP with traditional college admissions measures.*
In its present form, it is argued that, in essence, the content of NAEP assessments in Reading and Mathematics covers everything that traditional college admissions measures (e.g., ACT, SAT, etc.) do, but also more. It is claimed that NAEP content coverage is "broader." The Venn diagram below illustrates this claim:

![Venn Diagram](image)

**Figure 1**
*Hypothetical content coverage between NAEP Assessment and College Admissions Assessment*

Figure 1 illustrates (ignoring the relative size of the circles) the claim that NAEP is somewhat of an umbrella assessment in terms of content coverage compared to the traditional college admissions measures on which alignment research has already been conducted. However, it is not clear that the fact that an umbrella relationship exists unequivocally supports the claim that NAEP assessments capture the same things about college preparedness as the college admissions tests or, importantly, that conclusions based on such alignment can unambiguously be made with respect to preparedness. For example, it would be theoretically possible for an examinee could score "Proficient" on

---

1 The Venn diagram and the reference to the SAT are presented only illustrate content relationships between content coverage on assessments. The diagram is not intended to represent the actual proportional content coverage between NAEP and college admissions assessments, nor that of the SAT in particular.
NAEP Reading (and be deemed prepared for college) by getting very little of the "SAT-like" content correct on NAEP (that is, content deemed necessary for college success) and getting a lot of the “other” NAEP content correct (that is, the additional/broader content that may or may not necessarily be relevant to college preparedness).

3) To enhance the comprehensiveness of the proposed interpretations, I offer the following recommendation: NAGB should consider conducting additional research into the predictive validity of the NAEP with respect to college success.

Perhaps the most important variable assessed in the validation of traditional college admissions assessments is the ultimate criterion of college success—typically operationalized as first year GPA, persistence, or some other variable. Although the validity evidence gathered so far links NAEP scores to scores on other measures that are, in turn, linked to college success, the present validity case for NAEP preparedness does not do so directly. For the future, independent evaluations of direct evidence regarding the extent to which NAEP preparedness scores are associated with college criterion outcomes would substantially bolster the evidence in support of the intended score interpretations.

**Conclusion**

The logical and empirical evidence gathered to date provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear, coherent, and comprehensive. Recommendations were presented for future strategies to strengthen the validity
case. Nonetheless, based on the empirical evidence and logical rationales to date, there appear to be strong warrants for the intended interpretations regarding NAEP reporting and indications of college preparedness in Reading and Mathematics.

References


Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use, Psychological Methods, 17(1), 31-43.


Comments on the “Draft Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College”
Dated July 7, 2013

Mark D. Reckase
Michigan State University
July 10, 2013

Beginning in March, 2004, the National Assessment Governing Board (NAGB) began work to support the use of the 12th grade National Assessment of Educational Progress (NAEP) as a measure of “preparedness” of students for academic work at the college level. There are many challenges to this work, but one of the most important is to show that there is validity evidence to support the inference that students who are estimated to be above a specified level on the NAEP reporting score scale have the skills and knowledge to profit from credit-bearing, first-year college level coursework.

During the nine year period of this effort, the thinking about the way that validity evidence is collected and reported has had some significant changes. Particularly over the last few years, the work of Michael Kane (e.g., Kane, 2013) has provided guidance about how to present validity evidence for the interpretation of the results of an academic test in the form of what is now called a “validity argument.” The document that I reviewed was one of the first that I have seen that takes this approach to heart and makes a highly credible effort to apply this perspective on validation. In one sense, this is not surprising because work on NAEP has tended to be at the forefront of innovative psychometrics, be it on the use of item response theory procedures or standard setting. In another sense, it is surprising that NAGB has adopted this approach because there are few practical models for the creation of a validity argument. Even though there may have been some risk in being among the first to report support for an inference using the validity argument, this document is quite successful at providing a well supported validity argument. It gives other testing programs a very nice model for future reports on the validation of inferences from test scores.

My general view is that this document presents solid support for the inference that the proportion of the examinee population that is estimated to be above the specified cut score on the NAEP reporting score scale meets the definition of “preparedness for credit-bearing, first-year college coursework.” The evidence that was collected to support the inference is quite extensive and the connection of the evidence to the argument is logical and compelling. There are also appropriate cautions about over interpretation of results. It is very nice to see the areas of weakness in the supporting documents as well as the strengths. This adds credibility to the conclusions from the validity argument. This is not to say that the argument could not be tightened and elaborated, but this is an impressive example of a validity argument for a complex inference from a complex assessment.
A More Detailed Analysis

Although I have a very positive reaction to the report, it is important to probe the specifics of the argument and the claims being made. This may be interpreted as a desire for even more detail than is given in the report, but there is always a need for balance between detail and clear communication. The report is already long and detailed. I am reluctant to suggest adding more to it. But I do want to highlight some specific issues about some of the assumptions and claims in the argument.

The following statement is the basic inference that is the focus of the argument.

“The percentage of students in the NAEP distribution at or above a particular score level in reading or mathematics on 12th grade NAEP is a plausible, or reasonable, estimate of the percentage of 12th grade students who are academically prepared for college.” (P. 6)

This statement is very rich in meaning. To fully understand it, some background information is assumed to be known by the reader. Some of this background is listed here, but the list may not be comprehensive.

1. NAEP produces an accurate representation of the distribution of achievement of students in the areas of reading and mathematics.
2. The estimate of the proportion of students above a cut score on the NAEP reporting score scale is fairly accurate.
3. Students who are estimated to be above the specified cut score are likely to have high school grades and college admissions test scores that will make them eligible for admission to college.
4. Those students who are eligible for admission attend college and enroll in entry-level, credit-bearing courses.
5. The skills and knowledge in reading and mathematics are prerequisite to learning the content presented in the entry-level, credit-bearing courses.

The first two entries in the list are well supported by the technical documentation for NAEP. There are many years of research studies and analyses that show the technical quality of the assessment program. The last three of the entries in the list are more difficult to support because NAEP does not provide accurate student level scores and the individual students who participate are usually not identified so their academic history following the NAEP administration cannot be recorded. It is here that the special studies and data collections that have been done by NAGB are important to fill in links of the validity argument.

A Slight Variation on the Validity Argument

During the process of reviewing the report on the validity argument, I took notes on component parts of the argument. In some cases, the purpose of the notes was to highlight assumptions that were not explicitly stated. In other cases, the purpose was to elaborate on a step in the validity argument. A summary of these notes in the form of a slightly different validity argument than the one given in the report is given below. This is not meant to imply a problem
with the validity argument in the NAGB report, but rather to add some commentary on that argument.

1. There is a body of knowledge and skills that is taught at the secondary school level that is prerequisite to gaining admission into entry-level, credit-bearing courses at colleges and universities.
   a. There seems to be strong evidence for this from the America Diploma Project and the analysis of the admission and placement tests.
   b. It might be helpful to think of this in terms of a Venn diagram that shows the intersection and union of the content descriptions from all of these different sources. The argument should be made that NAEP is based on a reasonable sampling of content from the intersection or the union.

2. College admissions test scores and high school transcripts provide information about the prerequisite knowledge and skills and these are used to make decisions about admissions to the entry-level courses.
   a. This is easy to document, but it is not explicitly stated it in the argument. Of course, different institutions use the information in different ways.

3. The knowledge and skills reflected in college admissions tests and high school transcripts that are prerequisite to the entry-level college courses can be described in some detail to allow the design of a test to assess the knowledge and skills.
   a. This is clearly supported by the information from the studies.
   b. It would be useful to have a summary description of the common components from all of the parts.

4. NAEP assessments provide information about student acquisition of the knowledge and skills described above.
   a. This is the main thrust of all of the content analysis.
   b. The argument is compelling, but it would be helpful to have a general content description that is the result of all of the content analysis.

5. There is a threshold value for the knowledge and skills defined above. If students do not meet this threshold, they will not be ready to take the entry level courses.
   a. The comparative data make a good argument for the existence of the cut score.

6. A cut score on NAEP is consistent with the threshold.
   a. There is a good process for identifying a reasonable cut score on NAEP to correspond to #5.
   b. The combination of information from different tests results in strong support for parts of the argument.
7. The proportion of students estimated to be above the cut score on NAEP gives a good estimate of the proportion who exceed the threshold for admission into entry level courses.
   a. This is well supported by the statistical analysis procedures if the argument for an appropriate cut score is supported. In this case, there is reasonable support for the cut score from the connection to placement and admissions tests.

From this argument, I believe that the following inference from NAEP reported results is supported: The proportion of students estimated to be above the specified cut score on the NAEP reporting score scale is a reasonable estimate of the proportion of students who have the prerequisite knowledge and skills in mathematics and reading to profit from entry-level, credit-bearing college courses.

Reference

Introduction

Rationale for NAEP Reporting on 12th Grade Academic Preparedness
The National Assessment Governing Board is conducting a program of research to determine the feasibility of the National Assessment of Educational Progress (NAEP) reporting on the academic preparedness of U.S. 12th grade students, in reading and mathematics, for college and job training.

Since 1969, NAEP has reported to the public on the status and progress of student achievement in a wide range of key subjects at grades 4, 8, and 12. NAEP provides national and state-representative results, results for twenty-one urban districts, and results by subgroups of students (e.g., by race/ethnicity, gender, and for students with disabilities and English language learners). NAEP, by law, does not provide individual student results.

The Governing Board’s initiative on 12th grade academic preparedness began in March 2004, with the report of a blue-ribbon panel. The panel was composed of K-12 education leaders—the “producers” of high school graduates—and leaders in business, postsecondary education, and the military—the “consumers” of high school graduates.

The panel members recognized the importance of 12th grade as the gateway to postsecondary education and training, and viewed NAEP as a “truth teller” about student achievement. These distinguished state and national leaders recommended unanimously that “NAEP should report 12th grade students’ readiness for college-credit coursework, training for employment, and entrance into the military.” (National Commission on NAEP 12th Grade Assessment and Reporting; p. 6.). They stated that “America needs to know how well prepared its high school seniors are…[only NAEP] can provide this information…and it is necessary for our nation’s well-being that it be provided.” (Ibid. p. 2.).

The Governing Board approved this recommendation, with a minor modification. The term “readiness” was changed to “academic preparedness” and “entrance into the military” was subsumed by “job training.”

“Readiness” was changed to “academic preparedness” because “readiness” is broadly understood to include both academic preparedness and other characteristics needed for success in postsecondary education and training, such as habits of mind, time management, and persistence (Conley). NAEP does not purport to measure such characteristics. Rather, NAEP is designed to measure academic knowledge and skills.

1 The blue-ribbon panel was known officially as the National Commission on NAEP 12th Grade Assessment and Reporting.
“Entrance into the military” was subsumed by “job training” with the intention of identifying occupations with civilian and military counterparts and utilizing the military’s experience as the world’s largest occupational training organization and its extensive research on the relationship between performance on the Armed Service Vocational Aptitude Battery (ASVAB) and job training outcomes.

The Governing Board approved the 12th grade academic preparedness initiative because it believes that the academic preparation of high school students for postsecondary education and training is important to the nation’s economic well-being, national security, and democratic foundations (see Governing Board resolution of May 21, 2005 at http://www.nagb.org/content/nagb/assets/documents/policies/resolution-on-preparedness.pdf).

Indicators of many kinds are used to monitor critical aspects of national life and inform public policy. These include economic indicators (e.g., gross domestic product), health indicators (e.g., cancer rates), and demographic indicators (e.g., population trends by race/ethnicity and gender). The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would serve as a valuable indicator of the human capital potential of rising generations of citizens, a nation’s greatest resource.

The Governing Board is not alone in recognizing the importance of 12th grade academic preparedness for the nation. A wide array of state and national leaders has embraced the goal that 12th grade students graduate “college and career ready.” These include the leadership and members of the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), the Business Roundtable (BRT), the U.S. Chamber of Commerce (the Chamber), the Council on Foreign Relations, and the Obama Administration. The reason for this attention to 12th grade academic preparedness is well summarized by a statement of the Business Coalition for Student Achievement, an organization coordinated by BRT and the Chamber:

"Ensuring that all students graduate academically prepared for college, citizenship and the 21st century workplace…is necessary to provide a strong foundation for both U.S. competitiveness and for individuals to succeed in our rapidly changing world."

The NGA and CCSSO have collaborated to develop Common Core State Standards (CCSS) for mathematics and English language arts. These standards are aimed at fostering college and career readiness by the end of high school. The CCSS have been adopted formally by 45 states, several territories and the Department of Defense Education Activity. Viewing the need for rigor in education standards and outcomes through the lens of national security, a similar conclusion was made in the report of the Independent Task Force on U.S. Education Reform and National Security of the Council on Foreign Relations. The Task Force was co-chaired by former New York City School Chancellor Joel Klein and Former Secretary of State Condoleezza Rice. The Obama administration has stated that “educating every American student to graduate from high school prepared for college and for a career is a national imperative.” (Fields and Parsad; pp. 3–4).

Twelfth grade is the end of mandatory schooling for most students and represents the transition point to adult postsecondary pursuits. If it is essential for students to graduate from high school
academically prepared for college and job training, it is essential for the public and policymakers to know the degree to which this is occurring.

A trusted indicator is needed for reporting to the public and policymakers on the status of 12th grade academic preparedness in the U.S., but no such indicator exists. State tests at the high school level are typically administered at 10th and 11th grade. College admission tests, like the SAT and ACT, are administered before the 12th grade, generally to self-selected samples of students.

State tests and college admission tests do not provide a measure of what students know and can do at the very end of K-12 education. Even if these state tests and college admission tests were administered at the 12th grade, they could not be combined to produce nationally representative results.

NAEP is the only source of national and state-representative student achievement data at the 12th grade. As such, NAEP is uniquely positioned to serve as an indicator of 12th grade academic preparedness.

**Defining Academic Preparedness for College**

In the United States in 2013, there is no single, agreed upon definition of “academic preparedness for college” used by colleges for admission and placement. Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.

In this complex mix are 2-year institutions, 4-year public and private institutions with a wide range of selectivity, and proprietary schools. Institutions range from highly selective (i.e., with admission criteria including very high grade point averages, successful completion of rigorous high school coursework and very high SAT and/or ACT scores) to open admission (i.e., all applicants are admitted).

Even within institutions, requirements may vary across majors or programs of study. For example, the mathematics and science high school coursework and academic achievement needed for acceptance into an engineering program in a postsecondary institution may be more rigorous than the general requirements for admission to the institution or for a degree in elementary education in the institution.

In order to design the NAEP 12th grade preparedness research, a working definition of preparedness was needed. The Governing Board’s Technical Panel on 12th Grade Preparedness Research recommended use of the following working definition, which defines academic preparedness for college as

… the academic knowledge and skill levels in reading and mathematics necessary to be qualified for placement…into a credit-bearing entry-level general education course that fulfills requirements toward a two-year transfer degree or four-year undergraduate degree
This definition was intended to apply to the “typical” college, not to highly selective institutions, and thus, to the vast majority of prospective students, or about 80% of the college freshmen who enrolled in 2-year and 4-year institutions within 2 years following high school graduation (Ross, Kena, Rathbun, KewalRamani, Zhang, Kristapovich, and Manning, p 175). To make this clear, the definition is further elaborated as follows.

Academic preparedness for college refers to the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements (ECNRG) in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

This is consistent with the approach used by the College Board and ACT, Inc. in developing their respective college readiness benchmarks, which are used as external referents in the NAEP 12th grade preparedness research. The ACT benchmarks “represent predictive indicators of success for typical students at typical colleges (Allen and Scoring).” The SAT benchmarks are “an indication of college readiness at a typical college (College Board).”

Domain Definition for Academic Preparedness for College in Reading and Mathematics

The working definition described above set the stage for designing the preparedness research studies, but begged a basic question—What are the reading and mathematics knowledge and skills needed to qualify for placement into ECNRG and are they measured by NAEP? This question would be addressed by examining the degree of content match between NAEP and multiple widely accepted external sources that had developed domain definitions for academic preparedness for college in mathematics and reading.

A perfect match between two different sources could not be expected, but a sufficient content match between NAEP and each of a multiple of relevant widely accepted external sources would, collectively, support the inference that the needed knowledge and skills are measured by NAEP. Consequently, the Governing Board identified the following external sources for content comparison with NAEP: The American Diploma Project (ADP) benchmarks for mathematics and English, the ACT College Readiness Standards for Mathematics and Reading, and the ACT, SAT, and ACCUPLACER assessments for reading and mathematics. The results of the content comparison studies between NAEP and these other sources are described in the validity argument below.

The Central Issue: Validity

Having made the decision to determine the feasibility of NAEP reporting on 12th grade academic preparedness, the Governing Board recognized that the central concern would be establishing the validity of inferences about 12th grade academic preparedness that are to be made from NAEP scores and used in NAEP reports. The Governing Board would need to ensure that the content of NAEP 12th grade reading and mathematics assessments was appropriate for measuring academic preparedness and that research was conducted to collect evidence by which the validity of
proposed inferences could be evaluated. Finally, a formal validity argument would need to be
developed, specifying the proposed inference(s) for NAEP reporting, the underlying assumptions
or propositions, and the evidence related to the assumptions or propositions.

Accordingly, the Governing Board

- revised the NAEP assessment frameworks for the 2009 12th grade reading and
  mathematics with the explicit purpose of measuring academic preparedness for college
  and job training,
- appointed a special panel of technical experts to recommend a program of research on
  12th grade academic preparedness (National Assessment Governing Board, 2009),
- approved and conducted a comprehensive set of preparedness research studies, and
- adopted the model for a validity argument described by Michael Kane (Kane).

The first phase of the Governing Board’s program of preparedness research is completed. The
studies were conducted in connection with the 2009 NAEP 12th grade assessments in reading and
mathematics. More than 30 studies of five distinct types have been conducted. Study results are
available and the complete studies are posted at http://www.nagb.org/what-we-do/preparedness-
research.html. The National Center for Education Statistics (NCES) has provide additional data
drawn from analyses of the 2005 and 2009 High School Transcript Studies conducted in
connection with the NAEP 12th grade assessments in those years.

From this research, Governing Board staff developed a proposed interpretation of NAEP
performance in reading and mathematics related to 12th grade academic preparedness for college.
Following below is the validity evidence for the proposed interpretation, presented in the form of
a validity argument. The validity argument provides a statement of the proposed interpretation
and the main assumptions inherent in the proposed interpretation in terms of academic
preparedness for college. These assumptions are then evaluated using several lines of evidence,
which were found to converge for both reading and for mathematics.

Validity Argument

Overview
The National Assessment of Educational Progress (NAEP) program is designed to provide
information about student achievement in reading, mathematics and other content areas at the 4th,
8th, and 12th grades. The items for the assessments are developed according to content
frameworks and test specifications developed by the National Assessment Governing Board.
Scientific sampling procedures are used to produce estimates of score distributions representative
of the national population of students at each grade level, as well as estimates representative of
public school students in individual states and in 21 urban school districts. The NAEP results do
not produce scores for individual students, but rather, group estimates. The NAEP results are
reported, based on the estimated score distributions, by average score, percentiles, and in terms
of the percentages of students at or above three performance standards used for NAEP reporting,
called achievement levels, that are designated Basic, Proficient, and Advanced.
The purpose of the research reported here was to examine whether the interpretation of 12th grade NAEP results in reading and mathematics could be extended to include statements about the percentage of U.S. 12th graders who are academically prepared for college and, if such an interpretation were found to be defensible, to determine the specific statements about academic preparedness that were supportable by the research evidence. The specific statements would be based on the following general definition for academic preparedness, used in relation to the NAEP preparedness research:

the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

The NAEP assessment program is well-established and regularly evaluated, with ample technical documentation of the interpretation of the results at all three grade levels. Therefore, the technical quality, accuracy, and representativeness of the NAEP results in terms of the estimated distributions of U.S. 12th graders on the NAEP scales in reading and mathematics will be taken as a given and as a starting point for additional inferences about the academic preparedness of U.S. 12th graders for college.

In particular, the intent of this validity argument is to examine the evidence in support of statements related to academic preparedness for college for use in reporting NAEP 12th grade results that would have the following general form:

The percentage of students in the NAEP distribution at or above a particular score level in reading or mathematics on 12th grade NAEP is a plausible, or reasonable, estimate of the percentage of 12th grade students who are academically prepared for college.

This interpretation would depend on four prior claims (or assumptions):

1. The 12th grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.
2. Performance on 12th grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.
3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12\textsuperscript{th} grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

The first claim is supported by the combination of the content of the NAEP assessment frameworks and the NAEP test items, the NAEP sampling designs, and the statistical models used to generate estimates of score distributions at each grade level and in each content area. These claims are well-established, documented, and evaluated; therefore, the attention of the validity argument will be directed primarily to the second, third, and fourth claims.

The second claim is supported by a statistical relationship study that examined student performance on the NAEP 12\textsuperscript{th} grade reading and mathematics assessments to performance on the SAT reading and mathematics tests, as well as the respective college readiness benchmarks established by the College Board for these tests, which, in turn, are related to outcomes associated with academic preparedness for college.

The third claim was evaluated with multiple sources of evidence that were highly convergent. These include the SAT/NAEP statistical relationship study, a longitudinal study of Florida 12\textsuperscript{th} grade students, and analyses of the 2005 and 2009 NAEP High School Transcript Studies.

The fourth claim is supported by the fact that the Governing Board reviewed the NAEP 12\textsuperscript{th} grade reading and mathematics frameworks for the purpose of making NAEP a measure of academic preparedness for college; made changes to the frameworks accordingly; and conducted a comprehensive set of content alignment studies to determine the degree of match between NAEP and tests that are used for college admission and placement.

Further, the results from the examination of the NAEP content provide a counter argument to a possible falsifying claim about the positive relationships discussed in the second and third claims. The falsifying claim would be that the positive relationships between NAEP and the other indicators were merely statistical artifacts, due to factors extraneous to academic preparedness for college, akin to finding a high correlation between height and passing rates on a state driving test. The counter argument is that the relationships are meaningful because the NAEP 12\textsuperscript{th} grade reading and mathematics assessments were intentionally designed to measure academic preparedness for college and that the evidence supports the conclusion that the NAEP 12\textsuperscript{th} grade assessments do measure academic preparedness for college.
**Proposed Inferences**

**For reading:**

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

the percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.

**For mathematics:**

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

In contrast to the inference for reading, which is set at the Proficient level, the inference for mathematics is set at a score on the NAEP mathematics scale of 163. This score is strongly supported by the consistent research results across years and data sources, but is below and significantly different from the cut-score for the Proficient level for NAEP 12th grade mathematics, which is 176.

The research results for mathematics do support a related inference—that students in the distribution at or above the NAEP Proficient level in mathematics are likely to be academically prepared for college. However, the percentage of such students would be substantially less than the percentage in the distribution at or above 163, and thus, would underestimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.

For these reasons, and to have the proposed inferences for reading and mathematics as parallel as possible, the proposed inference for reading is formulated in relation to the Proficient achievement level and the proposed inference for mathematics is formulated in relation to the NAEP mathematics scale score of 163.

**Limitations on Interpretation and Other Caveats**

**False Negatives and False Positives**

Some proportion of 12th grade students scoring below Proficient on the 12th grade NAEP Reading or below a score of 163 on the Mathematics Assessment are
likely to be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and

not likely to need remedial/developmental courses in reading or mathematics in college, but with a lower probability than those at or above Proficient in reading or 163 in mathematics.

In addition, some proportion of 12th grade students scoring at or above Proficient on the 12th grade NAEP Reading or 163 on the Mathematics Assessment may not

- be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and
- need remedial/developmental courses in reading or mathematics in college.

Not a Preparedness Standard
The proposed inferences are not intended to represent or be used as standards for minimal academic preparedness for college. The proposed inferences are intended solely to add meaning to interpretations of the 12th grade NAEP reading and mathematics results in NAEP reports.

Academically Prepared for College
The proposed inferences are intended to apply to the typical degree-seeking entry-level college student at the typical college. Thus, “academically prepared for college” refers to the reading and mathematics knowledge and skills needed for placement into ECNRG courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

It is important to note the focus on “placement” rather than “admission.” This distinction is made because students who need remedial courses in reading, mathematics or writing may be admitted to college, but not placed into regular, credit-bearing courses. The criterion of importance is qualifying for regular credit-bearing courses, not admission.

The proposed inferences are not intended to reflect academic requirements for highly selective postsecondary institutions; to the additional academic requirements for specific majors or pre-professional programs, such as mathematics, engineering, or medicine; or to academic requirements applicable to entry into certificate or diploma programs for job training or professional development in postsecondary institutions.

The proposed inferences are focused on the first year of college; they do not support conclusions about college persistence beyond the first year or completion of a degree. The inferences will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.
GPA of B- or Better
The selection of “first-year GPA of B- or better” as a referent was made because of its use as a research-based criterion in defining college readiness benchmarks developed by an acknowledged leader in college testing programs—the College Board. The College Board had agreed to partner with the Governing Board in a study linking performance on 12th grade NAEP with the SAT. Another leader in college testing programs, ACT, Inc. has developed similar benchmarks for its college admission assessments using a similar criterion and similar methodology. Because they are based on credible research related to college outcomes, and because performance on the respective tests could be linked to performance on NAEP, the college readiness benchmarks used by these testing programs were embraced as relevant, useful points of reference for the NAEP preparedness research.

The College Board has set a score of 500 on the SAT Mathematics and Critical Reading tests as its college readiness benchmarks in those areas. Based on its research, the College Board has determined that the score of 500 predicts, with a probability of .65, attainment of a first-year overall GPA of B- or higher. Similarly, the ACT college readiness benchmarks are based on research indicating a .50 probability of attaining first-year grades in relevant courses (e.g., college algebra and courses requiring college level reading) of B or better and .75 probability of C or better.

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

Data Limitations
Although the preparedness research studies are comprehensive and the results consistent and mutually confirming, for reading they are limited to one year for data at the national level and to one state-based longitudinal study. For mathematics, there are two separate years of data at the national level and one state-based longitudinal study. Therefore, more evidence exists to support the plausibility of inferences related to mathematics than to reading.

Preparedness for Job Training
The completed research with respect to academic preparedness for job training does not support conclusions relative to the NAEP scale and will not be addressed at this time.
Discussion of the Claims and Evidence

1. The 12th-grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

The proposed inferences are premised in part on the capability of NAEP to report percentages of students scoring at or above a certain score on the NAEP 12th grade reading and mathematics scales. The technical qualities of the NAEP scales make them well suited to this purpose.

The NAEP sampling, scaling, IRT modeling, and statistical procedures are widely accepted, well documented (for example, see National Center for Education Statistics, pp. 70-71) and have been periodically evaluated over two decades (for example, see complete list of research conducted by the NAEP Validity Studies Panel at http://www.air.org/reports-products/index.cfm?fa=viewContent&content_id=890 and “Evaluation of the National Assessment of Educational Progress: Study Reports” at http://www2.ed.gov/rschstat/eval/other/naep/naep-complete.pdf).

Other than issues relating to the comparability among the state-level NAEP samples of inclusion rates of students with disabilities and students who are English language learners (about which the Governing Board and NAEP have taken and continue to take significant action), there is little dispute about the appropriateness of the NAEP sampling, scaling and statistical procedures for estimating the percentage of students scoring at or above a selected NAEP scale score.

This is relevant because the proposed inferences that are the subject of this validity argument are interpretations to add meaning to the reporting of NAEP 12th grade reading and mathematics results at particular score levels. The percentages of students at or above particular score levels (e.g., the NAEP achievement levels) have been estimated with accuracy and reported regularly, beginning with assessments in 1992. The proposed inference for reading would use the cut-score for 12th grade Proficient as the basis for reporting. The proposed inference for mathematics would use the score of 163 on the NAEP 12th grade scale as the basis for reporting, which is between the Basic and Proficient achievement levels. Clearly, reporting NAEP results using the proposed inferences will not impair the accuracy of the estimates of the percentages of students scoring at or above the identified points on the NAEP score scales.

2. Performance on 12th-grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

In designing the NAEP preparedness research program, the Governing Board determined that it would be essential to examine how performance on NAEP relates to performance on other
measures and outcomes associated with academic preparedness for college. The research program studied the relationship between performance on NAEP and performance on the SAT and ACT college admission tests, including the respective college readiness benchmarks that had been established by these testing programs.

The data sources for the analyses that were conducted are: the NAEP/SAT linking studies (see report at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/SAT-NAEP_Linking_Study.pdf); the Florida longitudinal study (see report at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/Florida_Statistical_Study.pdf); the 2005 and 2009 NAEP High School Transcript Studies; and the Governing Board’s survey of postsecondary education institutions’ use of tests and the cut-scores on those tests for determining whether incoming students need remedial instruction in reading and mathematics (Fields and Parsad).

In addition, the research program examined directly the relationship between performance on NAEP and postsecondary outcomes analyzing data from the Florida longitudinal study.

The results of these studies will be discussed both in this section and the next section of the validity argument. In this section, background is provided on the indicators that were examined and the results of the NAEP/SAT linking study. The NAEP/SAT linking study is discussed in this section because, as the most recent large-scale national study, it serves as a focal point for discussing the results of the other studies. Thus, in section 3, the results of the other statistical linking studies are discussed in relation to the NAEP/SAT linking study.

**Indicators: College Board and ACT College Readiness Benchmarks**

The College Board and ACT, Inc. have established college readiness benchmarks for the SAT and the ACT in a number of subjects tested, including reading and mathematics. The SAT College Readiness Benchmark for critical reading and mathematics is a score of 500 on the respective tests. According to the College Board’s research, a score of 500 predicts, with a .65 probability, a first-year GPA of B- or better. The ACT College Readiness Benchmark for reading is a score of 21. According to ACT’s research, a score of 21 predicts, with a .50 probability, a grade of B or better (or .75 probability of a C or better) in first year courses requiring college reading, such as history and the social sciences. A score of 22 on the ACT mathematics tests predicts a .50 probability of a grade of B or better in a first-year mathematics course, or a .75 probability of a grade of C or better. The College Board research and the ACT research are based on the first-year outcomes of their respective test takers.

**Indicators: First Year GPA of B- or Better and Remedial/non-Remedial Placement**

The Governing Board has a partnership with the state of Florida as a part of the Board’s program of preparedness research. Florida was one of 11 states that volunteered to provide state-
representative samples of 12th grade students for the 2009 NAEP reading and mathematics assessments. Under the partnership, the Florida 12th grade sample is being followed through the postsecondary years via the highly developed Florida longitudinal education data system. For comparability with the SAT College Readiness Benchmarks, the Governing Board analyzed the Florida data to determine the average score and interquartile range for the NAEP test takers with a first year GPA of B- or better. In addition, the Governing Board analyzed the Florida data to determine the average score and interquartile range for the NAEP test takers who were and who were not placed into remedial reading or remedial mathematics in their first year of college.

**Analysis of Results for Mathematics**

The statistical linking study examining performance on the NAEP 12th grade mathematics assessment and performance on the SAT mathematics test yielded a correlation of .91. This high correlation clearly supports inferences about NAEP performance in relation to SAT performance. The study also examined how performance on NAEP relates to the SAT College Readiness Benchmark for mathematics (i.e., a score on the SAT mathematics test of 500). The SAT benchmark provides “an indication of college readiness at a typical college (College Board).” This is consistent with the Governing Board’s definition of academic preparedness cited previously:

> Academic preparedness for college refers to the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

The SAT College Readiness Benchmark for mathematics is relevant to student outcomes in college, for it is “the SAT score associated with a 65 percent probability of earning a first-year GPA of B- (i.e., 2.67) or higher (College Board).” The average NAEP score of students scoring at the College Readiness Benchmark for mathematics was 163 (see Figure 1). As will be demonstrated in the discussion of the third claim, there are additional data corroborating this level of performance on the 12th grade NAEP mathematics assessment to outcomes in college.

**Analysis of Results for Reading**

The statistical linking study examining performance on the NAEP 12th grade reading assessment and the SAT critical reading test resulted in a correlation of .74. Although it may not be high enough to predict the performance of individual students from one test to another (which is not required to support the proposed inference for reading), it is sufficient to support the group-level inferences reported by NAEP.
Performance on NAEP was also examined in relation to the SAT College Readiness Benchmark for critical reading (i.e., a score on the SAT critical reading test of 500). The SAT benchmark provides “an indication of college readiness at a typical college (College Board).” This is consistent with the Governing Board’s definition of academic preparedness discussed in the results for mathematics above.

The SAT College Readiness Benchmark for critical reading is relevant to student outcomes in college, for it is “the SAT score associated with a 65 percent probability of earning a first-year GPA of B- (i.e., 2.67) or higher (College Board).” The average NAEP score of students scoring at the College Readiness Benchmark for reading was 301 (see Figure 2). As will be demonstrated in the discussion of the third claim, there are additional data corroborating this level of performance on the 12th grade NAEP reading assessment to outcomes in college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).

In addition to the NAEP/SAT Linking Studies (NSLS) described above, analyses were conducted using data from several other studies. There was a high degree of convergence found across the studies. The results are described below, first for mathematics and then for reading.

**Analysis of Results for Mathematics**

Companion statistical relationship studies to the NSLS for mathematics examined data from the 2005 and 2009 national NAEP High School Transcript Studies (HSTS) and from a longitudinal study under a partnership with the Florida Department of Education (FLS). In 2009, Florida was one of eleven states that volunteered to participate in 12th grade state NAEP in reading and mathematics. Using the highly developed Florida longitudinal data base, the students in the 12th grade NAEP samples were followed into postsecondary public institutions.

Analyzing data from the transcripts of NAEP test takers, the HSTS examined performance on 12th grade NAEP mathematics in relation to performance in mathematics on the SAT and ACT college admissions tests in 2005 and 2009. The FLS study examined performance on the NAEP 12th grade mathematics assessment in relation to the SAT and ACT college readiness benchmarks, first year overall college GPA, and whether students were placed into non-remedial college courses. The study results are displayed in Figure 1.

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) because it is the most recent of the national studies. The average NAEP score is 163 for students with an SAT score at the College Readiness Benchmark for mathematics of 500.
The other study results are consistently convergent with the NSLS results. The average NAEP mathematics scores for 12th grade students scoring at the SAT College Readiness Benchmark of 500 for mathematics are compared first for the 2005 HSTS and the 2009 NSLS. The average scores are 161 and 163 respectively.

As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for mathematics is defined somewhat differently than the SAT College Readiness Benchmark for mathematics. However, it is noteworthy that even with this different definition, the results from the 2005 HSTS, 2009 HSTS, and 2009 FLS analyses for the ACT (169, 166, and 164, respectively) are consistent and very similar to the results for the 2009 NSLS.

To answer the question, "What is the relationship between performance on NAEP and actual student outcomes?", we look to the FLS results. First we examine the average NAEP mathematics score for the 12th grade Florida NAEP test takers who attained a first-year GPA of B- or better. The average NAEP score for these students was 162. This is consistent with the SAT College Readiness Benchmark analyses and further supports the inference that students at
or above 163 on the 12th grade NAEP mathematics scale are likely to be academically prepared
and attain a first-year GPA of B- or better. It follows, of course, that students who are
academically prepared will not require remedial courses.

Thus, another outcome of interest is placement of entry-level students into remedial college
courses versus non-remedial credit-bearing courses. Here again, we look to the FLS as a data
source. The average NAEP mathematics score was 165 for the Florida NAEP test-takers not
placed into remedial courses, which is consistent with the NSLS score of 163 on the NAEP 12th
grade mathematics scale. Furthermore, the average NAEP score of students who were placed
into remedial mathematics courses in college was 136, much lower and significantly different
from the NSLS score of 163.

The FLS results, together with the SAT and ACT analyses, lend support to the conclusions that
students scoring at or above 163 on the 12th grade mathematics scale are likely to be
academically prepared for ECRNG college courses and not likely to need remedial courses in
mathematics.

These convergent, consistent results across years and across studies support the proposed
inference that the percentage of students scoring at or above a score of 163 on the Grade 12
NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students
who possess the knowledge, skills, and abilities in mathematics that would make them
academically prepared for college.

Analysis of Results for Reading

The companion statistical relationship study to the NSLS for reading examined data from a
longitudinal study under a partnership with the Florida Department of Education (FLS). In 2009,
Florida was one of eleven states that volunteered to participate in 12th grade state NAEP in
reading and mathematics. Using the highly developed Florida longitudinal data base, the
students in the 12th grade NAEP samples were followed into postsecondary public institutions.

The FLS study examined performance on the NAEP 12th grade reading assessment in relation to
the SAT and ACT college readiness benchmarks for reading, first year overall college GPA, and
whether students were placed into non-remedial college courses. The study results are displayed
in Figure 2.

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS)
for reading, because it is the most recent of the national studies. The average NAEP score is 301
for students with an SAT score at the College Readiness Benchmark for critical reading of 500.
A NAEP score of 301 in 12th grade reading is not significantly different from the cut-score for
the 12th grade Proficient achievement level (302).
The FLS results are consistently convergent with the NSLS results. The average NAEP reading score was 299 for the 12th grade Florida NAEP test takers who were not placed into remedial courses in their first year. The average score was 298 for those who had a first year overall GPA of a B- or better. These data, which show the relationship between performance on NAEP and actual student outcomes, provide strong confirmation that students scoring at or above Proficient on the NAEP 12th grade reading assessment are likely to be academically prepared for ECNRG college courses.

As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for reading is defined differently than the SAT College Readiness Benchmark for reading. However, it is noteworthy that even with this different definition, the ACT results from the 2009 FLS analysis are similar to the NSLS analysis and the FLS outcome data.

Taken together, these results support the inference that students scoring at or above Proficient on the NAEP 12th grade reading scale are likely to be academically prepared for ECNRG college courses.

In conclusion, these results suggest that the percentage of students at or above the Proficient level in reading on 12th grade NAEP would provide a plausible (or reasonable) estimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

- NAEP Assessment Frameworks Were Revised to Measure Academic Preparedness

The National Assessment Governing Board intentionally revised the NAEP 12th grade reading and mathematics assessment frameworks with the purpose of measuring academic preparedness for college.

On March 5, 2004, the Governing Board accepted the report of the Commission on NAEP 12th Grade Assessment and Reporting. The Commission recommended that “NAEP should report 12th grade students’ [academic preparedness] for college-credit coursework, training for employment, and entrance into the military.”

For NAEP to report on 12th grade academic preparedness for college, it must measure relevant content at the 12th grade. The content of each assessment is determined by the NAEP assessment frameworks, which the Governing Board is responsible for developing and approving. Accordingly, the Governing Board decided that the extant NAEP frameworks intended for the 2009 for reading and mathematics at the 12th grade would be reviewed. The review would identify changes needed to measure 12th grade academic preparedness for college. Examples of the changes made are described in the next two subsections.

Assessments at the 12th grade in reading and mathematics are conducted at least once every 4 years. In 2004, when the Board decided to proceed with the 12th grade academic preparedness initiative, 2009 was the next assessment year in which the 12th grade reading and mathematics assessments could be affected by framework changes.

In September 2004, the Governing Board contracted with Achieve, Inc. (Achieve) to review the NAEP 12th grade reading and mathematics assessment frameworks and identify where changes, if any, would be needed. Achieve had established the American Diploma Project (ADP) “…to improve postsecondary preparation by aligning high school standards, graduation requirements and assessment and accountability systems with the demands of college and careers (see www.achieve.org/adp-network).” The ADP had conducted research to identify key competencies in English and mathematics needed for high school graduates who aspire to higher education. They refer to these as the “ADP benchmarks.” The type of colleges that were the target for the ADP research was similar to the “typical colleges” in the Governing Board’s research. These were the “two- and four-year colleges and universities in each of the ADP partner states…[that] enroll the vast majority of high school graduates going on to college:

---

2 The review also addressed academic preparedness for job training, but that part of the NAEP preparedness initiative is not being addressed in this validity argument.
community colleges, as well as four-year state institutions, but generally not the more highly selective “flagship” campuses.” (Achieve, 2004, p. 107)

The research and expertise of the American Diploma Project was widely accepted and was brought to bear in reviewing the NAEP frameworks for 12th grade reading and mathematics. Achieve convened a panel of nationally recognized experts in reading and a panel of nationally recognized experts in mathematics. The panels were comprised of individuals from the K-12, postsecondary, research, and policy spheres, knowledgeable about academic preparedness for college reading and college mathematics. The panels compared the 12th grade NAEP reading and mathematics frameworks and the ADP benchmarks.

**Reading**

The Achieve reading panel found considerable similarity between NAEP and the ADP benchmarks for English, although not perfect agreement. This is displayed in the side-by-side chart on pages 30-40 of the Achieve Reading Report (http://www.nagb.org/content/nagb/assets/documents/commission/researchandresources/Achieve%20Reading%20Report.pdf). The English benchmarks have eight major components and objectives under each component. Three of these major components were deemed “Not Applicable” to the reading domain: writing, research, and media.

For almost all of the applicable objectives under the five major components that were applicable to the reading domain, the Achieve reading panel found matches in the NAEP 2009 reading framework. Overall, the panel concluded that “…the 2009 NAEP Reading Framework…was aligned to the ambitious [ADP] benchmarks” (Achieve Reading Report, p. 2).

The reading panel also listed items in the NAEP framework that are not found in the ADP English benchmarks. For example, under Argumentation and Persuasive Text, figurative language and rhetorical structure, including parallel structure and repetition, was present in the NAEP reading framework at grade 12, but not in the ADP benchmarks. Under Poetry, tone, complex symbolism, and extended metaphor and analogy were present in the NAEP reading framework but not the ADP benchmarks. A complete listing of the items in the NAEP framework not present in the ADP benchmarks appears on page 41 of the Achieve Reading Report.

Although the Achieve reading panel concluded that the 12th grade NAEP reading framework for 2009 was aligned with the ADP benchmarks applicable to reading, the panel’s report does include six recommendations. The Governing Board approved these recommendations on February 14, 2005. For example, the Achieve reading panel recommended increasing the percentage of informational text passages from 60% to 70% and to feature additional items that ask students to compare texts. The changes were modest, sufficiently so to permit continuation of the 12th grade trend line from its initiation in 1992.

The NAEP reading framework used for the 2009, 2011, and 2013 assessments contains the following statement
In May 2005, the Governing Board adopted a policy statement regarding NAEP and 12th-grade preparedness. The policy states that NAEP will pursue assessment and reporting on 12th-grade student achievement as it relates to preparedness for post-secondary education and training. This policy resulted from recommendations of the Board’s National Commission on NAEP 12th Grade Assessment and Reporting in March 2004. Subsequent studies and deliberations by the Board took place during 2004 and 2005.

In reading, the Board adopted minor modifications to the 2009 NAEP Reading Framework at grade 12 based on a comprehensive analysis of the framework conducted by Achieve, Inc. The current version of the reading framework incorporates these modifications at grade 12 to enable NAEP to measure and report on preparedness for postsecondary endeavors (National Assessment Governing Board, 2008, Reading Framework, p. v).

Mathematics
The mathematics review began with the 2007 NAEP mathematics framework, which was the most current and included the changes approved for the 2005 12th grade mathematics assessment. The Achieve panel examined the NAEP mathematics framework at the 12th grade in relation to the ADP benchmarks for mathematics. The Achieve panel developed proposed revisions to the assessment objectives for grade 12. While acknowledging differences in language and purpose, the Achieve mathematics panel concluded that the “overall mathematics frameworks of ADP and [12th grade] NAEP are remarkably similar” (see http://www.nagb.org/content/nagb/assets/documents/commission/researchandresources/Achieve-Mathematics-Report.pdf, Achieve Mathematics Report, p.9).

The Governing Board convened a panel of mathematicians and mathematics educators to review and revise the objectives in relation to the objectives for grades 4 and 8. The panel conducted focus groups with various NAEP constituents, using repeated rounds of reviews. The Governing Board approved the final set of grade 12 objectives on August 5, 2006. The changes to the framework were sufficiently modest to permit the continuation of the 12th grade trend line begun with the 2005 12th grade mathematics assessment under the previous 12th grade framework. Like the reading framework, the 2009/2013 mathematics framework for grade 12 states the Board’s intention to measure 12th grade academic preparedness (National Assessment Governing Board, 2008, Mathematics Framework, pp. 2-3).

Conclusion
The Governing Board, by official action, revised the NAEP 12th grade reading and mathematics frameworks with the explicit purpose of measuring 12th grade academic preparedness for college, beginning with the 2009 assessments. Setting forth the measurement purpose and making relevant revisions to the NAEP assessment frameworks are necessary elements of the validity argument; however, they are not sufficient. Evidence must be considered with respect to the alignment of the framework and the test questions administered to the measurement purpose. This will be addressed in the next section.
### Examples of Objectives added to the 2009 Grade 12 Mathematics Framework

**Number properties and operations**

b) * Analyze or interpret a proof by mathematical induction of a simple numerical relationship.

**Measurement**

d) Interpret and use the identity \( \sin^2 \theta + \cos^2 \theta = 1 \) for angles \( \theta \) between 0° and 90°; recognize this identity as a special representation of the Pythagorean theorem.

e) * Determine the radian measure of an angle and explain how radian measurement is related to a circle of radius 1.

f) * Use trigonometric formulas such as addition and double angle formulas.

g) * Use the law of cosines and the law of sines to find unknown sides and angles of a triangle.

**Geometry**

e) * Use vectors to represent velocity and direction; multiply a vector by a scalar and add vectors both algebraically and graphically.

g) * Graph ellipses and hyperbolas whose axes are parallel to the coordinate axes and demonstrate understanding of the relationship between their standard algebraic form and their graphical characteristics.

h) * Represent situations and solve problems involving polar coordinates.

**Data Analysis, Statistics, and Probability**

c) * Draw inferences from samples, such as estimates of proportions in a population, estimates of population means, or decisions about differences in means for two “treatments.”

e) * Recognize the differences in design and in conclusions between randomized experiments and observational studies.

k) * Use the binomial theorem to solve problems.

e) * Recognize and explain the potential errors caused by extrapolating from data.

**Algebra**

e) Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or trigonometric functions from tables, graphs, or equations.

j) * Given a function, determine its inverse if it exists and explain the contextual meaning of the inverse for a given situation.

h) * Analyze properties of exponential, logarithmic, and rational functions.

g) * Determine the sum of finite and infinite arithmetic and geometric series.
➢ **Content Alignment Studies Found Significant Overlap between NAEP and the ACT, SAT and ACCUPLACER**

The Governing Board conducted studies to determine the degree of content similarity between NAEP 12th grade reading and mathematics assessments and relevant tests used for college admissions and placement.

The studies had two objectives. The first objective was to determine the degree to which the content of 12th grade NAEP in reading and mathematics covers the reading and mathematics knowledge and skills needed for first year college work. The SAT, ACT, and ACCUPLACER are well-established tests that assess individual students’ reading and mathematics proficiency in relation to college level expectations.

The ACT is developed with the purpose of “…[measuring] as directly as possible the degree to which each student has developed the academic skills and knowledge that are important for success in college…” (ACT Technical Manual, p. 62).

The SAT is developed “to ensure that the topics measured on the SAT…reflect what is being taught in the nation’s high schools and what college professors consider to be required for college success.” (Kim, Wiley, and Packman, p.1)

The ACCUPLACER has the purpose of “… [determining] which course placements are appropriate for [incoming college] students and whether or not remedial work is needed.” (ACCUPLACER, p. A-2)

The SAT, ACT and ACCUPLACER in reading and mathematics are widely used for these purposes by admissions and placement professionals in postsecondary education institutions. These testing programs regularly conduct curriculum surveys, validity studies and other research to support their claims that the content measured is directly related to the reading and mathematics knowledge and skills needed to qualify for entry-level credit-bearing courses (e.g., see the ACT curriculum studies for 2012, 2009, 2005, and 2002 at [http://www.act.org/research-policy/national-curriculum-survey/](http://www.act.org/research-policy/national-curriculum-survey/), and the College Board National Curriculum Survey on English and Mathematics at [http://research.collegeboard.org/publications/content/2012/05/national-curriculum-survey-english-and-mathematics](http://research.collegeboard.org/publications/content/2012/05/national-curriculum-survey-english-and-mathematics).

- Therefore, with the assumption that the SAT, ACT, and ACCUPLACER do measure the content needed for college level work, significant content overlap between NAEP and these other assessments would support the conclusion that what NAEP measures covers the knowledge and skills needed by college freshmen to be placed into entry-level credit bearing courses.

The second reason for conducting the content alignment studies was to provide information for interpreting the results of planned statistical linking studies between NAEP and the other tests, which measure academic preparedness for college. The linking studies were designed to examine how performance on NAEP compares with performance on the other tests, with the
purpose of supporting inferences about academic preparedness for college. For NAEP to support inferences about academic preparedness for college based on the linking studies, a sufficient content match would be needed between NAEP and the other tests, not just a statistical relationship.

The Content Alignment Studies: Overview

The Governing Board conducted content alignment studies in reading and mathematics comparing the 2009 12th grade NAEP and the ACT, SAT, and ACCUPLACER reading and mathematics tests. Overall, considerable overlap was found between the ACT and NAEP and the SAT and NAEP, with some differences. NAEP was found to measure much of what is measured on the ACCUPLACER, but the reading and mathematics domains measured by NAEP were much broader than ACCUPLACER. More details are provided in the summaries of the individual studies below.

The general design for the content alignment studies was to compare the 12th grade NAEP frameworks in reading and mathematics with the analogous document for the other test, and then to compare the test items from one test to the framework/analogous document of the other test. The reviews were performed by subject specific (i.e., mathematics, reading) panels, composed of experts in mathematics or reading and English instruction at the high school and college levels.

Alignment studies that compare an assessment to the content standards on which it is based are relatively common and have well-established methodologies. However, this is not true for the types of alignment studies the Governing Board planned to conduct: content alignment studies comparing different assessment programs. Different assessment programs have different purposes, different approaches to describing the domain being measured, and, possibly, different “grain size” in the level of detail in describing the domain.

The Governing Board contracted with Norman Webb, a noted expert in content alignment studies, to prepare a design document for conducting the assessment to assessment alignment studies. The purpose was to put in place a methodology that considered the special challenges of assessment to assessment alignment studies and to foster comparability in the conduct of the studies and the reporting metrics across studies and contractors. The link to the Webb design document is at (http://www.nagb.org/content/nagb/assets/documents/publications/design-document-final.pdf).

The Webb design was developed after the ACT alignment studies were completed. It was used in conducting the SAT and ACCUPLACER content alignment studies.

In the following sections are summaries of the content alignment study results, excerpted from the study reports. The results for the three content alignment studies in reading are presented first, followed by the three content alignment studies for mathematics, along with summary discussions for the reading and mathematics results.
The Content Alignment Studies: Reading Results

**Reading: ACT**

The Governing Board contracted with ACT, Inc. to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the ACT reading test. The full report can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf).

The reading panel was composed of 7 members, with expertise in reading and/or English instruction at the high school and college levels. The panel was about evenly divided in terms of prior familiarity with either the ACT or NAEP reading domains.

The panel found considerable similarity in the content of the NAEP 12th grade reading assessment and the ACT. For example, the NAEP 12th grade reading framework was compared to the ACT reading domain and the ACT College Readiness Standards for reading. The ACT College Readiness Standards (CRS) are descriptions of the content (i.e., the knowledge and skills) measured by the ACT reading test in score bands along the ACT 1-36 point scale from 13-36 (see [http://www.act.org/standard/planact/reading/](http://www.act.org/standard/planact/reading/)). The panel concluded that

“All of the skills highlighted in the ACT [reading] domain and in the [ACT] College Readiness Standards [for reading] were identified within the NAEP Reading framework. In performing the comparison in the other direction—NAEP to ACT—it was the sense of the panel that the ACT measured primarily those skills that NAEP identifies as Locate/Recall and Integrate/Interpret skills, those that pertain primarily to finding explicit information in text (what the ACT would call Referring skills) and to making inferences, drawing conclusions, and making generalizations from information within text (what the ACT would call Reasoning skills). The panel saw less evidence of the higher-level analytical and evaluative Critique/Evaluate skills in the ACT domain, and attributed that to the multiple-choice format of the ACT [whereas NAEP includes constructed response items as well as multiple choice]. Another difference is that NAEP includes items and texts measuring how well an examinee can apply reading skills across texts, whereas the paired passage format is not a feature of the ACT. So, while the NAEP Reading framework and the ACT Reading domain, test specifications, and College Readiness Standards share similarities, important differences in what and how the assessments measure suggest caution when drawing comparisons between the assessments.” (p.17)

The reading panel also conducted an item classification study, in which the NAEP 12th grade reading items were classified in relation to the ACT College Readiness Standards for Reading.

“A total of 152 Reading items (comprising 17 blocks) were classified in [the reading] study. Of these, 97 were multiple-choice (MC). Nine were dichotomously-scored (“incorrect” or “correct”) short constructed-response (DSCR) items. Thirty-three were polytomously-scored short constructed-response (PSCR) items, each scored using a three-point scoring rubric. Thirteen were extended constructed-response (ECR) items,
An item or score category was deemed “classified” if there was majority agreement; that is, if at least 4 of the 7 panel members agreed about the score band to which an item (or creditable score category under an item rubric) was assigned.

Of the 211 determinations to be made, there was only one for which there was no majority agreement (the assignment of a PSCR rubric to a CRS score band). Of the remaining 210 determinations, 181 were unanimous.

The reading panel was able to classify 137 items or rubric categories (about two-thirds of the determinations to be made) to the CRS score bands. Of the 97 multiple choice items, 81 (or 84%) were classified. Of the 113 rubric score categories for items, 56 (or 50%) were classified. The reasons some multiple choice items and rubric score categories could not be classified were related to the differences in the ACT and NAEP reading domains described above. These reasons include the presence of constructed response items in NAEP but not the ACT, the presence of items involving multiple texts in NAEP but not the ACT, and the greater presence of “Critique/Evaluate” type items in NAEP than the ACT.

Of the 137 classifications, 24 were in the score bands from 13-19; 113 of the classifications were in the score bands from 20-36. This is noted because the ACT College Readiness Benchmark for reading is 21. The ACT College Readiness Benchmark signifies the score at which a student has a 50% chance of attaining a grade of B or better in a relevant subject and a 75% chance of a C or better. In addition, the Governing Board conducted a survey of postsecondary institutions’ use of tests in making entry-level decisions about placement into remedial or regular credit-bearing courses. With respect to the ACT, 18 was the mean reading score below which students were deemed to need remedial course work (Fields and Parsad, P. 19). Whereas this provides a context for the study results, it must be kept in mind that in making their judgments about item classifications, the panelists did not have data about NAEP item difficulty or data on how performance on NAEP compares with performance on the ACT.

Finally, although the study results support the conclusion that the 12th grade NAEP reading assessment measures content directly related to academic preparedness for college, it is noted that the study was conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in reading.

Reading: SAT
The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the SAT critical reading test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found
Overall, the study found similar content in the NAEP 12th grade reading assessment and the SAT critical reading test. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

What is the correspondence between the reading content domain assessed by NAEP and that assessed by SAT?

The greatest commonality between the two tests is their shared emphasis on the broad skills of integrating and interpreting both informational and literary texts. This is evident in the majority of items from both tests aligned to NAEP Standard 2, Integrate/Interpret,” including many to Goal 2.1, “Make complex inferences within and across both literary and informational texts.”

Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), there is also considerable overlap at the level of more specific skills.

To what extent is the emphasis of reading content on NAEP proportionally equal to that on SAT?

Both tests had many of their item alignments to the same NAEP “Integrate/Interpret” objectives, often with similar percentages of alignments. Although there were some differences in emphasis, both tests also had notable percentages of alignments to SAT Objectives B.1.1–B.1.3 and B.1.5. Skills with overlap include inferring/analyzing the following:

- the “main idea” and “author’s purpose” (SAT Objective B.1.1 and NAEP Objectives 2.3.a and 2.1.f);
- the “tone and attitude” of an author or character (NAEP Objectives 2.2.a and 2.2.c and SAT Objective B.1.4);
- the use of “rhetorical strategies” (NAEP Objective 2.1.d and SAT Objective B.1.2); and
- connections between ideas, perspectives, or problems (NAEP Objective 2.1.b and SAT Objectives B.1.3 and B.1.5).

Additionally, in the area of greatest content overlap—items on both tests aligned to objectives for NAEP “Integrate/Interpret” and aligned to SAT “Passage-Based Reading” Objectives B.1.1–B.1.5—both tests met the typical threshold criteria for depth of knowledge consistency…

Despite these similarities, there are some notable differences in emphasis between the two assessments. Both tests assess vocabulary skills. However, NAEP addresses vocabulary exclusively in the context of passage comprehension, while the majority of SAT vocabulary items are in a sentence-completion format, in which context plays a more limited role. This difference reflects NAEP’s emphasis on the understanding of
word meaning in context; the assessment is not intended to measure students’ prior knowledge of word definitions. The SAT sentence-completion items provide some context within the single sentence text, but in many cases, students’ success on the items almost certainly depends on their prior knowledge of word definitions.

In addition, panelists found considerably less emphasis in SAT than in NAEP on literal comprehension and critical evaluation, particularly the evaluation of the quality or effectiveness of an author’s writing, skills covered in the NAEP standards “Locate/Recall” (locating/recalling specific details and features of texts) and “Critique/Evaluate” (evaluating texts from a critical perspective), respectively. This difference suggests a greater emphasis on these skills in NAEP.

Even with the minimal coverage of NAEP “Locate/Recall” and “Critique/Evaluate” standards by SAT items, all NAEP items found a match in the SAT framework. However, the broad language of the SAT framework can encompass the range of the NAEP items. For example, SAT Goal B.2, “Literal Comprehension,” refers to items that “ask what is being said” in a “small but significant portion of a reading passage,” a description that can easily accommodate most NAEP “Locate/Recall” items and objectives. In fact, nearly all items on the NAEP short version that were coded to “Locate/Recall” objectives in the NAEP framework were matched to SAT Goal B.2 in the SAT framework.

Similarly, SAT Objective B.1.3, to which approximately one-quarter of NAEP items aligned, includes “Evaluation,” the primary focus of NAEP “Critique/Evaluate.” The description in SAT Objective B.1.3 of items that “ask the test taker to evaluate ideas or assumptions in a passage” is compatible at a very general level with NAEP “Critique/Evaluate” objectives addressing the author’s point of view, logic, or use of evidence. SAT Objective B.1.2, “Rhetorical Strategies,” is also broad enough in its language to make it a reasonable match for some NAEP “Critique/Evaluate” items focused on “author’s craft” or use of “literary devices.” In the NAEP short version, all items that aligned to “Critique/Evaluate” objectives in the NAEP framework were aligned to either SAT Objectives B.1.2 or B.1.3, or both.

**Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire reading subdomains are missing or not aligned?**

With regard to differences in content as described in the NAEP framework, SAT items had limited coverage of the knowledge and skills described by the NAEP standards “Locate/Recall” and “Critique/Evaluate.” This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.
With regard to differences in complexity, NAEP items and objectives had a range of depth of knowledge including items at DOK Levels 1, 2, and 3, while SAT items and objectives were coded primarily at Levels 2 and 3.

Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT “Passage-Based Reading” objectives and NAEP “Integrate/Interpret” objectives, as well as some important differences.

**Reading: ACCUPLACER**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the ACCUPLACER reading test. The ACCUPLACER is used specifically to determine whether entry-level students have the reading skills necessary for college level work or require remedial reading courses. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf).

Overall, the study found similar content in the NAEP 12th grade reading assessment and the ACCUPLACER reading test, although the content of NAEP is much broader and complex. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

*What is the correspondence between the reading content domain assessed by NAEP and that assessed by ACCUPLACER?*

The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts.

The shared emphasis on the comprehension and interpretation of informational text can also be seen in the alignments on both tests to the ACCUPLACER framework. Although the ACCUPLACER standards do not explicitly refer to text type, they focus almost exclusively on elements typical of informational text. A majority of both NAEP and ACCUPLACER items were matched to the ACCUPLACER standard “Inferences,” and both tests had notable percentages of alignments to “Direct statements and secondary ideas” and “Applications.” A smaller percentage of items on both tests were aligned to “Identifying main ideas.”
To what extent is the emphasis of reading content on NAEP proportionally equal to that on ACCUPLACER?

As previously discussed, the alignments both within and across frameworks show that both tests emphasize the comprehension and interpretation of informational text, particularly through the use of inference. Within this broad area of convergence, however, there are differences in emphasis revealed in the alignments to specific objectives within both frameworks. In relation to the NAEP framework, the NAEP short-version items showed a far greater emphasis on the comprehension of vocabulary in context (Objective 4.a) and on the analysis of an author’s use of language (Objective 1.d). In relation to the ACCUPLACER framework, NAEP items showed more emphasis on the use of inference to interpret text (“Inferences”). The higher percentage of NAEP items aligned to “Applications” also reflects the greater emphasis in NAEP on understanding authors’ use of language.

In relation to the ACCUPLACER framework, the ACCUPLACER items showed a greater emphasis than the NAEP items on the identification of main ideas. In relation to the NAEP framework, the ACCUPLACER items showed more emphasis on the recall of specific details, facts, and information (NAEP 1.1.a).

In general, in the cross-framework alignments, the matches found in each test to the other’s framework (NAEP to ACCUPLACER and ACCUPLACER to NAEP) tended to be for the most general objectives within that framework. For example, the great majority of hits for ACCUPLACER items to NAEP objectives for “Integrate/Interpret” were to two of the most broadly stated NAEP objectives, “Draw conclusions” (2.3.b) and “Compare or connect ideas” (2.1.b). Many of the more specific NAEP objectives for “Integrate/Interpret,” such as “Find evidence in support of an argument” (2.2.c), received far fewer or no hits from ACCUPLACER items. Compared to ACCUPLACER, the NAEP items were more evenly distributed among NAEP objectives.

The majority of alignments for NAEP items to ACCUPLACER standards were also to the broadest of those standards—“Inferences” and “Applications,” both of which overlap in content with a number of NAEP objectives but at a higher level of generality. The more specific ACCUPLACER standard, “Identifying main ideas,” received far fewer alignments from NAEP items.

Are there systematic differences in content and complexity between the NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between the NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire reading subdomains are missing or not aligned?

In regard to differences in content, NAEP addresses reading skills related to both literary and informational text, while ACCUPLACER does not address reading skills specific to literary text. As expected, based on the framework-to-specifications [review]… ACCUPLACER items had minimal matches to NAEP objectives for literary text. The main area of alignment of ACCUPLACER items to the NAEP framework, NAEP
objectives in “Locate/Recall” and “Integrate/Interpret,” applied to informational text only or to both informational and literary text.

The ACCUPLACER items also had minimal to no coverage of the NAEP standard “Critique/Evaluate.” … overall, the language of the ACCUPLACER objectives (“understand,” “comprehend,” “recognize”) places more emphasis on comprehension and interpretation of text (“distinguish the main idea from supporting ideas” or “perceive connections between ideas made—implicitly—in the passage”) than on critical analysis or evaluation (“Evaluate the strength and quality of evidence used by the author to support his or her position” in NAEP Objective 3.3.b, or “Judge the author's craft and technique” in NAEP Objective 3.1.a).

In regard to complexity, both assessments were found to meet the criteria for depth of knowledge consistency in relation to their own framework. In relation to the NAEP framework, however, only the NAEP items met the criteria for DOK consistency for all NAEP standards. The ACCUPLACER items met the criteria for depth of knowledge consistency only for NAEP “Locate/Recall.” Although the majority of the ACCUPLACER item alignments were to objectives for NAEP “Integrate/Interpret,” over half of these items were found to have a DOK level below that of the standard. In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the more in-depth reasoning required by DOK Level 3. NAEP, by contrast, includes much longer reading passages and both multiple-choice and constructed-response items.

NAEP covers skills specific to the comprehension and analysis of literary text while ACCUPLACER does not. In addition, NAEP covers the skills of evaluating and critiquing text, skills not addressed by ACCUPLACER. Finally, NAEP has a wider range of cognitive complexity than ACCUPLACER, with a substantially higher percentage of items at DOK Level 3, requiring more in-depth analysis or evaluation. However, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

Overall, the NAEP items covered a broader range of cognitive complexity than the ACCUPLACER items. This is also apparent in the frameworks. The three NAEP standards, defined in terms of three different “cognitive targets” (“Locate/Recall,” “Integrate/Interpret,” and “Critique/Evaluate”), cover a broader range of cognitive complexity supported by the use of longer reading passages and the inclusion of both short and extended constructed-response items. The language of the ACCUPLACER standards (“understand,” “comprehend,” “recognize”) places more emphasis on comprehension and interpretation of text (e.g., “distinguish the main idea from supporting ideas” in ACCUPLACER A, “Identifying main ideas,” or “perceive connections between ideas made—implicitly—in the passage” in ACCUPLACER C, “Inferences”) than on critical analysis or evaluation (e.g., “Evaluate the strength and quality of evidence” in NAEP 3.3.b, or “Judge the author’s craft” in NAEP 3.1.a). In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the cognitive complexity typical of DOK Level 3 items. Although the
NAEP items show a greater range of cognitive complexity and a greater emphasis on critical thinking, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

The Content Alignment Studies: Summary Discussion for Reading

The NAEP 12th grade reading framework, test questions, and, for constructed response items, the score category rubrics, were compared with the analogous domain descriptions and test questions for the ACT, SAT, and ACCUPLACER reading tests. These three tests are used for college admissions and placement. They are well established and have been used for these purposes for many years by professionals in postsecondary education. The test publishers regularly survey secondary and postsecondary educators about relevant content and have conducted research that supports the validity of the test content for the intended inferences and uses. The underlying assumption is that if the content of the 12th grade NAEP reading assessment is similar to the content of these reading tests, then the NAEP content is directly related to “academic preparedness for college.”

The ACT study found that “All of the skills highlighted in the ACT [reading] domain and in the [ACT] College Readiness Standards [for reading] were identified within the NAEP Reading framework.” At the same time, there was content measured by NAEP that was not present in the ACT reading test. In assigning 211 NAEP 12th grade reading items and rubric score categories to the ACT College Readiness Standards for reading, there were 137 positive classifications, or about 65% of the possible classifications. The multiple choice items and rubric score categories that could not be classified were those that measured content not measured by the ACT reading test.

The SAT study found that “Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT “Passage-Based Reading” objectives and NAEP “Integrate/Interpret” objectives, as well as some important differences.” With respect to the differences, “…SAT items had limited coverage of the knowledge and skills described by the NAEP standards “Locate/Recall” and “Critique/Evaluate.” This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.”

The ACCUPLACER study found that “The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts…Overall, the NAEP [frameworks and] items covered a broader range of cognitive complexity than the ACCUPLACER items…The three NAEP standards, defined in terms of three different “cognitive targets” (“Locate/Recall,” “Integrate/Interpret,” and “Critique/Evaluate”), cover a broader range of cognitive complexity supported by the use of
longer reading passages and the inclusion of both short and extended constructed-response items.”

The results across the three studies are consistent. In general, the content of the ACT, SAT, and ACCUPLACER reading tests are present in NAEP, but NAEP is generally broader. Alignment between NAEP and the other three respective assessments is substantial, but not perfect; perfect alignment is not expected. A component of the SAT critical reading assessment not present in NAEP is sentence completion, measuring vocabulary knowledge in a different way than NAEP does.

These results support the conclusion that
- The NAEP 12th grade reading assessment measures academic knowledge and skills that are also covered by other assessments designed and used to make judgments about the academic preparedness of college freshmen for placement into entry-level, credit bearing, non-remedial college courses that meet general education degree requirements, and
- NAEP 12th grade reading test items and rubric scoring categories for items are appropriate for obtaining evidence of test takers’ possession of knowledge and skills needed for college freshmen to be placed into ECNRG courses requiring college level reading.

The Content Alignment Studies: Mathematics Results

Mathematics: ACT
The Governing Board contracted with ACT, Inc. to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACT mathematics test. The full report can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf.

The mathematics panel was composed of 7 members, with expertise in mathematics instruction at the high school and college levels. The panel was about evenly divided in terms of prior familiarity with either the ACT or NAEP mathematics domains.

The panel found considerable similarity in the content of the NAEP 12th grade mathematics assessment and the ACT. For example, the NAEP 12th grade mathematics framework was compared to the ACT mathematics domain and the ACT College Readiness Standards for mathematics. The ACT College Readiness Standards (CRS) are descriptions of the content (i.e., the knowledge and skills) measured by the ACT mathematics test in score bands along the ACT 1-36 point scale from 13-36 (see http://www.act.org/standard/planact/math/index.html). The panel concluded that

“… the two assessments have much of their content domains in common. However, in the NAEP-to-ACT comparison, the difference in specificity with which the domains are articulated in the assessment documents left the panel uncertain as to whether a number
of NAEP content topics—those pertaining to transformations, probability, statistics, and data analysis—are assessed by the ACT. In addition, there was some uncertainty within the panel on the degree to which higher-order analytic skills were assessed, and it was the sense of the panel that the ACT Mathematics Test contained few items involving high mathematical complexity, at least as the NAEP defines it. With regard to the ACT to-NAEP comparison, the Mathematics panel found nearly all of the ACT Mathematics domain and College Readiness Standards reflected in the NAEP Mathematics domain, but determined that a number of the lower-level topics in the ACT Pre-Algebra subdomain were more consistent with Grade 8 NAEP topics. All of these points suggest that while there may be substantial overlap in what the two assessments measure and how they measure it, there are areas of difference, as well. (p. 17)

The mathematics panel also conducted an item classification study, in which the NAEP 12th grade mathematics items were classified in relation to the ACT College Readiness Standards for Mathematics.

An item or score category was deemed “classified” if there was majority agreement; that is, if at least 4 of the 7 panel members agreed about the score band to which an item (or creditable score category under an item rubric) was assigned.

Of the 229 determinations to be made, panel members believed that every item or rubric category could be classified to some CRS score range. However, there were 39 for which there was no majority agreement (17 multiple choice items and 22 rubric categories) on what the classification should be; therefore those items were not considered assigned to a CRS score band. Of the remaining 190 determinations, 24 were unanimous, 142 involved classifications to adjacent score ranges and 24 involved classifications to non-adjacent score ranges.

Of the 108 multiple choice items, 91 (or 84%) were classified. Of the 121 rubric score categories for items, 99 (or 82%) were classified.

Of the 190 classifications, 10 were in the score bands from 13-19; 180 of the classifications were in the score bands from 20-36. This is noted because the ACT College Readiness Benchmark for mathematics is 22. The ACT College Readiness Benchmark signifies the score at which a student has a 50% chance of attaining a grade of B or better in a relevant subject and a 75% change of a C or better. In addition, the Governing Board conducted a survey of postsecondary institutions’ use of tests in making entry-level decisions about placement into remedial or regular credit-bearing courses. With respect to the ACT, 19 was the mean mathematics score below which students were deemed to need remedial course work in mathematics (Fields and Parsad, p. 13). Although this provides a context for the study results, it must be kept in mind that in making their judgments about content, the panelists did not have data about NAEP item difficulty or data on how performance on NAEP compares with performance on the ACT.

Finally, although the study results support the conclusion that the 12th grade NAEP mathematics assessment measures content that is also covered by other assessments designed and used to make judgments about academic preparedness for college, it is noted that the study was
conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in mathematics.

**Mathematics: SAT**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the SAT mathematics test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf.

Overall, the study found similar content in the NAEP 12th grade mathematics assessment and the SAT mathematics test. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

“What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by SAT?

At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard. Each framework contains both general and specific objectives, although the SAT objectives, which are presented as content topics without indication of the cognitive level at which that content would be assessed, may be interpreted as more general than the NAEP objectives.

Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions.

To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on SAT?

The greatest commonality between the two tests is their emphasis at the standard level. This is evident in the distribution of percentages of total hits from both assessments matched to each set of standards. Although there are some differences of emphasis, such as the full NAEP item pool’s greater proportion of alignment to SAT “Data analysis, statistics, and probability,” and the SAT short-version’s greater proportion of alignment to SAT “Geometry and measurement,” the proportions of alignments to “Algebra and functions” and “Number and operations” are comparable. There is also considerable overlap among some specific skills, with both assessments addressing many of the same NAEP “Number properties and operations” objectives and SAT objectives…
Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), it is clear that both assessments emphasize a number of the same or closely related skills. These include properties, equivalence, and operations on rational numbers (included in NAEP Goals 1.1 and 1.3 and included in SAT Objective N.2) and properties of two-dimensional shapes (included in NAEP Goals 3.1 and 3.3 and included in SAT Objective G.6).

Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework. The language of the objectives in the SAT framework is sufficiently broad to encompass the range of the NAEP items. For example, SAT Objective A.10, “Basic concepts of algebraic functions,” may accommodate most of the items aligning to the seven objectives within NAEP Goal 5.1, “Patterns, relations, and functions.” Finally, some NAEP items were found to be uncodable to the SAT objectives. These items assessed skills not present in the SAT framework.

The two tests are also similar in the average DOK [Depth of Knowledge] levels of items. However, while most items in both tests were found to be at DOK Level 2, NAEP items had a wider range of DOK than did SAT items, with more NAEP items coded to Levels 1 and 3. The Level 3 NAEP items often involved application of concepts through short or extended constructed-response items. Both tests also met depth-of-knowledge consistency overall (with each not meeting this criterion for only one standard as rated by one panel).

Overall, despite differences in alignment at the detailed specific objective level, differences in emphasis at the standard level, and a small difference in ranges of depth of knowledge, there is considerable overlap of content and complexity between [the NAEP 12th grade mathematics assessment and the SAT mathematics test].”

Mathematics: ACCUPLACER
The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test. The ACCUPLACER is used specifically to determine whether entry-level students have the mathematic knowledge and skills necessary for college level work or require remedial mathematics courses.
WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf.

Overall, the study found similar content in the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test, although the content of NAEP is much broader and complex. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

“What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?

The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra (included in NAEP’s “Number properties and operations” and “Algebra” standards and in ACCUPLACER’s “Arithmetic,” “Elementary algebra,” and “College level math” standards), although their respective degrees of alignment and focus in these subdomains vary. Whereas the NAEP items focus primarily on number or number operations and algebra content at the grade 12 level, with an emphasis on problem solving and application of concepts at that grade level, the ACCUPLACER items span a wider developmental and grade-level range (from basic to more advanced). This difference in focus is consistent with the purposes of the two assessments and their frameworks. The NAEP objectives are written to describe assessable content for grade 12 mathematics; thus, the 130 objectives tend to address the skills and concepts specific to that grade. The purpose of ACCUPLACER is to help determine appropriate placement for an individual student, and so the 87 ACCUPLACER objectives are spread more broadly across grade levels and are intended to be more general.

To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?

Regarding alignment to the NAEP framework, within the “Number properties and operations” and “Algebra” standards, NAEP items had broader overall coverage of the NAEP objectives than did ACCUPLACER. The 42 NAEP items (the short version used for within-framework alignment) aligned to 72 NAEP objectives, whereas the 105 ACCUPLACER items (one complete form of each of the three ACCUPLACER Mathematics Core tests) aligned to only 56 NAEP objectives, with 44% of the ACCUPLACER item alignments aligning to only three NAEP objectives (all in “Number properties and operations” and “Algebra”). These differences in breadth and emphasis between the two assessments were evident across all NAEP standards. For example, in each assessment, items were aligned to four NAEP “Algebra” objectives for which the other assessment had no alignments, reflecting differences in emphasis within that standard.

Regarding alignment to the ACCUPLACER framework, ACCUPLACER items in the short version of 45 items covered all three standards—“Arithmetic,” “Elementary algebra,” and “College level math”—with a relatively even distribution, although
“College level math” had the lowest percentage of item alignments. NAEP items in the full pool of 164 items also covered “Arithmetic,” “Elementary algebra,” and “College level math,” with a fairly even distribution of approximately one-third of NAEP codable items aligned to each standard, although “Elementary algebra” received somewhat fewer item alignments. Despite these differences in emphasis, however, considering only codable items, the percentages of alignments to each ACCUPLACER standard were relatively evenly distributed in both assessments and similar in distribution across assessments. At the objective level, the distribution of item alignments to objectives was relatively even on both tests, although each assessment was aligned to some objectives to which the other was not.

In summarizing cross-framework alignment, there was somewhat less even distribution of items than observed in within-framework alignment. The majority of items on each test were found to align to objectives on the other test. However, the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from “Number properties and operations” in the NAEP framework, and four of the five goals in “Algebra.” Conversely, the NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives.

Are there systematic differences in content and complexity between NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives (20 were rated uncodable by the majority of panelists in each panel). For other NAEP items that were aligned to an ACCUPLACER objective, there were often parts of those items not addressed by the objective. These items were coded as aligned, since they do assess an ACCUPLACER objective, but parts of the items also cover other skills not included in the ACCUPLACER framework.

Regarding differences in alignment of complexity, the items from both tests that aligned to the NAEP standards met the typical depth-of-knowledge (DOK) consistency threshold; that is, the items assessed the objectives at or above the DOK level of the objective. The items from both tests that aligned to the ACCUPLACER standards had somewhat different ranges of DOK. The ACCUPLACER short-version items were divided fairly evenly between Level 1 and Level 2. The NAEP items aligned to the ACCUPLACER framework had a wider range of DOK, with items at Level 1, 2, and 3, and a greater emphasis on Level 2 than was in the ACCUPLACER items.”
The Content Alignment Studies: Summary Discussion for Mathematics

The NAEP 12th grade mathematics framework, test questions, and, for constructed response items, the score category rubrics, were compared with the analogous domain descriptions and test questions for the ACT, SAT, and ACCUPLACER mathematics tests. These three tests are used for college admissions and placement. They are well established and have been used for these purposes for many years by professionals in postsecondary education. The test publishers regularly survey secondary and postsecondary educators about relevant content and have conducted research that supports the validity of the test content for the intended inferences and uses. The underlying assumption is that if the content of the 12th grade NAEP mathematics assessment is similar to the content of these mathematics tests, then the NAEP content is directly related to “academic preparedness for college.”

The ACT study found that “With regard to the ACT to-NAEP comparison…nearly all of the ACT Mathematics domain and College Readiness Standards [are] reflected in the NAEP Mathematics domain, but…a number of the lower-level topics in the ACT Pre-Algebra subdomain were more consistent with Grade 8 NAEP topics.” In the NAEP-to ACT comparison, there was uncertainty about “…whether a number of NAEP content topics—those pertaining to transformations, probability, statistics, and data analysis—are assessed by the ACT…and the degree to which higher-order analytic skills were assessed…and it was the sense of the panel that the ACT Mathematics Test contained few items involving high mathematical complexity, at least as the NAEP defines it.”

The SAT study found similar content in the NAEP 12th grade mathematics assessment and the SAT mathematics test. “At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard…Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions…While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework.

The ACCUPLACER study found that “The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra…although their respective degrees of alignment and focus in these subdomains vary…the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from “Number properties and operations” in the NAEP framework, and four of the five goals in “Algebra.” Conversely, the
NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives. Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives.

The results across the three studies are consistent. In general, the content of the ACT, SAT, and ACCUPLACER mathematics tests are present in NAEP, but NAEP is generally broader. Alignment between NAEP and the other three respective assessments is substantial, but not perfect; perfect alignment is not expected.

These results support the conclusion that

- The NAEP 12th grade mathematics assessment measures academic knowledge and skills that is also covered by other assessments designed and used to make judgments about the academic preparedness of college freshmen for placement into entry-level, credit bearing, non-remedial college courses that meet general education degree requirements for mathematics, and
- NAEP 12th grade mathematics test items and rubric scoring categories for items are appropriate for obtaining evidence of test takers’ possession of knowledge and skills needed for college freshmen to be placed into ECRNG college mathematics courses.

**Discussion of Test Uses and Consequences in Relation to the Proposed Inferences**

The National Assessment of Educational Progress is an independent monitor of student academic achievement in the United States. It reports on achievement at specific points in time and trends in achievement over time. NAEP reports to the public, national and state policymakers, and education leaders. It assesses student achievement at grades 4, 8, and 12 in important subjects. NAEP is used to compare performance across states and for 21 urban school districts. NAEP results are reported by gender, race/ethnicity, socioeconomic status, and for students with disabilities and students who are English language learners.

The audiences and the uses of NAEP are well established. They will not change as a result of the added meaning afforded by the inferences proposed in this validity argument. However, providing familiar external referents for performance on 12th grade NAEP will greatly enhance the understanding of NAEP results by its audiences.

Currently, there are either no or very low stakes consequences associated with the use of NAEP results. NAEP is not used as a basis for evaluating or diagnosing individual students, classroom or school performance, the effectiveness of individual teachers or administrators, or for any other accountability purpose. This will not change as a consequence of the inferences proposed in this validity argument.
Although the uses and consequences of NAEP will not change, employing the proposed inferences for NAEP reporting will bring a potential for misinterpretation. NAEP reports should include text explaining the limitations on interpretation and other caveats that were discussed in detail on pages 8-10 above.

Summary and Conclusion

The National Assessment Governing Board decided to determine the feasibility of transforming NAEP into a measure of academic preparedness for college. Consequently, the Governing Board made changes to the NAEP 12th grade reading and mathematics frameworks with the explicit purpose of measuring academic preparedness for college. The Governing Board conducted research that established a high degree of overlap between the content of the NAEP 12th grade reading and mathematics assessments and the content of widely used college admissions and placement tests.

Through a partnership with the College Board, performance on 12th grade NAEP was compared with performance on the SAT mathematics and critical reading assessments, with correlations of .91 and .74 respectively. Analyses of these data examined the average NAEP scores and interquartile ranges for students scoring “at” and “at or above” the College Board College Readiness Benchmarks for reading and mathematics. Similar analyses were conducted using data from the 2005 and 2009 NAEP High School Transcript Studies, using the college readiness benchmarks developed by ACT and by the College Board. A longitudinal study was conducted in partnership with the Florida Department of Education, following the 12th grade students in the state NAEP sample into Florida public postsecondary institutions, employing Florida’s longitudinal data base. The average NAEP scores and interquartile ranges were calculated for the Florida students in relation to the ACT or SAT college readiness benchmarks, whether they achieved a first-year GPA of B- or better, and whether they were placed into a remedial course in their first year of college.

The results of these analyses were consistent across studies and across years. They support the conclusions that students in the NAEP 12th grade distribution at or above the Proficient achievement level in reading and at or above 163 on the NAEP score scale for mathematics are

- likely to be academically prepared for entry-level, credit-bearing non-remedial courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions, and

- not likely to need remedial/developmental courses in reading or mathematics in college

That the NAEP sampling, scaling and statistical procedures yield accurate estimates of the percentage of students scoring at or above a selected cut-score (i.e., NAEP achievement level) is well established as a result of numerous validity studies and evaluations.

Thus, the NAEP 12th grade preparedness research results support the inferences that
For reading:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

the percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.

For mathematics:

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

Including these inferences in NAEP 12th grade reports will add meaning to the interpretation of the NAEP 12th grade results. However, steps must be taken to avoid potential misinterpretation. NAEP reports using these inferences must also include the limitations on interpretation and caveats described previously in this validity argument. In addition, the reports should explain the rationale for NAEP reporting on academic preparedness and describe appropriate and inappropriate uses of the results.
References


WestEd ACCUPLACER Reading Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf).

WestEd SAT Mathematics Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pdf).

WestEd SAT Reading Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf).

<table>
<thead>
<tr>
<th>Year</th>
<th>National</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>Reading</td>
<td>Reading (4, 8)</td>
</tr>
<tr>
<td></td>
<td>MATHEMATICS</td>
<td>MATH (4, 8)</td>
</tr>
<tr>
<td></td>
<td>Science</td>
<td>Science (4, 8)</td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>U.S. History</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Civics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ECONOMICS (12)</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Reading (4, 8)</td>
<td>Reading (4, 8)</td>
</tr>
<tr>
<td></td>
<td>Mathematics (4, 8)</td>
<td>Math (4, 8)</td>
</tr>
<tr>
<td></td>
<td>Writing (8, 12)</td>
<td>Writing (8)</td>
</tr>
<tr>
<td>2008</td>
<td>Arts (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long-term trend</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>READING Mathematics*</td>
<td>READING (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>SCIENCE**</td>
<td>Math (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td>SCIENCE (4, 8)</td>
</tr>
<tr>
<td>2010</td>
<td>U.S. History</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Civics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geography</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Reading (4, 8)</td>
<td>Reading (4, 8)</td>
</tr>
<tr>
<td></td>
<td>Mathematics (4, 8)</td>
<td>Math (4, 8)</td>
</tr>
<tr>
<td></td>
<td>Science (8)**</td>
<td>Science (8)</td>
</tr>
<tr>
<td></td>
<td>WRITING (8, 12)**</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Economics (8, 12)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long-term trend</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Reading</td>
<td>Reading (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>Math (4, 8, 12)</td>
</tr>
<tr>
<td>2014</td>
<td>U.S. History (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Civics (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Geography (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TECHNOLOGY AND ENGINEERING LITERACY (8) **</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Reading</td>
<td>Reading (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>Math (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>Science**</td>
<td>Science (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>High School Transcript Study</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>Arts (8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long-term trend</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>Reading</td>
<td>Reading (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>Mathematics</td>
<td>Math (4, 8, 12)</td>
</tr>
<tr>
<td></td>
<td>Writing**</td>
<td>Writing (4, 8, 12)</td>
</tr>
</tbody>
</table>

*New framework for grade 12 only.

**Assessments involving test administration by computer.

NOTES:
1. Grades tested are 4, 8, and 12 unless otherwise indicated, except that long-term trend assessments sample students at ages 9, 13, and 17 and are conducted in reading and mathematics.
2. Subjects in **BOLD ALL CAPS** indicate the year in which a new framework is implemented or assessment year for which the Board will decide whether a new or updated framework is needed.
3. In 2009, 12th grade assessments in reading and mathematics at the state level were conducted as a pilot in 11 volunteering states (AR, CT, FL, IA, ID, IL, MA, NH, NJ, SD, WV). For 2013, 13 states agreed to participate (with MI and TN added).
4. The Governing Board intends to conduct assessments at the 12th grade in World History and Foreign Language during the assessment period 2018-2022.
Dr. Lillian M. Lowery became Maryland State Superintendent of Schools and Secretary-Treasurer of the State Board on July 1, 2012. She moved to Maryland from Delaware, where she served as Secretary of Education since January 2009. As the State Chief, she facilitated a broad-based statewide strategic planning and grant application process, resulting in Delaware being selected one of the first states to be awarded a coveted federal Race to the Top grant. She led efforts to establish Teacher and Leader Effectiveness Units as well as Turnaround Units within the Department, which have assisted in the implementation of the grant and the improvement of instructional opportunities for students.

Dr. Lowery has worked in various education institutions and programs since 1976. Prior to becoming Secretary of Education, she served for three years as the Superintendent of the Christina School District in New Castle County, DE. She has served as an Assistant Superintendent in Fairfax County, VA, and an Area Superintendent in Fort Wayne, IN. She held other administrative positions, and began her career as an English teacher at the middle and high school levels in Virginia.

A graduate in English and Secondary Education at North Carolina Central University, Dr. Lowery received her Master’s of Education in Curriculum and Instruction at the University of North Carolina, and her Doctorate in Educational Leadership and Policy Studies from Virginia Polytechnic and State University.
# National Assessment Governing Board
## Committee on Standards, Design and Methodology

**August 2, 2013
10:00 am – 12:30 pm**

## AGENDA

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Attachment</th>
</tr>
</thead>
</table>
| 10:00 – 10:10 am | Introductions and Welcome to Sharyn Rosenberg, the New Governing Board Assistant Director for Psychometrics  
*Lou Fabrizio, COSDAM Chair* | Attachment A |
| 10:10 – 10:20 am | Committee Questions on Information Items (see below)  
*Cornelia Orr, Executive Director, NAGB  
Michelle Blair, Senior Research Associate, NAGB* |            |
| 10:20 – 11:20 am | Interpreting NAEP Proficient using Preparedness Research Findings  
- Summary of Recent Feedback on the Validity Argument  
- Report Mock-up Presentation  
*Lou Fabrizio, COSDAM Chair  
Ray Fields, Assistant Director for Policy and Research, NAGB* | Attachment B |
| 11:20 – 12:25  | Discussion on Achievement Level Setting (ALS) on the 2014 NAEP Technology and Engineering Literacy (TEL) Assessment  
- 2013 TEL Pilot Scaling Analyses and Plans  
- Implications for Planning the TEL ALS  
*Cornelia Orr, Executive Director, NAGB  
Andreas Oranje, ETS* | Attachment C |
| 12:25 – 12:30 pm | Other Issues or Questions  
*COSDAM Members* |            |

## Closed Session

- Discussion on Achievement Level Setting (ALS) on the 2014 NAEP Technology and Engineering Literacy (TEL) Assessment
  - 2013 TEL Pilot Scaling Analyses and Plans
  - Implications for Planning the TEL ALS

## Open Session

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Attachment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Item</td>
<td>Update on Evaluation of NAEP Achievement Levels Procurement</td>
<td>Attachment D</td>
</tr>
</tbody>
</table>
| Information Item | NAEP 12th Grade Academic Preparedness Research: Phase 2 Research Updates  
- Course Content Analysis Research  
- National and State Partnerships  
- Research with Frameworks | Attachment E |
Welcome to Sharyn Rosenberg,
New Governing Board Assistant Director for Psychometrics

The Governing Board welcomes Dr. Sharyn Rosenberg to the NAGB staff as the new Assistant Director of Psychometrics. Dr. Rosenberg officially began working with the NAGB staff on July 15, 2013, and she is well-steeped in knowledge about NAEP through her previous work with NCES and NAEP as Senior Research Scientist/Psychometrician at the American Institutes for Research (AIR).

Abbreviated Professional Biography for Sharyn Rosenberg

Sharyn Rosenberg has an extensive background in education, which began nearly 20 years ago in a school reform class at Brown University with Theodore Sizer. She chose a major in Educational Studies as a direct result of taking this course. Sharyn received her M.A. and Ph.D. degrees in Educational Psychology, Measurement, and Evaluation at UNC-Chapel Hill. Greg Cizek, a former NAGB member, directed her studies and dissertation. The focus of her graduate work was on measurement and quantitative methods, including sampling theory, research methods, and advanced statistical/psychometric methodologies. She also earned a Certificate in Survey Methodology from the Odum Institute. In 2011, she and Cizek co-authored a chapter entitled, “Psychometric Methods and High Stakes Assessment: Contexts and Methods for Promoting Ethics in Testing,” which appears in the Handbook of Ethics in Quantitative Methodology.

Her work experiences include Horizon Research, where she conducted complex data analyses and provided psychometric expertise for projects, and the American Institutes for Research (AIR), where she provided research and psychometric support for NAEP. At AIR, Sharyn most recently served as the Project Director for the NAEP research and technical support team where she managed and conceptualized NAEP research studies, as well as responded to technical requests from the NCES Assessment Division. Her knowledge of NAEP and the work of NAGB are extensive.
Interpreting NAEP Results Using Preparedness Research Findings

At the August 2013 meeting, the Governing Board will discuss the way forward on reporting the NAEP 12th grade results from the 2013 reading and mathematics assessments. As background for the discussion, included in this tab are:

- a draft of a prototype chapter for the report (Attachment B-1; new document)
- the independent technical reviews of the preparedness validity argument by Gregory Cizek and Mark Reckase (Attachments B-2 and B-3; new documents)
- the preparedness validity argument (Attachment B-4; included in the May 2013 COSDAM briefing materials, but changes were made to the proposed inferences as described below and indicated in highlighting on pages B32 and B65)

The draft prototype chapter was prepared as an example of what NAEP reporting on academic preparedness for college would look like in the report of the 2013 12th grade assessment results.

As previously reported to the Governing Board, the Board staff and NCES staff have been working collaboratively since March 2013 to develop options for reporting NAEP 12th grade results based upon the preparedness research findings. The options ranged from merely providing information about the 12th grade preparedness research and findings to reporting 12th grade results using statements (inferences) about 12th grade students' academic preparedness.

After the May 2013 Board meeting, at which the Board reviewed the draft validity argument, the two staffs met and agreed that the next step should be to use the guidance from the Board discussion on the validity argument and prepare a prototype chapter for the report. This would provide something specific and concrete as a basis for further Board discussion.

The Board staff drew two main conclusions from the Board discussion in May about the validity argument:

- While finding the validity argument supportive, the Board wanted to consider the independent technical reviews that were to be presented at the August 2013 meeting to inform its decision making.
- The Board found the inferences that were being proposed to be “not quite there yet.”

The inference proposed in May was of the form “12th grade students scoring at or above Proficient are likely to be academically prepared…” Because “likely” was not quantitatively defined, the Board found this formulation ambiguous and potentially confusing to the public. During the discussion, Board member Andrew Ho said he was proposing a solution that he would share with staff. Mr. Ho proposed an inference of the general form as follows:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness, the percentage of students scoring at or above Proficient on Grade 12 NAEP is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities that would make them academically prepared for college.
Mr. Ho’s formulation for the preparedness inference was shared with Michael Kane, who is advising Board staff on the validity argument. Mr. Kane supported using this formulation in place of the one originally proposed and suggested adding “or reasonable” after “plausible.” Board staff revised the validity argument accordingly and it was this formulation that was considered by the independent technical reviewers of the validity argument.

**Question for Board Consideration:**

With the understanding that additional work will be required in collaboration with NCES, along with additional guidance from the Board, is the general approach exemplified in the prototype chapter (Attachment B-1) an acceptable basis for moving forward with reporting on academic preparedness for college as a part of the reporting of the NAEP 12th grade reading and mathematics assessment results for 2013?
Towards NAEP as an Indicator of Academic Preparedness for College and Job Training
Ray Fields July 18, 2013

For over a decade, the National Assessment Governing Board has been conducting research to enable 12th grade NAEP to serve as an indicator of academic preparedness for college and job training. This chapter provides the rationale for pursuing this goal; the research results from studies conducted in connection with the 2009 administration of 12th grade NAEP; and the implications for NAEP 12th grade reporting.

INTRODUCTION
Indicators of many kinds are used to monitor critical aspects of national life and inform public policy. These include economic indicators (e.g., gross domestic product), health indicators (e.g., cancer rates), and demographic indicators (e.g., population trends by race/ethnicity and gender).

NAEP serves the public as a national and state indicator of education achievement at the elementary and secondary levels. NAEP monitors student achievement at key points in the elementary/secondary progression: grades 4, 8, and 12.

According to the National Assessment Governing Board, the 4th grade is the point at which the foundations for further learning are expected to be in place (e.g., when “learning to read” becomes “reading to learn”)

The 8th grade is the typical transition point to high school.

The 12th grade is the end of the K-12 education experience, the transition point for most students to postsecondary education, training, the military, and the work force. (Draft Policy Statement on NAEP).

**NAEP is the only source of nationally representative 12th grade student achievement results.** State tests of academic achievement are usually administered before 12th grade and are quite different across the country. Likewise, college admission tests like the ACT and SAT are generally taken before 12th grade by a self-selected sample and therefore, are not representative of all 12th graders.

Consequently, NAEP is uniquely positioned to serve as an indicator of academic preparedness for college and job training at grade 12—the point that represents the end of mandatory schooling for most students and the start of postsecondary education and training for adult pursuits.

A wide array of state and national leaders has embraced the goal that 12th grade students graduate “college and career ready.” These include the leadership and members of the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), the Business Roundtable (BRT), the U.S. Chamber of Commerce (the Chamber), a task force on education reform of the Council on Foreign Relations, and state and national political leaders. (Fields and Parsad).

**NAEP and ACADEMIC PREPAREDNESS**
The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would afford an invaluable public service: providing an indicator of the human capital potential of today’s and future generations of the nation’s population.
The Board began this initiative in 2004, after receiving recommendations from a distinguished blue-ribbon panel that had examined whether NAEP should continue assessing at the 12th grade.

The panel stated that “America needs to know how well prepared its high school seniors are...[only NAEP] can provide this information...and it is necessary for our nation’s well-being that it be provided.” The panel recommended that NAEP continue to assess at grade 12 and that the 12th grade assessment be transformed to measure preparedness for college, job training, and the military. (National Commission on NAEP 12th Grade Assessment and Reporting; p. 2.)

To transform 12th grade NAEP into an indicator of academic preparedness, the Governing Board took several significant steps.

1. The Board determined that measuring academic preparedness for college and job training should be an intended purpose of 12th grade NAEP.

2. The Board contracted with Achieve, Inc., in 2005 to review the NAEP 12th grade reading and mathematics assessment frameworks and identify where changes, if any, would be needed. Modest changes were recommended.

3. Accordingly, the Board made changes to the frameworks to be used for the administrations of the 12th grade assessments, scheduled for 2009 and 2013.

4. In 2006, the Governing Board assembled a team of noted psychometricians, industrial/organizational psychologists, and K-12 and postsecondary researchers to serve as a technical panel, advising on validity research to conduct.

5. In 2008, the technical panel recommended a comprehensive program of research. The validity of statements about academic preparedness in NAEP reports would be affected by the degree to which the results were mutually confirming.

Figure 1 presents a model of the research program, with five types of research displayed, the interrelationships that would be examined, and the potential meaning of the research results in terms of the NAEP score scale.

Figure 1 about here (see page 8)

6. The Governing Board began contracting for the research studies in 2008, in connection with the 2009 administration of the 12th grade reading and mathematics assessments. More than 30 research studies were completed during the period 2009-2012.

The Research Findings
The research findings were consistent across studies and across years. For example, the content of the 12th grade NAEP reading and mathematics assessments was found to be similar to widely recognized tests used for college admission and placement (see http://www.nagb.org/what-we-do/preparedness-research/types-of-research/content-alignment.html).

Performance by the same students on NAEP and the SAT mathematics and reading tests was correlated at 0.91 and 0.74, respectively.
Statistical linking studies examining performance on NAEP and the college admission tests found that the college readiness benchmarks set for the ACT and SAT reading and mathematics were in a range around the Proficient achievement levels on the 12th grade NAEP reading and mathematics assessments. For example, the average NAEP reading score of students scoring at the SAT benchmark was 301, not significantly different from the cut-score for Proficient of 302 (see Fig. 2 and 3).

A longitudinal study followed a representative sample of Florida 12th grade NAEP test-takers into the state’s public colleges (see Fig. 2 and 3). The longitudinal study permitted an analysis of performance on NAEP and actual student outcomes. In the first year of this study, an analysis was conducted of performance on NAEP and (1) enrollment in regular versus remedial courses, and (2) first year overall college grade point average (GPA). As with the other statistical studies, the average NAEP score of the students who were not placed into remedial courses or who had a first year college GPA of B- or better was in a range around the 12th grade reading and mathematics Proficient achievement levels.

Results from the more than 30 studies were used to develop a validity argument to support proposed inferences (claims) about academic preparedness for college in relation to student performance on 12th grade NAEP. The validity argument was reviewed by two independent technical reviewers. The technical reviewers concluded that the validity argument supports the proposed inferences.

Although the research results support inferences about NAEP performance and academic preparedness for college, the research results to date do not support inferences about NAEP performance and academic preparedness for job training.

A second phase of NAEP preparedness research began in 2013 and is expected to be completed in time for reporting 12th grade results in 2015. The second phase of research results will be examined to determine the degree to which they confirm existing results.

### A TRANSITION TO REPORTING ON ACADEMIC PREPAREDNESS

The reporting of the 12th grade results for 2013 represents a transition point for NAEP.

The interpretations of the 2013 NAEP 12th grade reading and mathematics results related to academic preparedness for college set forth in this report are considered foundational and subject to adjustment in the future.

These interpretations are included in this report because the independent technical reviewers found them to be technically defensible, but more importantly, to promote public discussion about their meaningfulness and utility.

### The Context for Academic Preparedness for College

In the United States in 2013, there is no single, agreed upon definition of “academic preparedness for college” used by colleges for admission and placement (Fields and Parsad). Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.

[http://www.nagb.org/what-we-do/preparedness-research.html]
In this complex mix are 2-year institutions, 4-year public and private institutions with a wide range of selectivity, and proprietary schools. Institutions range from highly selective (i.e., with admission criteria including very high grade point averages, successful completion of rigorous high school coursework and very high SAT and/or ACT scores) to open admission (i.e., all applicants are admitted).

Even within institutions, requirements may vary across majors or programs of study. For example, the mathematics and science high school coursework and academic achievement needed for acceptance into an engineering program in a postsecondary institution may be more rigorous than the general requirements for admission to the institution or for a degree in elementary education in that institution.

**Defining Academic Preparedness for College**

Given the diversity of postsecondary education institutions, it is essential to provide a reasonable definition of academic preparedness for NAEP reporting. The definition should be relevant to NAEP’s purpose of providing group estimates of achievement. (It is important to note that NAEP does not provide individual student results.) The definition should be meaningful to NAEP’s primary audiences: the general public and national and state policymakers.

The definition proposed in this report is intended to apply to the typical degree-seeking entry-level student at the typical college. For NAEP reporting, “academically prepared for college” refers to the reading and mathematics knowledge and skills needed for placement into entry-level, credit bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

It is important to note the focus on “placement” rather than “admission.” This distinction is made because students who need remedial courses in reading, mathematics or writing may be admitted to college, but not placed into regular, credit-bearing courses. The criterion of importance is qualifying for regular credit-bearing courses, not admission.

The definition is not intended to reflect:
- academic requirements for highly selective postsecondary institutions;
- the additional academic requirements for specific majors or pre-professional programs, such as mathematics, engineering, or medicine; or
- academic requirements applicable to entry into certificate or diploma programs for job training or professional development in postsecondary institutions.

The definition is focused on the first year of college; it does not address college persistence beyond the first year or completion of a degree. The definition will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.

**Proposed Inferences for NAEP Reporting**

The NAEP preparedness research does not affect the NAEP results in any way. The distribution of student achievement is unchanged. That is, the average scores, the percentiles, and the achievement level results are not impacted by the NAEP preparedness research.
The independent technical reviewers confirmed that the research findings support inferences about performance on NAEP 12th grade results in reading and mathematics in relation to academic preparedness for college.

**Proposed Inferences**

In the NAEP/SAT linking study for reading (Figure 2), the average NAEP score for 12th grade students scoring at the SAT college readiness benchmark for critical reading is 301, not significantly different from the Proficient cut-score of 302. The results from the Florida longitudinal study are confirmatory.

These data, together with the content analyses that found NAEP reading content to be similar to college admission and placement tests, support the inference for reading that

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

- **the percentage of students scoring at or above a score of 302 (Proficient) on Grade 12 NAEP in reading is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.**

In 2013, XX% of 12th graders nationally scored at or above 302 (Proficient) in reading.

The study results support these inferences. However, there will be students scoring at or above Proficient who are not academically prepared and students scoring below Proficient who are academically prepared (i.e., false positives and false negatives). This will be true for any assessment program that sets cut-scores for a similar purpose.

In the NAEP/SAT linking study for mathematics (Figure 3), the average NAEP score for 12th grade students scoring at the SAT college readiness benchmark for mathematics is 163, lower than and significantly different from the Proficient cut-score of 176. The results from the High School Transcript Study and the Florida longitudinal study are confirmatory.

These data, together with the content analyses that found NAEP mathematics content to be similar to college admission and placement tests, support the inference for reading that

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships
between NAEP scores and NAEP content to other relevant measures of college academic preparedness, the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

In 2013, XX% of 12th graders nationally scored at or above 163 in mathematics.

To consider the plausibility of these estimates, comparisons can be made with the percentages of students who met the ACT or SAT college readiness benchmarks.

Information is available about students who were seniors in 2009 (ACT) and in 2010 (SAT). Thus, the ACT data are for the same student cohort as the NAEP data, but the SAT data are for a cohort that followed one year later.

It also must be noted that, unlike the NAEP results, neither the ACT nor the SAT results represent all 12th graders. Further, there is overlap among ACT and SAT test-takers, with about 20% estimated to take both tests.

Assuming that a substantial portion of students who do not take either test are not academically prepared for college, it is not inconsistent that the NAEP percentages are lower than those for the respective college readiness benchmarks.

| Percentages* Scoring at/above ACT and SAT College Readiness Benchmarks and at/above Proficient in Reading on NAEP and at/above 163 in Mathematics on NAEP |
|-------------------------------------------------|-------------------------------------------------|
| | Reading | Mathematics |
| ACT (2009) | 53 | 42 |
| SAT (2010) | 50 | 54 |
| NAEP (2009) | 38 | 40 |

* About 48% of 12th graders took the ACT or SAT. NAEP represents 100% of 12th graders.

Limitations on Interpretation and Other Caveats

False Negatives and False Positives
Some proportion of 12th grade students scoring below Proficient on the 12th grade NAEP Reading or below a score of 163 on the Mathematics Assessment are

- likely to be academically prepared for college
- not likely to need remedial/developmental courses in reading or mathematics in college,

but with a lower probability than those at or above Proficient in reading or 163 in mathematics.

In addition, some proportion of 12th grade students scoring at or above Proficient on the 12th grade NAEP Reading or 163 on the Mathematics Assessment may not

- be academically prepared for college
- need remedial/developmental courses in reading or mathematics in college.

Not a Preparedness Standard
The proposed inferences are not intended to represent or be used as standards for minimal academic preparedness for college. The proposed inferences are intended solely to add meaning to interpretations of the 12th
grade NAEP reading and mathematics results in NAEP reports.

GPA of B- or Better
The variable “first-year GPA of B- or better” was selected because of its use as a research-based criterion in defining college readiness benchmarks developed for the SAT by the College Board. The College Board had agreed to partner with the Governing Board in a study linking performance on 12th grade NAEP with the SAT. Another leader in college testing programs, ACT, Inc. has developed similar benchmarks for its college admission assessments using a similar criterion and similar methodology. Because they are based on credible research related to college outcomes, and because performance on the respective tests could be linked to performance on NAEP, the college readiness benchmarks used by these testing programs were relevant, useful points of reference for the NAEP preparedness research.

The College Board has set a score of 500 on the SAT Mathematics and Critical Reading tests as its college readiness benchmarks in those areas. Based on its research, the College Board has determined that the score of 500 predicts, with a probability of .65, attainment of a first-year overall GPA of B- or higher. Similarly, the ACT college readiness benchmarks are based on research indicating a .50 probability of attaining first-year grades in relevant courses (e.g., college algebra and courses requiring college level reading) of B or better and .75 probability of C or better.

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

Data Limitations
The NAEP preparedness research studies are comprehensive and the results consistent and mutually confirming, but, for reading the statistical studies are limited to one year for data at the national level and to one state-based longitudinal study. For mathematics, there are two separate years of data at the national level and one state-based longitudinal study. Therefore, more evidence exists to support the plausibility of inferences related to mathematics than to reading.

Preparedness for Job Training
The completed research with respect to academic preparedness for job training does not support conclusions relative to the NAEP scale. Plans for future research will be reviewed by the Governing Board.

Conclusion
The independent technical reviewers found the Governing Board’s preparedness research to be methodical, rigorous, and comprehensive. They concluded that the research findings support the use of the proposed inferences in NAEP reports about 12th graders’ academic preparedness for college.

The interpretations of NAEP results in relation to academic preparedness for college are being reported on a preliminary basis. They are provided to help foster public
understanding and policy discussions about defining, measuring, validating and reporting on academic preparedness for college by NAEP and more broadly.

Including these inferences in NAEP 12th grade reports is intended to add meaning to the interpretation of the NAEP 12th grade results. However, the potential for misinterpretation exists. For these reasons, the section above on limitations on interpretation and other caveats is included in this chapter.

The Governing Board will monitor the use of these inferences as well as unintended consequences arising from their use as a part of the next phase of the preparedness research.

The next phase of the preparedness research is being conducted in connection with the NAEP reading and mathematics assessments administered in 2013. The research results will be used as additional validity evidence in relation to NAEP reporting on 12th grade academic preparedness.

Figure 1. Model of the Preparedness Research Program
Figure 2.

NAEP 12th-Grade Preparedness Research: Reading
Average Scores and Inter-quartile Ranges for Selected Variables, SAT and ACT College Readiness Benchmarks From the 2009 NAEP/SAT Linking Study and 2009 Florida Longitudinal Study

Figure 3.

NAEP 12th-Grade Preparedness Research: Mathematics
Average Scores and Inter-quartile Ranges for Selected Variables, SAT and ACT College Readiness Benchmarks From the 2009 NAEP/SAT Linking Study, 2005 High School Transcript Study, 2009 High School Transcript Study, and 2009 Florida Longitudinal Study
NAEP 12th Grade Reading and Mathematics Report Card: DRAFT Chapter “X”

References to be added.
Review and Comment on

*Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College*

Prepared for:

National Assessment Governing Board

Prepared by:

Gregory J. Cizek, PhD
Professor of Educational Measurement and Evaluation
University of North Carolina at Chapel Hill
cizek@unc.edu

June 30, 2013
Review and Comment on

Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College

Introduction

The National Assessment Governing Board (NAGB) sought input on the constellation of logical and empirical evidence it has amassed in support of certain claims centering on how scores on the 12th Grade National Assessment of Educational Progress (NAEP) might be interpreted with respect to college preparedness. The logic underlying those claims and the logical and empirical support for the claims can be referred to as the validity argument.

According to Kane (2013):

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (p. 1)

The remainder of this paper presents the preparedness score interpretation claims proposed for the 12th grade NAEP scores and an overall an evaluation of the plausibility of those claims.

To produce this evaluation, I relied primarily on two documents that presented the NAEP preparedness validity argument and evidence (Fields, 2013a, 2013b). A draft response to Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College (Fields,
2013a) was submitted to the National Assessment Governing Board on May 29, 2013 (Cizek, 2013). This paper is a response to a revision of *Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College* (Fields, 2013b)

**The Proposed Interpretations and Claims**

The proposed score interpretations related to college preparedness for NAEP Reading and Mathematics are the following:

**READING** – "The percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college."

**MATHEMATICS** – "The percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college." (Fields, 2013b, p. 8)

The proposed interpretations are grounded in four claims (taken from Fields, 2013b):

1. The 12th grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

2. Performance on 12th grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12\textsuperscript{th} grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

**Evaluation of Validity Evidence in Support of the Proposed Interpretations**

Overall, my review and analysis leads me to conclude that the logical and empirical evidence amassed provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear and coherent. The proposed interpretations are warranted in two ways: 1) by the accumulation of confirming evidence that is uniformly in the direction that would be hypothesized by the proposed interpretations; and 2) by the paucity of disconfirming evidence. On this point, it is noteworthy that the present validation effort appeared to be searching, objective, and contemplated the potential for disconfirming evidence.

It is my opinion, based on the evidence provided, that future NAEP reporting can provide reasonably confident and accurate indications of college preparedness in Reading and Mathematics.

It should be recognized, of course, that validation efforts typically should not be considered final or complete at any given juncture (see Cizek, 2012). Additional data can be gathered; additional experience with the test is gained; theory related to (in this case) college preparedness evolves; and new relationships among variables can be explored. The following three recommendations suggest additional validation strategies or evidential sources that may have the potential to strengthen warrants for the intended preparedness score interpretations:

1) To enhance the clarity of the proposed interpretations, I offer the following recommendation: *NAGB should consider making the score interpretations parallel by specifying the NAEP scale score associated with preparedness in Reading.*
As currently worded, a defensible and specific scale score associated with preparedness is offered for NAEP Mathematics score interpretations; however, the interpretation for Reading is phrased as an achievement level: “The percentage of students in the 12th grade NAEP distribution at or above (Proficient for reading and a score of 163 for mathematics) is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in (reading or mathematics) that would make them academically prepared for college.”

The lack of parallelism in construction seems awkward, unnecessary, and potentially confusing to readers and users of this information. I recommend expressing both the Reading and Mathematics interpretations as NAEP scale scores, with elaboration as achievement levels if desired. An example of a slightly reworded interpretation along these lines would be:

“The percentage of students in the 12th grade NAEP distribution at or above a scaled score of XXX (Proficient) in Reading and a score of 163 in Mathematics is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in those subjects that would make them academically prepared for college.”

2) To enhance the coherence of the proposed interpretations, I offer the following recommendation: NAGB should consider conducting additional research into the content coverage of the NAEP and the alignment of NAEP with traditional college admissions measures.
In its present form, it is argued that, in essence, the content of NAEP assessments in Reading and Mathematics covers everything that traditional college admissions measures (e.g., ACT, SAT, etc.) do, but also more. It is claimed that NAEP content coverage is "broader." The Venn diagram below illustrates this claim:\(^1\)

![Figure 1](image)

**Figure 1**

*Hypothetical content coverage between NAEP Assessment and College Admissions Assessment*

Figure 1 illustrates (ignoring the relative size of the circles) the claim that NAEP is somewhat of an umbrella assessment in terms of content coverage compared to the traditional college admissions measures on which alignment research has already been conducted. However, it is not clear that the fact that an umbrella relationship exists unequivocally supports the claim that NAEP assessments capture the same things about college preparedness as the college admissions tests or, importantly, that conclusions based on such alignment can unambiguously be made with respect to preparedness. For example, it would be theoretically possible for an examinee could score "Proficient" on

---

\(^1\) The Venn diagram and the reference to the SAT are presented only illustrate content relationships between content coverage on assessments. The diagram is not intended to represent the actual proportional content coverage between NAEP and college admissions assessments, nor that of the SAT in particular.
NAEP Reading (and be deemed prepared for college) by getting very little of the "SAT-like" content correct on NAEP (that is, content deemed necessary for college success) and getting a lot of the “other” NAEP content correct (that is, the additional/broader content that may or may not necessarily be relevant to college preparedness).

3) To enhance the comprehensiveness of the proposed interpretations, I offer the following recommendation: *NAGB should consider conducting additional research into the predictive validity of the NAEP with respect to college success.*

Perhaps the most important variable assessed in the validation of traditional college admissions assessments is the ultimate criterion of college success—typically operationalized as first year GPA, persistence, or some other variable. Although the validity evidence gathered so far links NAEP scores to scores on other measures that are, in turn, linked to college success, the present validity case for NAEP preparedness does not do so directly. For the future, independent evaluations of direct evidence regarding the extent to which NAEP preparedness scores are associated with college criterion outcomes would substantially bolster the evidence in support of the intended score interpretations.

**Conclusion**

The logical and empirical evidence gathered to date provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear, coherent, and comprehensive. Recommendations were presented for future strategies to strengthen the validity
case. Nonetheless, based on the empirical evidence and logical rationales to date, there appear to be strong warrants for the intended interpretations regarding NAEP reporting and indications of college preparedness in Reading and Mathematics.

References


Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use, Psychological Methods, 17(1), 31-43.


Beginning in March, 2004, the National Assessment Governing Board (NAGB) began work to support the use of the 12th grade National Assessment of Educational Progress (NAEP) as a measure of “preparedness” of students for academic work at the college level. There are many challenges to this work, but one of the most important is to show that there is validity evidence to support the inference that students who are estimated to be above a specified level on the NAEP reporting score scale have the skills and knowledge to profit from credit-bearing, first-year college level coursework.

During the nine year period of this effort, the thinking about the way that validity evidence is collected and reported has had some significant changes. Particularly over the last few years, the work of Michael Kane (e.g., Kane, 2013) has provided guidance about how to present validity evidence for the interpretation of the results of an academic test in the form of what is now called a “validity argument.” The document that I reviewed was one of the first that I have seen that takes this approach to heart and makes a highly credible effort to apply this perspective on validation. In one sense, this is not surprising because work on NAEP has tended to be at the forefront of innovative psychometrics, be it on the use of item response theory procedures or standard setting. In another sense, it is surprising that NAGB has adopted this approach because there are few practical models for the creation of a validity argument. Even though there may have been some risk in being among the first to report support for an inference using the validity argument, this document is quite successful at providing a well supported validity argument. It gives other testing programs a very nice model for future reports on the validation of inferences from test scores.

My general view is that this document presents solid support for the inference that the proportion of the examinee population that is estimated to be above the specified cut score on the NAEP reporting score scale meets the definition of “preparedness for credit-bearing, first-year college coursework.” The evidence that was collected to support the inference is quite extensive and the connection of the evidence to the argument is logical and compelling. There are also appropriate cautions about over interpretation of results. It is very nice to see the areas of weakness in the supporting documents as well as the strengths. This adds credibility to the conclusions from the validity argument. This is not to say that the argument could not be tightened and elaborated, but this is an impressive example of a validity argument for a complex inference from a complex assessment.
A More Detailed Analysis

Although I have a very positive reaction to the report, it is important to probe the specifics of the argument and the claims being made. This may be interpreted as a desire for even more detail than is given in the report, but there is always a need for balance between detail and clear communication. The report is already long and detailed. I am reluctant to suggest adding more to it. But I do want to highlight some specific issues about some of the assumptions and claims in the argument.

The following statement is the basic inference that is the focus of the argument.

“The percentage of students in the NAEP distribution at or above a particular score level in reading or mathematics on 12th grade NAEP is a plausible, or reasonable, estimate of the percentage of 12th grade students who are academically prepared for college.” (P. 6)

This statement is very rich in meaning. To fully understand it, some background information is assumed to be known by the reader. Some of this background is listed here, but the list may not be comprehensive.

1. NAEP produces an accurate representation of the distribution of achievement of students in the areas of reading and mathematics.
2. The estimate of the proportion of students above a cut score on the NAEP reporting score scale is fairly accurate.
3. Students who are estimated to be above the specified cut score are likely to have high school grades and college admissions test scores that will make them eligible for admission to college.
4. Those students who are eligible for admission attend college and enroll in entry-level, credit-bearing courses.
5. The skills and knowledge in reading and mathematics are prerequisite to learning the content presented in the entry-level, credit-bearing courses.

The first two entries in the list are well supported by the technical documentation for NAEP. There are many years of research studies and analyses that show the technical quality of the assessment program. The last three of the entries in the list are more difficult to support because NAEP does not provide accurate student level scores and the individual students who participate are usually not identified so their academic history following the NAEP administration cannot be recorded. It is here that the special studies and data collections that have been done by NAGB are important to fill in links of the validity argument.

A Slight Variation on the Validity Argument

During the process of reviewing the report on the validity argument, I took notes on component parts of the argument. In some cases, the purpose of the notes was to highlight assumptions that were not explicitly stated. In other cases, the purpose was to elaborate on a step in the validity argument. A summary of these notes in the form of a slightly different validity argument than the one given in the report is given below. This is not meant to imply a problem
with the validity argument in the NAGB report, but rather to add some commentary on that argument.

1. There is a body of knowledge and skills that is taught at the secondary school level that is prerequisite to gaining admission into entry-level, credit-bearing courses at colleges and universities.
   a. There seems to be strong evidence for this from the America Diploma Project and the analysis of the admission and placement tests.
   b. It might be helpful to think of this in terms of a Venn diagram that shows the intersection and union of the content descriptions from all of these different sources. The argument should be made that NAEP is based on a reasonable sampling of content from the intersection or the union.

2. College admissions test scores and high school transcripts provide information about the prerequisite knowledge and skills and these are used to make decisions about admissions to the entry-level courses.
   a. This is easy to document, but it is not explicitly stated it in the argument. Of course, different institutions use the information in different ways.

3. The knowledge and skills reflected in college admissions tests and high school transcripts that are prerequisite to the entry-level college courses can be described in some detail to allow the design of a test to assess the knowledge and skills.
   a. This is clearly supported by the information from the studies.
   b. It would be useful to have a summary description of the common components from all of the parts.

4. NAEP assessments provide information about student acquisition of the knowledge and skills described above.
   a. This is the main thrust of all of the content analysis.
   b. The argument is compelling, but it would be helpful to have a general content description that is the result of all of the content analysis.

5. There is a threshold value for the knowledge and skills defined above. If students do not meet this threshold, they will not be ready to take the entry level courses.
   a. The comparative data make a good argument for the existence of the cut score.

6. A cut score on NAEP is consistent with the threshold.
   a. There is a good process for identifying a reasonable cut score on NAEP to correspond to #5.
   b. The combination of information from different tests results in strong support for parts of the argument.
7. The proportion of students estimated to be above the cut score on NAEP gives a good estimate of the proportion who exceed the threshold for admission into entry level courses.
   a. This is well supported by the statistical analysis procedures if the argument for an appropriate cut score is supported. In this case, there is reasonable support for the cut score from the connection to placement and admissions tests.

From this argument, I believe that the following inference from NAEP reported results is supported: The proportion of students estimated to be above the specified cut score on the NAEP reporting score scale is a reasonable estimate of the proportion of students who have the prerequisite knowledge and skills in mathematics and reading to profit from entry-level, credit-bearing college courses.

Reference

Draft Validity Argument for
NAEP Reporting on 12th Grade Academic Preparedness for College
Ray Fields – July 7, 2013

Introduction

Rationale for NAEP Reporting on 12th Grade Academic Preparedness
The National Assessment Governing Board is conducting a program of research to determine the feasibility of the National Assessment of Educational Progress (NAEP) reporting on the academic preparedness of U.S. 12th grade students, in reading and mathematics, for college and job training.

Since 1969, NAEP has reported to the public on the status and progress of student achievement in a wide range of key subjects at grades 4, 8, and 12. NAEP provides national and state-representative results, results for twenty-one urban districts, and results by subgroups of students (e.g., by race/ethnicity, gender, and for students with disabilities and English language learners). NAEP, by law, does not provide individual student results.

The Governing Board’s initiative on 12th grade academic preparedness began in March 2004, with the report of a blue-ribbon panel. The panel was composed of K-12 education leaders—the “producers” of high school graduates—and leaders in business, postsecondary education, and the military—the “consumers” of high school graduates.

The panel members recognized the importance of 12th grade as the gateway to postsecondary education and training, and viewed NAEP as a “truth teller” about student achievement. These distinguished state and national leaders recommended unanimously that “NAEP should report 12th grade students’ readiness for college-credit coursework, training for employment, and entrance into the military.” (National Commission on NAEP 12th Grade Assessment and Reporting; p. 6.). They stated that “America needs to know how well prepared its high school seniors are… [only NAEP] can provide this information…and it is necessary for our nation’s well-being that it be provided.” (Ibid. p. 2.).

The Governing Board approved this recommendation, with a minor modification. The term “readiness” was changed to “academic preparedness” and “entrance into the military” was subsumed by “job training.”

“Readiness” was changed to “academic preparedness” because “readiness” is broadly understood to include both academic preparedness and other characteristics needed for success in postsecondary education and training, such as habits of mind, time management, and persistence (Conley). NAEP does not purport to measure such characteristics. Rather, NAEP is designed to measure academic knowledge and skills.

1 The blue-ribbon panel was known officially as the National Commission on NAEP 12th Grade Assessment and Reporting.
“Entrance into the military” was subsumed by “job training” with the intention of identifying occupations with civilian and military counterparts and utilizing the military’s experience as the world’s largest occupational training organization and its extensive research on the relationship between performance on the Armed Service Vocational Aptitude Battery (ASVAB) and job training outcomes.

The Governing Board approved the 12th grade academic preparedness initiative because it believes that the academic preparation of high school students for postsecondary education and training is important to the nation’s economic well-being, national security, and democratic foundations (see Governing Board resolution of May 21, 2005 at http://www.nagb.org/content/nagb/assets/documents/policies/resolution-on-preparedness.pdf).

Indicators of many kinds are used to monitor critical aspects of national life and inform public policy. These include economic indicators (e.g., gross domestic product), health indicators (e.g., cancer rates), and demographic indicators (e.g., population trends by race/ethnicity and gender). The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would serve as a valuable indicator of the human capital potential of rising generations of citizens, a nation’s greatest resource.

The Governing Board is not alone in recognizing the importance of 12th grade academic preparedness for the nation. A wide array of state and national leaders has embraced the goal that 12th grade students graduate “college and career ready.” These include the leadership and members of the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), the Business Roundtable (BRT), the U.S. Chamber of Commerce (the Chamber), the Council on Foreign Relations, and the Obama Administration. The reason for this attention to 12th grade academic preparedness is well summarized by a statement of the Business Coalition for Student Achievement, an organization coordinated by BRT and the Chamber:

“Ensuring that all students graduate academically prepared for college, citizenship and the 21st century workplace…is necessary to provide a strong foundation for both U.S. competitiveness and for individuals to succeed in our rapidly changing world.”

The NGA and CCSSO have collaborated to develop Common Core State Standards (CCSS) for mathematics and English language arts. These standards are aimed at fostering college and career readiness by the end of high school. The CCSS have been adopted formally by 45 states, several territories and the Department of Defense Education Activity. Viewing the need for rigor in education standards and outcomes through the lens of national security, a similar conclusion was made in the report of the Independent Task Force on U.S. Education Reform and National Security of the Council on Foreign Relations. The Task Force was co-chaired by former New York City School Chancellor Joel Klein and Former Secretary of State Condoleezza Rice. The Obama administration has stated that “educating every American student to graduate from high school prepared for college and for a career is a national imperative.” (Fields and Parsad; pp. 3-4).

Twelfth grade is the end of mandatory schooling for most students and represents the transition point to adult postsecondary pursuits. If it is essential for students to graduate from high school
academically prepared for college and job training, it is essential for the public and policymakers to know the degree to which this is occurring.

A trusted indicator is needed for reporting to the public and policymakers on the status of 12th grade academic preparedness in the U.S., but no such indicator exists. State tests at the high school level are typically administered at 10th and 11th grade. College admission tests, like the SAT and ACT, are administered before the 12th grade, generally to self-selected samples of students.

State tests and college admission tests do not provide a measure of what students know and can do at the very end of K-12 education. Even if these state tests and college admission tests were administered at the 12th grade, they could not be combined to produce nationally representative results.

NAEP is the only source of national and state-representative student achievement data at the 12th grade. As such, NAEP is uniquely positioned to serve as an indicator of 12th grade academic preparedness.

**Defining Academic Preparedness for College**

In the United States in 2013, there is no single, agreed upon definition of “academic preparedness for college” used by colleges for admission and placement. Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.

In this complex mix are 2-year institutions, 4-year public and private institutions with a wide range of selectivity, and proprietary schools. Institutions range from highly selective (i.e., with admission criteria including very high grade point averages, successful completion of rigorous high school coursework and very high SAT and/or ACT scores) to open admission (i.e., all applicants are admitted).

Even within institutions, requirements may vary across majors or programs of study. For example, the mathematics and science high school coursework and academic achievement needed for acceptance into an engineering program in a postsecondary institution may be more rigorous than the general requirements for admission to the institution or for a degree in elementary education in the institution.

In order to design the NAEP 12th grade preparedness research, a working definition of preparedness was needed. The Governing Board’s Technical Panel on 12th Grade Preparedness Research recommended use of the following working definition, which defines academic preparedness for college as

… the academic knowledge and skill levels in reading and mathematics necessary to be qualified for placement…into a credit-bearing entry-level general education course that fulfills requirements toward a two-year transfer degree or four-year undergraduate degree
at a postsecondary institution [without the need for remedial coursework in those subjects]. (National Assessment Governing Board, 2009; p.3.)

This definition was intended to apply to the “typical” college, not to highly selective institutions, and thus, to the vast majority of prospective students, or about 80% of the college freshmen who enrolled in 2-year and 4-year institutions within 2 years following high school graduation (Ross, Kena, Rathbun, KewalRamani, Zhang, Kristapovich, and Manning, p 175). To make this clear, the definition is further elaborated as follows.

Academic preparedness for college refers to the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements (ECNRG) in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

This is consistent with the approach used by the College Board and ACT, Inc. in developing their respective college readiness benchmarks, which are used as external referents in the NAEP 12th grade preparedness research. The ACT benchmarks “represent predictive indicators of success for typical students at typical colleges (Allen and Sconing).” The SAT benchmarks are “an indication of college readiness at a typical college (College Board).”

**Domain Definition for Academic Preparedness for College in Reading and Mathematics**

The working definition described above set the stage for designing the preparedness research studies, but begged a basic question—What are the reading and mathematics knowledge and skills needed to qualify for placement into ECNRG and are they measured by NAEP? This question would be addressed by examining the degree of content match between NAEP and multiple widely accepted external sources that had developed domain definitions for academic preparedness for college in mathematics and reading.

A perfect match between two different sources could not be expected, but a sufficient content match between NAEP and each of a multiple of relevant widely accepted external sources would, collectively, support the inference that the needed knowledge and skills are measured by NAEP. Consequently, the Governing Board identified the following external sources for content comparison with NAEP: The American Diploma Project (ADP) benchmarks for mathematics and English, the ACT College Readiness Standards for Mathematics and Reading, and the ACT, SAT, and ACCUPLACER assessments for reading and mathematics. The results of the content comparison studies between NAEP and these other sources are described in the validity argument below.

**The Central Issue: Validity**

Having made the decision to determine the feasibility of NAEP reporting on 12th grade academic preparedness, the Governing Board recognized that the central concern would be establishing the validity of inferences about 12th grade academic preparedness that are to be made from NAEP scores and used in NAEP reports. The Governing Board would need to ensure that the content of NAEP 12th grade reading and mathematics assessments was appropriate for measuring academic preparedness and that research was conducted to collect evidence by which the validity of
proposed inferences could be evaluated. Finally, a formal validity argument would need to be developed, specifying the proposed inference(s) for NAEP reporting, the underlying assumptions or propositions, and the evidence related to the assumptions or propositions.

Accordingly, the Governing Board

- revised the NAEP assessment frameworks for the 2009 12th grade reading and mathematics with the explicit purpose of measuring academic preparedness for college and job training,
- appointed a special panel of technical experts to recommend a program of research on 12th grade academic preparedness (National Assessment Governing Board, 2009),
- approved and conducted a comprehensive set of preparedness research studies, and
- adopted the model for a validity argument described by Michael Kane (Kane).

The first phase of the Governing Board’s program of preparedness research is completed. The studies were conducted in connection with the 2009 NAEP 12th grade assessments in reading and mathematics. More than 30 studies of five distinct types have been conducted. Study results are available and the complete studies are posted at [http://www.nagb.org/what-we-do/preparedness-research.html](http://www.nagb.org/what-we-do/preparedness-research.html). The National Center for Education Statistics (NCES) has provide additional data drawn from analyses of the 2005 and 2009 High School Transcript Studies conducted in connection with the NAEP 12th grade assessments in those years.

From this research, Governing Board staff developed a proposed interpretation of NAEP performance in reading and mathematics related to 12th grade academic preparedness for college. Following below is the validity evidence for the proposed interpretation, presented in the form of a validity argument. The validity argument provides a statement of the proposed interpretation and the main assumptions inherent in the proposed interpretation in terms of academic preparedness for college. These assumptions are then evaluated using several lines of evidence, which were found to converge for both reading and for mathematics.

### Validity Argument

#### Overview
The National Assessment of Educational Progress (NAEP) program is designed to provide information about student achievement in reading, mathematics and other content areas at the 4th, 8th, and 12th grades. The items for the assessments are developed according to content frameworks and test specifications developed by the National Assessment Governing Board. Scientific sampling procedures are used to produce estimates of score distributions representative of the national population of students at each grade level, as well as estimates representative of public school students in individual states and in 21 urban school districts. The NAEP results do not produce scores for individual students, but rather, group estimates. The NAEP results are reported, based on the estimated score distributions, by average score, percentiles, and in terms of the percentages of students at or above three performance standards used for NAEP reporting, called achievement levels, that are designated Basic, Proficient, and Advanced.
The purpose of the research reported here was to examine whether the interpretation of 12th grade NAEP results in reading and mathematics could be extended to include statements about the percentage of U.S. 12th graders who are academically prepared for college and, if such an interpretation were found to be defensible, to determine the specific statements about academic preparedness that were supportable by the research evidence. The specific statements would be based on the following general definition for academic preparedness, used in relation to the NAEP preparedness research:

the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

The NAEP assessment program is well-established and regularly evaluated, with ample technical documentation of the interpretation of the results at all three grade levels. Therefore, the technical quality, accuracy, and representativeness of the NAEP results in terms of the estimated distributions of U.S. 12th graders on the NAEP scales in reading and mathematics will be taken as a given and as a starting point for additional inferences about the academic preparedness of U.S. 12th graders for college.

In particular, the intent of this validity argument is to examine the evidence in support of statements related to academic preparedness for college for use in reporting NAEP 12th grade results that would have the following general form:

The percentage of students in the NAEP distribution at or above a particular score level in reading or mathematics on 12th grade NAEP is a plausible, or reasonable, estimate of the percentage of 12th grade students who are academically prepared for college.

This interpretation would depend on four prior claims (or assumptions):

1. The 12th grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

2. Performance on 12th grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

The first claim is supported by the combination of the content of the NAEP assessment frameworks and the NAEP test items, the NAEP sampling designs, and the statistical models used to generate estimates of score distributions at each grade level and in each content area. These claims are well-established, documented, and evaluated; therefore, the attention of the validity argument will be directed primarily to the second, third, and fourth claims.

The second claim is supported by a statistical relationship study that examined student performance on the NAEP 12th grade reading and mathematics assessments to performance on the SAT reading and mathematics tests, as well as the respective college readiness benchmarks established by the College Board for these tests, which, in turn, are related to outcomes associated with academic preparedness for college.

The third claim was evaluated with multiple sources of evidence that were highly convergent. These include the SAT/NAEP statistical relationship study, a longitudinal study of Florida 12th grade students, and analyses of the 2005 and 2009 NAEP High School Transcript Studies.

The fourth claim is supported by the fact that the Governing Board reviewed the NAEP 12th grade reading and mathematics frameworks for the purpose of making NAEP a measure of academic preparedness for college; made changes to the frameworks accordingly; and conducted a comprehensive set of content alignment studies to determine the degree of match between NAEP and tests that are used for college admission and placement.

Further, the results from the examination of the NAEP content provide a counter argument to a possible falsifying claim about the positive relationships discussed in the second and third claims. The falsifying claim would be that the positive relationships between NAEP and the other indicators were merely statistical artifacts, due to factors extraneous to academic preparedness for college, akin to finding a high correlation between height and passing rates on a state driving test. The counter argument is that the relationships are meaningful because the NAEP 12th grade reading and mathematics assessments were intentionally designed to measure academic preparedness for college and that the evidence supports the conclusion that the NAEP 12th grade assessments do measure academic preparedness for college.
Proposed Inferences

For reading:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

the percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.

For mathematics:

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

In contrast to the inference for reading, which is set at the Proficient level, the inference for mathematics is set at a score on the NAEP mathematics scale of 163. This score is strongly supported by the consistent research results across years and data sources, but is below and significantly different from the cut-score for the Proficient level for NAEP 12th grade mathematics, which is 176.

The research results for mathematics do support a related inference—that students in the distribution at or above the NAEP Proficient level in mathematics are likely to be academically prepared for college. However, the percentage of such students would be substantially less than the percentage in the distribution at or above 163, and thus, would underestimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.

For these reasons, and to have the proposed inferences for reading and mathematics as parallel as possible, the proposed inference for reading is formulated in relation to the Proficient achievement level and the proposed inference for mathematics is formulated in relation to the NAEP mathematics scale score of 163.

Limitations on Interpretation and Other Caveats

False Negatives and False Positives
Some proportion of 12th grade students scoring below Proficient on the 12th grade NAEP Reading or below a score of 163 on the Mathematics Assessment are
• likely to be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and
• not likely to need remedial/developmental courses in reading or mathematics in college, but with a lower probability than those at or above Proficient in reading or 163 in mathematics.

In addition, some proportion of 12th grade students scoring at or above Proficient on the 12th grade NAEP Reading or 163 on the Mathematics Assessment may not
• be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and
• need remedial/developmental courses in reading or mathematics in college.

Not a Preparedness Standard
The proposed inferences are not intended to represent or be used as standards for minimal academic preparedness for college. The proposed inferences are intended solely to add meaning to interpretations of the 12th grade NAEP reading and mathematics results in NAEP reports.

Academically Prepared for College
The proposed inferences are intended to apply to the typical degree-seeking entry-level college student at the typical college. Thus, “academically prepared for college” refers to the reading and mathematics knowledge and skills needed for placement into ECNRG courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

It is important to note the focus on “placement” rather than “admission.” This distinction is made because students who need remedial courses in reading, mathematics or writing may be admitted to college, but not placed into regular, credit-bearing courses. The criterion of importance is qualifying for regular credit-bearing courses, not admission.

The proposed inferences are not intended to reflect academic requirements for highly selective postsecondary institutions; to the additional academic requirements for specific majors or pre-professional programs, such as mathematics, engineering, or medicine; or to academic requirements applicable to entry into certificate or diploma programs for job training or professional development in postsecondary institutions.

The proposed inferences are focused on the first year of college; they do not support conclusions about college persistence beyond the first year or completion of a degree. The inferences will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.
GPA of B- or Better
The selection of “first-year GPA of B- or better” as a referent was made because of its use as a research-based criterion in defining college readiness benchmarks developed by an acknowledged leader in college testing programs—the College Board. The College Board had agreed to partner with the Governing Board in a study linking performance on 12th grade NAEP with the SAT. Another leader in college testing programs, ACT, Inc. has developed similar benchmarks for its college admission assessments using a similar criterion and similar methodology. Because they are based on credible research related to college outcomes, and because performance on the respective tests could be linked to performance on NAEP, the college readiness benchmarks used by these testing programs were embraced as relevant, useful points of reference for the NAEP preparedness research.

The College Board has set a score of 500 on the SAT Mathematics and Critical Reading tests as its college readiness benchmarks in those areas. Based on its research, the College Board has determined that the score of 500 predicts, with a probability of .65, attainment of a first-year overall GPA of B- or higher. Similarly, the ACT college readiness benchmarks are based on research indicating a .50 probability of attaining first-year grades in relevant courses (e.g., college algebra and courses requiring college level reading) of B or better and .75 probability of C or better.

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

Data Limitations
Although the preparedness research studies are comprehensive and the results consistent and mutually confirming, for reading they are limited to one year for data at the national level and to one state-based longitudinal study. For mathematics, there are two separate years of data at the national level and one state-based longitudinal study. Therefore, more evidence exists to support the plausibility of inferences related to mathematics than to reading.

Preparedness for Job Training
The completed research with respect to academic preparedness for job training does not support conclusions relative to the NAEP scale and will not be addressed at this time.
Discussion of the Claims and Evidence

1. The 12th-grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

The proposed inferences are premised in part on the capability of NAEP to report percentages of students scoring at or above a certain score on the NAEP 12th grade reading and mathematics scales. The technical qualities of the NAEP scales make them well suited to this purpose.

The NAEP sampling, scaling, IRT modeling, and statistical procedures are widely accepted, well documented (for example, see National Center for Education Statistics, pp. 70-71) and have been periodically evaluated over two decades (for example, see complete list of research conducted by the NAEP Validity Studies Panel at http://www.air.org/reports-products/index.cfm?fa=viewContent&content_id=890 and “Evaluation of the National Assessment of Educational Progress: Study Reports” at http://www2.ed.gov/rschstat/eval/other/naep/naep-complete.pdf).

Other than issues relating to the comparability among the state-level NAEP samples of inclusion rates of students with disabilities and students who are English language learners (about which the Governing Board and NAEP have taken and continue to take significant action), there is little dispute about the appropriateness of the NAEP sampling, scaling and statistical procedures for estimating the percentage of students scoring at or above a selected NAEP scale score.

This is relevant because the proposed inferences that are the subject of this validity argument are interpretations to add meaning to the reporting of NAEP 12th grade reading and mathematics results at particular score levels. The percentages of students at or above particular score levels (e.g., the NAEP achievement levels) have been estimated with accuracy and reported regularly, beginning with assessments in 1992. The proposed inference for reading would use the cut-score for 12th grade Proficient as the basis for reporting. The proposed inference for mathematics would use the score of 163 on the NAEP 12th grade scale as the basis for reporting, which is between the Basic and Proficient achievement levels. Clearly, reporting NAEP results using the proposed inferences will not impair the accuracy of the estimates of the percentages of students scoring at or above the identified points on the NAEP score scales.

2. Performance on 12th-grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

In designing the NAEP preparedness research program, the Governing Board determined that it would be essential to examine how performance on NAEP relates to performance on other
measures and outcomes associated with academic preparedness for college. The research program studied the relationship between performance on NAEP and performance on the SAT and ACT college admission tests, including the respective college readiness benchmarks that had been established by these testing programs.

The data sources for the analyses that were conducted are: the NAEP/SAT linking studies (see report at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/SAT-NAEP_Linking_Study.pdf); the Florida longitudinal study (see report at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/Florida_Statistical_Study.pdf); the 2005 and 2009 NAEP High School Transcript Studies; and the Governing Board’s survey of postsecondary education institutions’ use of tests and the cut-scores on those tests for determining whether incoming students need remedial instruction in reading and mathematics (Fields and Parsad).

In addition, the research program examined directly the relationship between performance on NAEP and postsecondary outcomes analyzing data from the Florida longitudinal study.

The results of these studies will be discussed both in this section and the next section of the validity argument. In this section, background is provided on the indicators that were examined and the results of the NAEP/SAT linking study. The NAEP/SAT linking study is discussed in this section because, as the most recent large-scale national study, it serves as a focal point for discussing the results of the other studies. Thus, in section 3, the results of the other statistical linking studies are discussed in relation to the NAEP/SAT linking study.

**Indicators: College Board and ACT College Readiness Benchmarks**

The College Board and ACT, Inc. have established college readiness benchmarks for the SAT and the ACT in a number of subjects tested, including reading and mathematics. The SAT College Readiness Benchmark for critical reading and mathematics is a score of 500 on the respective tests. According to the College Board’s research, a score of 500 predicts, with a .65 probability, a first-year GPA of B- or better. The ACT College Readiness Benchmark for reading is a score of 21. According to ACT’s research, a score of 21 predicts, with a .50 probability, a grade of B or better (or .75 probability of a C or better) in first year courses requiring college reading, such as history and the social sciences. A score of 22 on the ACT mathematics tests predicts a .50 probability of a grade of B or better in a first-year mathematics course, or a .75 probability of a grade of C or better. The College Board research and the ACT research are based on the first-year outcomes of their respective test takers.

**Indicators: First Year GPA of B- or Better and Remedial/non-Remedial Placement**

The Governing Board has a partnership with the state of Florida as a part of the Board’s program of preparedness research. Florida was one of 11 states that volunteered to provide state-
representative samples of 12th grade students for the 2009 NAEP reading and mathematics assessments. Under the partnership, the Florida 12th grade sample is being followed through the postsecondary years via the highly developed Florida longitudinal education data system. For comparability with the SAT College Readiness Benchmarks, the Governing Board analyzed the Florida data to determine the average score and interquartile range for the NAEP test takers with a first year GPA of B- or better. In addition, the Governing Board analyzed the Florida data to determine the average score and interquartile range for the NAEP test takers who were and who were not placed into remedial reading or remedial mathematics in their first year of college.

**Analysis of Results for Mathematics**

The statistical linking study examining performance on the NAEP 12th grade mathematics assessment and performance on the SAT mathematics test yielded a correlation of .91. This high correlation clearly supports inferences about NAEP performance in relation to SAT performance. The study also examined how performance on NAEP relates to the SAT College Readiness Benchmark for mathematics (i.e., a score on the SAT mathematics test of 500). The SAT benchmark provides “an indication of college readiness at a typical college (College Board).” This is consistent with the Governing Board’s definition of academic preparedness cited previously:

> Academic preparedness for college refers to the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

The SAT College Readiness Benchmark for mathematics is relevant to student outcomes in college, for it is “the SAT score associated with a 65 percent probability of earning a first-year GPA of B- (i.e., 2.67) or higher (College Board).” The average NAEP score of students scoring at the College Readiness Benchmark for mathematics was 163 (see Figure 1). As will be demonstrated in the discussion of the third claim, there are additional data corroborating this level of performance on the 12th grade NAEP mathematics assessment to outcomes in college.

**Analysis of Results for Reading**

The statistical linking study examining performance on the NAEP 12th grade reading assessment and the SAT critical reading test resulted in a correlation of .74. Although it may not be high enough to predict the performance of individual students from one test to another (which is not required to support the proposed inference for reading), it is sufficient to support the group-level inferences reported by NAEP.
Performance on NAEP was also examined in relation to the SAT College Readiness Benchmark for critical reading (i.e., a score on the SAT critical reading test of 500). The SAT benchmark provides “an indication of college readiness at a typical college (College Board).” This is consistent with the Governing Board’s definition of academic preparedness discussed in the results for mathematics above.

The SAT College Readiness Benchmark for critical reading is relevant to student outcomes in college, for it is “the SAT score associated with a 65 percent probability of earning a first-year GPA of B- (i.e., 2.67) or higher (College Board).” The average NAEP score of students scoring at the College Readiness Benchmark for reading was 301(see Figure 2). As will be demonstrated in the discussion of the third claim, there are additional data corroborating this level of performance on the 12th grade NAEP reading assessment to outcomes in college.

3. **There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).**

In addition to the NAEP/SAT Linking Studies (NSLS) described above, analyses were conducted using data from several other studies. There was a high degree of convergence found across the studies. The results are described below, first for mathematics and then for reading.

**Analysis of Results for Mathematics**

Companion statistical relationship studies to the NSLS for mathematics examined data from the 2005 and 2009 national NAEP High School Transcript Studies (HSTS) and from a longitudinal study under a partnership with the Florida Department of Education (FLS). In 2009, Florida was one of eleven states that volunteered to participate in 12th grade state NAEP in reading and mathematics. Using the highly developed Florida longitudinal data base, the students in the 12th grade NAEP samples were followed into postsecondary public institutions.

Analyzing data from the transcripts of NAEP test takers, the HSTS examined performance on 12th grade NAEP mathematics in relation to performance in mathematics on the SAT and ACT college admissions tests in 2005 and 2009. The FLS study examined performance on the NAEP 12th grade mathematics assessment in relation to the SAT and ACT college readiness benchmarks, first year overall college GPA, and whether students were placed into non-remedial college courses. The study results are displayed in Figure 1.

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) because it is the most recent of the national studies. The average NAEP score is 163 for students with an SAT score at the College Readiness Benchmark for mathematics of 500.
The other study results are consistently convergent with the NSLS results. The average NAEP mathematics scores for 12th grade students scoring at the SAT College Readiness Benchmark of 500 for mathematics are compared first for the 2005 HSTS and the 2009 NSLS. The average scores are 161 and 163 respectively.

These results are confirmed by the FLS. The average NAEP mathematics score for the 12th grade Florida NAEP test takers who scored at the SAT College Readiness Benchmark of 500 was 160, much like the 2009 NSLS results and the 2005 HSTS results.

As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for mathematics is defined somewhat differently than the SAT College Readiness Benchmark for mathematics. However, it is noteworthy that even with this different definition, the results from the 2005 HSTS, 2009 HSTS, and 2009 FLS analyses for the ACT (169, 166, and 164, respectively) are consistent and very similar to the results for the 2009 NSLS.

To answer the question, "What is the relationship between performance on NAEP and actual student outcomes?", we look to the FLS results. First we examine the average NAEP mathematics score for the 12th grade Florida NAEP test takers who attained a first-year GPA of B- or better. The average NAEP score for these students was 162. This is consistent with the SAT College Readiness Benchmark analyses and further supports the inference that students at
or above 163 on the 12th grade NAEP mathematics scale are likely to be academically prepared and attain a first-year GPA of B- or better. It follows, of course, that students who are academically prepared will not require remedial courses.

Thus, another outcome of interest is placement of entry-level students into remedial college courses versus non-remedial credit-bearing courses. Here again, we look to the FLS as a data source. The average NAEP mathematics score was 165 for the Florida NAEP test-takers not placed into remedial courses, which is consistent with the NSLS score of 163 on the NAEP 12th grade mathematics scale. Furthermore, the average NAEP score of students who were placed into remedial mathematics courses in college was 136, much lower and significantly different from the NSLS score of 163.

The FLS results, together with the SAT and ACT analyses, lend support to the conclusions that students scoring at or above 163 on the 12th grade mathematics scale are likely to be academically prepared for ECRNG college courses and not likely to need remedial courses in mathematics.

These convergent, consistent results across years and across studies support the proposed inference that the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

**Analysis of Results for Reading**

The companion statistical relationship study to the NSLS for reading examined data from a longitudinal study under a partnership with the Florida Department of Education (FLS). In 2009, Florida was one of eleven states that volunteered to participate in 12th grade state NAEP in reading and mathematics. Using the highly developed Florida longitudinal data base, the students in the 12th grade NAEP samples were followed into postsecondary public institutions.

The FLS study examined performance on the NAEP 12th grade reading assessment in relation to the SAT and ACT college readiness benchmarks for reading, first year overall college GPA, and whether students were placed into non-remedial college courses. The study results are displayed in Figure 2.

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) for reading, because it is the most recent of the national studies. The average NAEP score is 301 for students with an SAT score at the College Readiness Benchmark for critical reading of 500. A NAEP score of 301 in 12th grade reading is not significantly different from the cut-score for the 12th grade Proficient achievement level (302).
The FLS results are consistently convergent with the NSLS results. The average NAEP reading score was 299 for the 12th grade Florida NAEP test takers who were not placed into remedial courses in their first year. The average score was 298 for those who had a first year overall GPA of a B- or better. These data, which show the relationship between performance on NAEP and actual student outcomes, provide strong confirmation that students scoring at or above Proficient on the NAEP 12th grade reading assessment are likely to be academically prepared for ECNRG college courses.

As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for reading is defined differently than the SAT College Readiness Benchmark for reading. However, it is noteworthy that even with this different definition, the ACT results from the 2009 FLS analysis are similar to the NSLS analysis and the FLS outcome data.

Taken together, these results support the inference that students scoring at or above Proficient on the NAEP 12th grade reading scale are likely to be academically prepared for ECNRG college courses.

In conclusion, these results suggest that the percentage of students at or above the Proficient level in reading on 12th grade NAEP would provide a plausible (or reasonable) estimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

- NAEP Assessment Frameworks Were Revised to Measure Academic Preparedness

The National Assessment Governing Board intentionally revised the NAEP 12th grade reading and mathematics assessment frameworks with the purpose of measuring academic preparedness for college.

On March 5, 2004, the Governing Board accepted the report of the Commission on NAEP 12th Grade Assessment and Reporting. The Commission recommended that “NAEP should report 12th grade students’ [academic preparedness] for college-credit coursework, training for employment, and entrance into the military.”

For NAEP to report on 12th grade academic preparedness for college, it must measure relevant content at the 12th grade. The content of each assessment is determined by the NAEP assessment frameworks, which the Governing Board is responsible for developing and approving. Accordingly, the Governing Board decided that the extant NAEP frameworks intended for the 2009 for reading and mathematics at the 12th grade would be reviewed. The review would identify changes needed to measure 12th grade academic preparedness for college. Examples of the changes made are described in the next two subsections.

Assessments at the 12th grade in reading and mathematics are conducted at least once every 4 years. In 2004, when the Board decided to proceed with the 12th grade academic preparedness initiative, 2009 was the next assessment year in which the 12th grade reading and mathematics assessments could be affected by framework changes.

In September 2004, the Governing Board contracted with Achieve, Inc. (Achieve) to review the NAEP 12th grade reading and mathematics assessment frameworks and identify where changes, if any, would be needed. Achieve had established the American Diploma Project (ADP) “…to improve postsecondary preparation by aligning high school standards, graduation requirements and assessment and accountability systems with the demands of college and careers (see www.achieve.org/adp-network).” The ADP had conducted research to identify key competencies in English and mathematics needed for high school graduates who aspire to higher education. They refer to these as the “ADP benchmarks.” The type of colleges that were the target for the ADP research was similar to the “typical colleges” in the Governing Board’s research. These were the “two- and four-year colleges and universities in each of the ADP partner states…[that] enroll the vast majority of high school graduates going on to college:

---

\[ \text{(Footnote)} \]

The review also addressed academic preparedness for job training, but that part of the NAEP preparedness initiative is not being addressed in this validity argument.
community colleges, as well as four-year state institutions, but generally not the more highly selective “flagship” campuses.” (Achieve, 2004, p. 107)

The research and expertise of the American Diploma Project was widely accepted and was brought to bear in reviewing the NAEP frameworks for 12th grade reading and mathematics. Achieve convened a panel of nationally recognized experts in reading and a panel of nationally recognized experts in mathematics. The panels were comprised of individuals from the K-12, postsecondary, research, and policy spheres, knowledgeable about academic preparedness for college reading and college mathematics. The panels compared the 12th grade NAEP reading and mathematics frameworks and the ADP benchmarks.

Reading

The Achieve reading panel found considerable similarity between NAEP and the ADP benchmarks for English, although not perfect agreement. This is displayed in the side-by-side chart on pages 30-40 of the Achieve Reading Report (http://www.nagb.org/content/nagb/assets/documents/commission/researchandresources/Achieve%20Reading%20Report.pdf). The English benchmarks have eight major components and objectives under each component. Three of these major components were deemed “Not Applicable” to the reading domain: writing, research, and media.

For almost all of the applicable objectives under the five major components that were applicable to the reading domain, the Achieve reading panel found matches in the NAEP 2009 reading framework. Overall, the panel concluded that “…the 2009 NAEP Reading Framework…was aligned to the ambitious [ADP] benchmarks” (Achieve Reading Report, p. 2).

The reading panel also listed items in the NAEP framework that are not found in the ADP English benchmarks. For example, under Argumentation and Persuasive Text, figurative language and rhetorical structure, including parallel structure and repetition, was present in the NAEP reading framework at grade 12, but not in the ADP benchmarks. Under Poetry, tone, complex symbolism, and extended metaphor and analogy were present in the NAEP reading framework but not the ADP benchmarks. A complete listing of the items in the NAEP framework not present in the ADP benchmarks appears on page 41 of the Achieve Reading Report.

Although the Achieve reading panel concluded that the 12th grade NAEP reading framework for 2009 was aligned with the ADP benchmarks applicable to reading, the panel’s report does include six recommendations. The Governing Board approved these recommendations on February 14, 2005. For example, the Achieve reading panel recommended increasing the percentage of informational text passages from 60% to 70% and to feature additional items that ask students to compare texts. The changes were modest, sufficiently so to permit continuation of the 12th grade trend line from its initiation in 1992.

The NAEP reading framework used for the 2009, 2011, and 2013 assessments contains the following statement
In May 2005, the Governing Board adopted a policy statement regarding NAEP and 12th-grade preparedness. The policy states that NAEP will pursue assessment and reporting on 12th-grade student achievement as it relates to preparedness for post-secondary education and training. This policy resulted from recommendations of the Board’s National Commission on NAEP 12th Grade Assessment and Reporting in March 2004. Subsequent studies and deliberations by the Board took place during 2004 and 2005.

In reading, the Board adopted minor modifications to the 2009 NAEP Reading Framework at grade 12 based on a comprehensive analysis of the framework conducted by Achieve, Inc. The current version of the reading framework incorporates these modifications at grade 12 to enable NAEP to measure and report on preparedness for postsecondary endeavors (National Assessment Governing Board, 2008, Reading Framework, p. v).

Mathematics
The mathematics review began with the 2007 NAEP mathematics framework, which was the most current and included the changes approved for the 2005 12th grade mathematics assessment. The Achieve panel examined the NAEP mathematics framework at the 12th grade in relation to the ADP benchmarks for mathematics. The Achieve panel developed proposed revisions to the assessment objectives for grade 12. While acknowledging differences in language and purpose, the Achieve mathematics panel concluded that the “overall mathematics frameworks of ADP and [12th grade] NAEP are remarkably similar” (see http://www.nagb.org/content/nagb/assets/documents/commission/researchandresources/Achieve-Mathematics-Report.pdf, Achieve Mathematics Report, p.9).

The Governing Board convened a panel of mathematicians and mathematics educators to review and revise the objectives in relation to the objectives for grades 4 and 8. The panel conducted focus groups with various NAEP constituents, using repeated rounds of reviews. The Governing Board approved the final set of grade 12 objectives on August 5, 2006. The changes to the framework were sufficiently modest to permit the continuation of the 12th grade trend line begun with the 2005 12th grade mathematics assessment under the previous 12th grade framework. Like the reading framework, the 2009/2013 mathematics framework for grade 12 states the Board’s intention to measure 12th grade academic preparedness (National Assessment Governing Board, 2008, Mathematics Framework, pp. 2-3).

Conclusion
The Governing Board, by official action, revised the NAEP 12th grade reading and mathematics frameworks with the explicit purpose of measuring 12th grade academic preparedness for college, beginning with the 2009 assessments. Setting forth the measurement purpose and making relevant revisions to the NAEP assessment frameworks are necessary elements of the validity argument; however, they are not sufficient. Evidence must be considered with respect to the alignment of the framework and the test questions administered to the measurement purpose. This will be addressed in the next section.
Examples of Objectives added to the 2009 Grade 12 Mathematics Framework

**Number properties and operations**

b) * Analyze or interpret a proof by mathematical induction of a simple numerical relationship.

**Measurement**

d) Interpret and use the identity \(\sin^2 \theta + \cos^2 \theta = 1\) for angles \(\theta\) between 0° and 90°; recognize this identity as a special representation of the Pythagorean theorem.

e) * Determine the radian measure of an angle and explain how radian measurement is related to a circle of radius 1.

f) * Use trigonometric formulas such as addition and double angle formulas.

g) * Use the law of cosines and the law of sines to find unknown sides and angles of a triangle.

**Geometry**

e) * Use vectors to represent velocity and direction; multiply a vector by a scalar and add vectors both algebraically and graphically.

g) * Graph ellipses and hyperbolas whose axes are parallel to the coordinate axes and demonstrate understanding of the relationship between their standard algebraic form and their graphical characteristics.

h) * Represent situations and solve problems involving polar coordinates.

**Data Analysis, Statistics, and Probability**

c) * Draw inferences from samples, such as estimates of proportions in a population, estimates of population means, or decisions about differences in means for two “treatments.”

e) * Recognize the differences in design and in conclusions between randomized experiments and observational studies.

k) * Use the binomial theorem to solve problems.

e) * Recognize and explain the potential errors caused by extrapolating from data.

**Algebra**

e) Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or trigonometric functions from tables, graphs, or equations.

j) * Given a function, determine its inverse if it exists and explain the contextual meaning of the inverse for a given situation.

h) * Analyze properties of exponential, logarithmic, and rational functions.

g) * Determine the sum of finite and infinite arithmetic and geometric series.
Content Alignment Studies Found Significant Overlap between NAEP and the ACT, SAT and ACCUPLACER

The Governing Board conducted studies to determine the degree of content similarity between NAEP 12\textsuperscript{th} grade reading and mathematics assessments and relevant tests used for college admissions and placement.

The studies had two objectives. The first objective was to determine the degree to which the content of 12\textsuperscript{th} grade NAEP in reading and mathematics covers the reading and mathematics knowledge and skills needed for first year college work. The SAT, ACT, and ACCUPLACER are well-established tests that assess individual students’ reading and mathematics proficiency in relation to college level expectations.

The ACT is developed with the purpose of “…[measuring] as directly as possible the degree to which each student has developed the academic skills and knowledge that are important for success in college…” (ACT Technical Manual, p. 62).

The SAT is developed “to ensure that the topics measured on the SAT…reflect what is being taught in the nation’s high schools and what college professors consider to be required for college success.” (Kim, Wiley, and Packman, p.1)

The ACCUPLACER has the purpose of “…[determining] which course placements are appropriate for [incoming college] students and whether or not remedial work is needed.” (ACCUPLACER, p. A-2)

The SAT, ACT and ACCUPLACER in reading and mathematics are widely used for these purposes by admissions and placement professionals in postsecondary education institutions. These testing programs regularly conduct curriculum surveys, validity studies and other research to support their claims that the content measured is directly related to the reading and mathematics knowledge and skills needed to qualify for entry-level credit-bearing courses (e.g., see the ACT curriculum studies for 2012, 2009, 2005, and 2002 at http://www.act.org/research-policy/national-curriculum-survey/, and the College Board National Curriculum Survey on English and Mathematics at http://research.collegeboard.org/publications/content/2012/05/national-curriculum-survey-english-and-mathematics).

- Therefore, with the assumption that the SAT, ACT, and ACCUPLACER do measure the content needed for college level work, significant content overlap between NAEP and these other assessments would support the conclusion that what NAEP measures covers the knowledge and skills needed by college freshmen to be placed into entry-level credit bearing courses.

The second reason for conducting the content alignment studies was to provide information for interpreting the results of planned statistical linking studies between NAEP and the other tests, which measure academic preparedness for college. The linking studies were designed to examine how performance on NAEP compares with performance on the other tests, with the
purpose of supporting inferences about academic preparedness for college. For NAEP to support inferences about academic preparedness for college based on the linking studies, a sufficient content match would be needed between NAEP and the other tests, not just a statistical relationship.

The Content Alignment Studies: Overview
The Governing Board conducted content alignment studies in reading and mathematics comparing the 2009 12th grade NAEP and the ACT, SAT, and ACCUPLACER reading and mathematics tests. Overall, considerable overlap was found between the ACT and NAEP and the SAT and NAEP, with some differences. NAEP was found to measure much of what is measured on the ACCUPLACER, but the reading and mathematics domains measured by NAEP were much broader than ACCUPLACER. More details are provided in the summaries of the individual studies below.

The general design for the content alignment studies was to compare the 12th grade NAEP frameworks in reading and mathematics with the analogous document for the other test, and then to compare the test items from one test to the framework/analogous document of the other test. The reviews were performed by subject specific (i.e., mathematics, reading) panels, composed of experts in mathematics or reading and English instruction at the high school and college levels.

Alignment studies that compare an assessment to the content standards on which it is based are relatively common and have well-established methodologies. However, this is not true for the types of alignment studies the Governing Board planned to conduct: content alignment studies comparing different assessment programs. Different assessment programs have different purposes, different approaches to describing the domain being measured, and, possibly, different “grain size” in the level of detail in describing the domain.

The Governing Board contracted with Norman Webb, a noted expert in content alignment studies, to prepare a design document for conducting the assessment to assessment alignment studies. The purpose was to put in place a methodology that considered the special challenges of assessment to assessment alignment studies and to foster comparability in the conduct of the studies and the reporting metrics across studies and contractors. The link to the Webb design document is at [http://www.nagb.org/content/nagb/assets/documents/publications/design-document-final.pdf](http://www.nagb.org/content/nagb/assets/documents/publications/design-document-final.pdf).

The Webb design was developed after the ACT alignment studies were completed. It was used in conducting the SAT and ACCUPLACER content alignment studies.

In the following sections are summaries of the content alignment study results, excerpted from the study reports. The results for the three content alignment studies in reading are presented first, followed by the three content alignment studies for mathematics, along with summary discussions for the reading and mathematics results.
The Content Alignment Studies: Reading Results

**Reading: ACT**

The Governing Board contracted with ACT, Inc. to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the ACT reading test. The full report can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf).

The reading panel was composed of 7 members, with expertise in reading and/or English instruction at the high school and college levels. The panel was about evenly divided in terms of prior familiarity with either the ACT or NAEP reading domains.

The panel found considerable similarity in the content of the NAEP 12th grade reading assessment and the ACT. For example, the NAEP 12th grade reading framework was compared to the ACT reading domain and the ACT College Readiness Standards for reading. The ACT College Readiness Standards (CRS) are descriptions of the content (i.e., the knowledge and skills) measured by the ACT reading test in score bands along the ACT 1-36 point scale from 13-36 (see [http://www.act.org/standard/planact/reading/](http://www.act.org/standard/planact/reading/)). The panel concluded that

“All of the skills highlighted in the ACT [reading] domain and in the [ACT] College Readiness Standards [for reading] were identified within the NAEP Reading framework. In performing the comparison in the other direction—NAEP to ACT—it was the sense of the panel that the ACT measured primarily those skills that NAEP identifies as Locate/Recall and Integrate/Interpret skills, those that pertain primarily to finding explicit information in text (what the ACT would call Referring skills) and to making inferences, drawing conclusions, and making generalizations from information within text (what the ACT would call Reasoning skills). The panel saw less evidence of the higher-level analytical and evaluative Critique/Evaluate skills in the ACT domain, and attributed that to the multiple-choice format of the ACT [whereas NAEP includes constructed response items as well as multiple choice]. Another difference is that NAEP includes items and texts measuring how well an examinee can apply reading skills across texts, whereas the paired passage format is not a feature of the ACT. So, while the NAEP Reading framework and the ACT Reading domain, test specifications, and College Readiness Standards share similarities, important differences in what and how the assessments measure suggest caution when drawing comparisons between the assessments.” (p.17)

The reading panel also conducted an item classification study, in which the NAEP 12th grade reading items were classified in relation to the ACT College Readiness Standards for Reading.

“A total of 152 Reading items (comprising 17 blocks) were classified in [the reading] study. Of these, 97 were multiple-choice (MC). Nine were dichotomously-scored (“incorrect” or “correct”) short constructed-response (DSCR) items. Thirty-three were polytomously-scored short constructed-response (PSCR) items, each scored using a three-point scoring rubric. Thirteen were extended constructed-response (ECR) items,
each scored using a four-point rubric. Each DSCR had one creditable score category, each PSCR had two, and each ECR had three. Each Reading panelist, therefore, assigned a total of 211 classifications to the NAEP Reading items [and rubric scoring categories].”

(p.54)

An item or score category was deemed “classified” if there was majority agreement; that is, if at least 4 of the 7 panel members agreed about the score band to which an item (or creditable score category under an item rubric) was assigned.

Of the 211 determinations to be made, there was only one for which there was no majority agreement (the assignment of a PSCR rubric to a CRS score band). Of the remaining 210 determinations, 181 were unanimous.

The reading panel was able to classify 137 items or rubric categories (about two-thirds of the determinations to be made) to the CRS score bands. Of the 97 multiple choice items, 81 (or 84%) were classified. Of the 113 rubric score categories for items, 56 (or 50%) were classified. The reasons some multiple choice items and rubric score categories could not be classified were related to the differences in the ACT and NAEP reading domains described above. These reasons include the presence of constructed response items in NAEP but not the ACT, the presence of items involving multiple texts in NAEP but not the ACT, and the greater presence of “Critique/Evaluate” type items in NAEP than the ACT.

Of the 137 classifications, 24 were in the score bands from 13-19; 113 of the classifications were in the score bands from 20-36. This is noted because the ACT College Readiness Benchmark for reading is 21. The ACT College Readiness Benchmark signifies the score at which a student has a 50% chance of attaining a grade of B or better in a relevant subject and a 75% change of a C or better. In addition, the Governing Board conducted a survey of postsecondary institutions’ use of tests in making entry-level decisions about placement into remedial or regular credit-bearing courses. With respect to the ACT, 18 was the mean reading score below which students were deemed to need remedial course work (Fields and Parsad, P. 19). Whereas this provides a context for the study results, it must be kept in mind that in making their judgments about item classifications, the panelists did not have data about NAEP item difficulty or data on how performance on NAEP compares with performance on the ACT.

Finally, although the study results support the conclusion that the 12th grade NAEP reading assessment measures content directly related to academic preparedness for college, it is noted that the study was conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in reading.

**Reading: SAT**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the SAT critical reading test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found
Overall, the study found similar content in the NAEP 12th grade reading assessment and the SAT critical reading test. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

**What is the correspondence between the reading content domain assessed by NAEP and that assessed by SAT?**

The greatest commonality between the two tests is their shared emphasis on the broad skills of integrating and interpreting both informational and literary texts. This is evident in the majority of items from both tests aligned to NAEP Standard 2, Integrate/Interpret,” including many to Goal 2.1, “Make complex inferences within and across both literary and informational texts.”

Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), there is also considerable overlap at the level of more specific skills.

**To what extent is the emphasis of reading content on NAEP proportionally equal to that on SAT?**

Both tests had many of their item alignments to the same NAEP “Integrate/Interpret” objectives, often with similar percentages of alignments. Although there were some differences in emphasis, both tests also had notable percentages of alignments to SAT Objectives B.1.1–B.1.3 and B.1.5. Skills with overlap include inferring/analyzing the following:

- the “main idea” and “author’s purpose” (SAT Objective B.1.1 and NAEP Objectives 2.3.a and 2.1.f);
- the “tone and attitude” of an author or character (NAEP Objectives 2.2.a and 2.2.c and SAT Objective B.1.4);
- the use of “rhetorical strategies” (NAEP Objective 2.1.d and SAT Objective B.1.2); and
- connections between ideas, perspectives, or problems (NAEP Objective 2.1.b and SAT Objectives B.1.3 and B.1.5).

Additionally, in the area of greatest content overlap—items on both tests aligned to objectives for NAEP “Integrate/Interpret” and aligned to SAT “Passage-Based Reading” Objectives B.1.1–B.1.5—both tests met the typical threshold criteria for depth of knowledge consistency…

Despite these similarities, there are some notable differences in emphasis between the two assessments. Both tests assess vocabulary skills. However, NAEP addresses vocabulary exclusively in the context of passage comprehension, while the majority of SAT vocabulary items are in a sentence-completion format, in which context plays a more limited role. This difference reflects NAEP’s emphasis on the understanding of
word meaning in context; the assessment is not intended to measure students’ prior knowledge of word definitions. The SAT sentence-completion items provide some context within the single sentence text, but in many cases, students’ success on the items almost certainly depends on their prior knowledge of word definitions.

In addition, panelists found considerably less emphasis in SAT than in NAEP on literal comprehension and critical evaluation, particularly the evaluation of the quality or effectiveness of an author’s writing, skills covered in the NAEP standards “Locate/Recall” (locating/recalling specific details and features of texts) and “Critique/Evaluate” (evaluating texts from a critical perspective), respectively. This difference suggests a greater emphasis on these skills in NAEP.

Even with the minimal coverage of NAEP “Locate/Recall” and “Critique/Evaluate” standards by SAT items, all NAEP items found a match in the SAT framework. However, the broad language of the SAT framework can encompass the range of the NAEP items. For example, SAT Goal B.2, “Literal Comprehension,” refers to items that “ask what is being said” in a “small but significant portion of a reading passage,” a description that can easily accommodate most NAEP “Locate/Recall” items and objectives. In fact, nearly all items on the NAEP short version that were coded to “Locate/Recall” objectives in the NAEP framework were matched to SAT Goal B.2 in the SAT framework.

Similarly, SAT Objective B.1.3, to which approximately one-quarter of NAEP items aligned, includes “Evaluation,” the primary focus of NAEP “Critique/Evaluate.” The description in SAT Objective B.1.3 of items that “ask the test taker to evaluate ideas or assumptions in a passage” is compatible at a very general level with NAEP “Critique/Evaluate” objectives addressing the author’s point of view, logic, or use of evidence. SAT Objective B.1.2, “Rhetorical Strategies,” is also broad enough in its language to make it a reasonable match for some NAEP “Critique/Evaluate” items focused on “author’s craft” or use of “literary devices.” In the NAEP short version, all items that aligned to “Critique/Evaluate” objectives in the NAEP framework were aligned to either SAT Objectives B.1.2 or B.1.3, or both.

**Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire reading subdomains are missing or not aligned?**

With regard to differences in content as described in the NAEP framework, SAT items had limited coverage of the knowledge and skills described by the NAEP standards “Locate/Recall” and “Critique/Evaluate.” This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.
With regard to differences in complexity, NAEP items and objectives had a range of depth of knowledge including items at DOK Levels 1, 2, and 3, while SAT items and objectives were coded primarily at Levels 2 and 3.

Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT “Passage-Based Reading” objectives and NAEP “Integrate/Interpret” objectives, as well as some important differences.

Reading: ACCUPLACER
The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the ACCUPLACER reading test. The ACCUPLACER is used specifically to determine whether entry-level students have the reading skills necessary for college level work or require remedial reading courses. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf.

Overall, the study found similar content in the NAEP 12th grade reading assessment and the ACCUPLACER reading test, although the content of NAEP is much broader and complex. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

*What is the correspondence between the reading content domain assessed by NAEP and that assessed by ACCUPLACER?*

The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts.

The shared emphasis on the comprehension and interpretation of informational text can also be seen in the alignments on both tests to the ACCUPLACER framework. Although the ACCUPLACER standards do not explicitly refer to text type, they focus almost exclusively on elements typical of informational text. A majority of both NAEP and ACCUPLACER items were matched to the ACCUPLACER standard “Inferences,” and both tests had notable percentages of alignments to “Direct statements and secondary ideas” and “Applications.” A smaller percentage of items on both tests were aligned to “Identifying main ideas.”
To what extent is the emphasis of reading content on NAEP proportionally equal to that on ACCUPLACER?

As previously discussed, the alignments both within and across frameworks show that both tests emphasize the comprehension and interpretation of informational text, particularly through the use of inference. Within this broad area of convergence, however, there are differences in emphasis revealed in the alignments to specific objectives within both frameworks. In relation to the NAEP framework, the NAEP short-version items showed a far greater emphasis on the comprehension of vocabulary in context (Objective 4.a) and on the analysis of an author’s use of language (Objective 1.d). In relation to the ACCUPLACER framework, NAEP items showed more emphasis on the use of inference to interpret text (“Inferences”). The higher percentage of NAEP items aligned to “Applications” also reflects the greater emphasis in NAEP on understanding authors’ use of language.

In relation to the ACCUPLACER framework, the ACCUPLACER items showed a greater emphasis than the NAEP items on the identification of main ideas. In relation to the NAEP framework, the ACCUPLACER items showed more emphasis on the recall of specific details, facts, and information (NAEP 1.1.a).

In general, in the cross-framework alignments, the matches found in each test to the other’s framework (NAEP to ACCUPLACER and ACCUPLACER to NAEP) tended to be for the most general objectives within that framework. For example, the great majority of hits for ACCUPLACER items to NAEP objectives for “Integrate/Interpret” were to two of the most broadly stated NAEP objectives, “Draw conclusions” (2.3.b) and “Compare or connect ideas” (2.1.b). Many of the more specific NAEP objectives for “Integrate/Interpret,” such as “Find evidence in support of an argument” (2.2.c), received far fewer or no hits from ACCUPLACER items. Compared to ACCUPLACER, the NAEP items were more evenly distributed among NAEP objectives.

The majority of alignments for NAEP items to ACCUPLACER standards were also to the broadest of those standards—“Inferences” and “Applications,” both of which overlap in content with a number of NAEP objectives but at a higher level of generality. The more specific ACCUPLACER standard, “Identifying main ideas,” received far fewer alignments from NAEP items.

Are there systematic differences in content and complexity between the NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between the NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire reading subdomains are missing or not aligned?

In regard to differences in content, NAEP addresses reading skills related to both literary and informational text, while ACCUPLACER does not address reading skills specific to literary text. As expected, based on the framework-to-specifications [review]… ACCUPLACER items had minimal matches to NAEP objectives for literary text. The main area of alignment of ACCUPLACER items to the NAEP framework, NAEP...
objectives in “Locate/Recall” and “Integrate/Interpret,” applied to informational text only or to both informational and literary text.

The ACCUPLACER items also had minimal to no coverage of the NAEP standard “Critique/Evaluate.” … overall, the language of the ACCUPLACER objectives (“understand,” “comprehend,” “recognize”) places more emphasis on comprehension and interpretation of text (“distinguish the main idea from supporting ideas” or “perceive connections between ideas made—implicitly—in the passage”) than on critical analysis or evaluation (“Evaluate the strength and quality of evidence used by the author to support his or her position” in NAEP Objective 3.3.b, or “Judge the author's craft and technique” in NAEP Objective 3.1.a).

In regard to complexity, both assessments were found to meet the criteria for depth of knowledge consistency in relation to their own framework. In relation to the NAEP framework, however, only the NAEP items met the criteria for DOK consistency for all NAEP standards. The ACCUPLACER items met the criteria for depth of knowledge consistency only for NAEP “Locate/Recall.” Although the majority of the ACCUPLACER item alignments were to objectives for NAEP “Integrate/Interpret,” over half of these items were found to have a DOK level below that of the standard. In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the more in-depth reasoning required by DOK Level 3. NAEP, by contrast, includes much longer reading passages and both multiple-choice and constructed-response items.

NAEP covers skills specific to the comprehension and analysis of literary text while ACCUPLACER does not. In addition, NAEP covers the skills of evaluating and critiquing text, skills not addressed by ACCUPLACER. Finally, NAEP has a wider range of cognitive complexity than ACCUPLACER, with a substantially higher percentage of items at DOK Level 3, requiring more in-depth analysis or evaluation. However, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

Overall, the NAEP items covered a broader range of cognitive complexity than the ACCUPLACER items. This is also apparent in the frameworks. The three NAEP standards, defined in terms of three different “cognitive targets” (“Locate/Recall,” “Integrate/Interpret,” and “Critique/Evaluate”), cover a broader range of cognitive complexity supported by the use of longer reading passages and the inclusion of both short and extended constructed-response items. The language of the ACCUPLACER standards (“understand,” “comprehend,” “recognize”) places more emphasis on comprehension and interpretation of text (e.g., “distinguish the main idea from supporting ideas” in ACCUPLACER A, “Identifying main ideas,” or “perceive connections between ideas made—implicitly—in the passage” in ACCUPLACER C, “Inferences”) than on critical analysis or evaluation (e.g., “Evaluate the strength and quality of evidence” in NAEP 3.3.b, or “Judge the author’s craft” in NAEP 3.1.a). In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the cognitive complexity typical of DOK Level 3 items. Although the
NAEP items show a greater range of cognitive complexity and a greater emphasis on critical thinking, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

The Content Alignment Studies: Summary Discussion for Reading

The NAEP 12th grade reading framework, test questions, and, for constructed response items, the score category rubrics, were compared with the analogous domain descriptions and test questions for the ACT, SAT, and ACCUPLACER reading tests. These three tests are used for college admissions and placement. They are well established and have been used for these purposes for many years by professionals in postsecondary education. The test publishers regularly survey secondary and postsecondary educators about relevant content and have conducted research that supports the validity of the test content for the intended inferences and uses. The underlying assumption is that if the content of the 12th grade NAEP reading assessment is similar to the content of these reading tests, then the NAEP content is directly related to “academic preparedness for college.”

The ACT study found that “All of the skills highlighted in the ACT [reading] domain and in the [ACT] College Readiness Standards [for reading] were identified within the NAEP Reading framework.” At the same time, there was content measured by NAEP that was not present in the ACT reading test. In assigning 211 NAEP 12th grade reading items and rubric score categories to the ACT College Readiness Standards for reading, there were 137 positive classifications, or about 65% of the possible classifications. The multiple choice items and rubric score categories that could not be classified were those that measured content not measured by the ACT reading test.

The SAT study found that “Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT “Passage-Based Reading” objectives and NAEP “Integrate/Interpret” objectives, as well as some important differences.” With respect to the differences, “…SAT items had limited coverage of the knowledge and skills described by the NAEP standards “Locate/Recall” and “Critique/Evaluate.” This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.”

The ACCUPLACER study found that “The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts…Overall, the NAEP [frameworks and] items covered a broader range of cognitive complexity than the ACCUPLACER items…The three NAEP standards, defined in terms of three different “cognitive targets” (“Locate/Recall,” “Integrate/Interpret,” and “Critique/Evaluate”), cover a broader range of cognitive complexity supported by the use of
longer reading passages and the inclusion of both short and extended constructed-response items.”

The results across the three studies are consistent. In general, the content of the ACT, SAT, and ACCUPLACER reading tests are present in NAEP, but NAEP is generally broader. Alignment between NAEP and the other three respective assessments is substantial, but not perfect; perfect alignment is not expected. A component of the SAT critical reading assessment not present in NAEP is sentence completion, measuring vocabulary knowledge in a different way than NAEP does.

These results support the conclusion that

- The NAEP 12th grade reading assessment measures academic knowledge and skills that are also covered by other assessments designed and used to make judgments about the academic preparedness of college freshmen for placement into entry-level, credit bearing, non-remedial college courses that meet general education degree requirements, and
- NAEP 12th grade reading test items and rubric scoring categories for items are appropriate for obtaining evidence of test takers’ possession of knowledge and skills needed for college freshmen to be placed into ECNRG courses requiring college level reading.

**The Content Alignment Studies: Mathematics Results**

**Mathematics: ACT**

The Governing Board contracted with ACT, Inc. to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACT mathematics test. The full report can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf).

The mathematics panel was composed of 7 members, with expertise in mathematics instruction at the high school and college levels. The panel was about evenly divided in terms of prior familiarity with either the ACT or NAEP mathematics domains.

The panel found considerable similarity in the content of the NAEP 12th grade mathematics assessment and the ACT. For example, the NAEP 12th grade mathematics framework was compared to the ACT mathematics domain and the ACT College Readiness Standards for mathematics. The ACT College Readiness Standards (CRS) are descriptions of the content (i.e., the knowledge and skills) measured by the ACT mathematics test in score bands along the ACT 1-36 point scale from 13-36 (see [http://www.act.org/standard/planact/math/index.html](http://www.act.org/standard/planact/math/index.html)). The panel concluded that

“… the two assessments have much of their content domains in common. However, in the NAEP-to-ACT comparison, the difference in specificity with which the domains are articulated in the assessment documents left the panel uncertain as to whether a number
of NAEP content topics—those pertaining to transformations, probability, statistics, and data analysis—are assessed by the ACT. In addition, there was some uncertainty within the panel on the degree to which higher-order analytic skills were assessed, and it was the sense of the panel that the ACT Mathematics Test contained few items involving high mathematical complexity, at least as the NAEP defines it. With regard to the ACT to-NAEP comparison, the Mathematics panel found nearly all of the ACT Mathematics domain and College Readiness Standards reflected in the NAEP Mathematics domain, but determined that a number of the lower-level topics in the ACT Pre-Algebra subdomain were more consistent with Grade 8 NAEP topics. All of these points suggest that while there may be substantial overlap in what the two assessments measure and how they measure it, there are areas of difference, as well. (p. 17)

The mathematics panel also conducted an item classification study, in which the NAEP 12th grade mathematics items were classified in relation to the ACT College Readiness Standards for Mathematics.

An item or score category was deemed “classified” if there was majority agreement; that is, if at least 4 of the 7 panel members agreed about the score band to which an item (or creditable score category under an item rubric) was assigned.

Of the 229 determinations to be made, panel members believed that every item or rubric category could be classified to some CRS score range. However, there were 39 for which there was no majority agreement (17 multiple choice items and 22 rubric categories) on what the classification should be; therefore those items were not considered assigned to a CRS score band. Of the remaining 190 determinations, 24 were unanimous, 142 involved classifications to adjacent score ranges and 24 involved classifications to non-adjacent score ranges.

Of the 108 multiple choice items, 91 (or 84%) were classified. Of the 121 rubric score categories for items, 99 (or 82%) were classified.

Of the 190 classifications, 10 were in the score bands from 13-19; 180 of the classifications were in the score bands from 20-36. This is noted because the ACT College Readiness Benchmark for mathematics is 22. The ACT College Readiness Benchmark signifies the score at which a student has a 50% chance of attaining a grade of B or better in a relevant subject and a 75% chance of a C or better. In addition, the Governing Board conducted a survey of postsecondary institutions’ use of tests in making entry-level decisions about placement into remedial or regular credit-bearing courses. With respect to the ACT, 19 was the mean mathematics score below which students were deemed to need remedial course work in mathematics (Fields and Parsad, p. 13). Although this provides a context for the study results, it must be kept in mind that in making their judgments about content, the panelists did not have data about NAEP item difficulty or data on how performance on NAEP compares with performance on the ACT.

Finally, although the study results support the conclusion that the 12th grade NAEP mathematics assessment measures content that is also covered by other assessments designed and used to make judgments about academic preparedness for college, it is noted that the study was
conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in mathematics.

**Mathematics: SAT**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the SAT mathematics test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf).

Overall, the study found similar content in the NAEP 12th grade mathematics assessment and the SAT mathematics test. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

> **What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by SAT?**

At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard. Each framework contains both general and specific objectives, although the SAT objectives, which are presented as content topics without indication of the cognitive level at which that content would be assessed, may be interpreted as more general than the NAEP objectives.

Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions.

**To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on SAT?**

The greatest commonality between the two tests is their emphasis at the standard level. This is evident in the distribution of percentages of total hits from both assessments matched to each set of standards. Although there are some differences of emphasis, such as the full NAEP item pool’s greater proportion of alignment to SAT “Data analysis, statistics, and probability,” and the SAT short-version’s greater proportion of alignment to SAT “Geometry and measurement,” the proportions of alignments to “Algebra and functions” and “Number and operations” are comparable. There is also considerable overlap among some specific skills, with both assessments addressing many of the same NAEP “Number properties and operations” objectives and SAT objectives…
Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), it is clear that both assessments emphasize a number of the same or closely related skills. These include properties, equivalence, and operations on rational numbers (included in NAEP Goals 1.1 and 1.3 and included in SAT Objective N.2) and properties of two-dimensional shapes (included in NAEP Goals 3.1 and 3.3 and included in SAT Objective G.6).

Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework. The language of the objectives in the SAT framework is sufficiently broad to encompass the range of the NAEP items. For example, SAT Objective A.10, “Basic concepts of algebraic functions,” may accommodate most of the items aligning to the seven objectives within NAEP Goal 5.1, “Patterns, relations, and functions.” Finally, some NAEP items were found to be uncodable to the SAT objectives. These items assessed skills not present in the SAT framework.

The two tests are also similar in the average DOK [Depth of Knowledge] levels of items. However, while most items in both tests were found to be at DOK Level 2, NAEP items had a wider range of DOK than did SAT items, with more NAEP items coded to Levels 1 and 3. The Level 3 NAEP items often involved application of concepts through short or extended constructed-response items. Both tests also met depth-of-knowledge consistency overall (with each not meeting this criterion for only one standard as rated by one panel).

Overall, despite differences in alignment at the detailed specific objective level, differences in emphasis at the standard level, and a small difference in ranges of depth of knowledge, there is considerable overlap of content and complexity between [the NAEP 12th grade mathematics assessment and the SAT mathematics test].”

**Mathematics: ACCUPLACER**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test. The ACCUPLACER is used specifically to determine whether entry-level students have the mathematic knowledge and skills necessary for college level work or require remedial mathematics courses.
WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf.

Overall, the study found similar content in the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test, although the content of NAEP is much broader and complex. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

“What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?”

The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra (included in NAEP’s “Number properties and operations” and “Algebra” standards and in ACCUPLACER’s “Arithmetic,” “Elementary algebra,” and “College level math” standards), although their respective degrees of alignment and focus in these subdomains vary. Whereas the NAEP items focus primarily on number or number operations and algebra content at the grade 12 level, with an emphasis on problem solving and application of concepts at that grade level, the ACCUPLACER items span a wider developmental and grade-level range (from basic to more advanced). This difference in focus is consistent with the purposes of the two assessments and their frameworks. The NAEP objectives are written to describe assessable content for grade 12 mathematics; thus, the 130 objectives tend to address the skills and concepts specific to that grade. The purpose of ACCUPLACER is to help determine appropriate placement for an individual student, and so the 87 ACCUPLACER objectives are spread more broadly across grade levels and are intended to be more general.

To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?

Regarding alignment to the NAEP framework, within the “Number properties and operations” and “Algebra” standards, NAEP items had broader overall coverage of the NAEP objectives than did ACCUPLACER. The 42 NAEP items (the short version used for within-framework alignment) aligned to 72 NAEP objectives, whereas the 105 ACCUPLACER items (one complete form of each of the three ACCUPLACER Mathematics Core tests) aligned to only 56 NAEP objectives, with 44% of the ACCUPLACER item alignments aligning to only three NAEP objectives (all in “Number properties and operations” and “Algebra”). These differences in breadth and emphasis between the two assessments were evident across all NAEP standards. For example, in each assessment, items were aligned to four NAEP “Algebra” objectives for which the other assessment had no alignments, reflecting differences in emphasis within that standard.

Regarding alignment to the ACCUPLACER framework, ACCUPLACER items in the short version of 45 items covered all three standards—“Arithmetic,” “Elementary algebra,” and “College level math”—with a relatively even distribution, although
“College level math” had the lowest percentage of item alignments. NAEP items in the full pool of 164 items also covered “Arithmetic,” “Elementary algebra,” and “College level math,” with a fairly even distribution of approximately one-third of NAEP codable items aligned to each standard, although “Elementary algebra” received somewhat fewer item alignments. Despite these differences in emphasis, however, considering only codable items, the percentages of alignments to each ACCUPLACER standard were relatively evenly distributed in both assessments and similar in distribution across assessments. At the objective level, the distribution of item alignments to objectives was relatively even on both tests, although each assessment was aligned to some objectives to which the other was not.

In summarizing cross-framework alignment, there was somewhat less even distribution of items than observed in within-framework alignment. The majority of items on each test were found to align to objectives on the other test. However, the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from “Number properties and operations” in the NAEP framework, and four of the five goals in “Algebra.” Conversely, the NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives.

Are there systematic differences in content and complexity between NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives (20 were rated uncodable by the majority of panelists in each panel). For other NAEP items that were aligned to an ACCUPLACER objective, there were often parts of those items not addressed by the objective. These items were coded as aligned, since they do assess an ACCUPLACER objective, but parts of the items also cover other skills not included in the ACCUPLACER framework.

Regarding differences in alignment of complexity, the items from both tests that aligned to the NAEP standards met the typical depth-of-knowledge (DOK) consistency threshold; that is, the items assessed the objectives at or above the DOK level of the objective. The items from both tests that aligned to the ACCUPLACER standards had somewhat different ranges of DOK. The ACCUPLACER short-version items were divided fairly evenly between Level 1 and Level 2. The NAEP items aligned to the ACCUPLACER framework had a wider range of DOK, with items at Level 1, 2, and 3, and a greater emphasis on Level 2 than was in the ACCUPLACER items.”
The Content Alignment Studies: Summary Discussion for Mathematics

The NAEP 12th grade mathematics framework, test questions, and, for constructed response items, the score category rubrics, were compared with the analogous domain descriptions and test questions for the ACT, SAT, and ACCUPLACER mathematics tests. These three tests are used for college admissions and placement. They are well established and have been used for these purposes for many years by professionals in postsecondary education. The test publishers regularly survey secondary and postsecondary educators about relevant content and have conducted research that supports the validity of the test content for the intended inferences and uses. The underlying assumption is that if the content of the 12th grade NAEP mathematics assessment is similar to the content of these mathematics tests, then the NAEP content is directly related to “academic preparedness for college.”

The ACT study found that “With regard to the ACT to-NAEP comparison…nearly all of the ACT Mathematics domain and College Readiness Standards [are] reflected in the NAEP Mathematics domain, but…a number of the lower-level topics in the ACT Pre-Algebra subdomain were more consistent with Grade 8 NAEP topics.” In the NAEP-to ACT comparison, there was uncertainty about “…whether a number of NAEP content topics—those pertaining to transformations, probability, statistics, and data analysis—are assessed by the ACT….and the degree to which higher-order analytic skills were assessed…and it was the sense of the panel that the ACT Mathematics Test contained few items involving high mathematical complexity, at least as the NAEP defines it.”

The SAT study found similar content in the NAEP 12th grade mathematics assessment and the SAT mathematics test. “At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard… Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions… While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework.

The ACCUPLACER study found that “The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra…although their respective degrees of alignment and focus in these subdomains vary… the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from “Number properties and operations” in the NAEP framework, and four of the five goals in “Algebra.” Conversely, the
NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives…Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives…”

The results across the three studies are consistent. In general, the content of the ACT, SAT, and ACCUPLACER mathematics tests are present in NAEP, but NAEP is generally broader. Alignment between NAEP and the other three respective assessments is substantial, but not perfect; perfect alignment is not expected.

These results support the conclusion that
- The NAEP 12th grade mathematics assessment measures academic knowledge and skills that is also covered by other assessments designed and used to make judgments about the academic preparedness of college freshmen for placement into entry-level, credit bearing, non-remedial college courses that meet general education degree requirements for mathematics, and
- NAEP 12th grade mathematics test items and rubric scoring categories for items are appropriate for obtaining evidence of test takers’ possession of knowledge and skills needed for college freshmen to be placed into ECRNG college mathematics courses.

**Discussion of Test Uses and Consequences in Relation to the Proposed Inferences**

The National Assessment of Educational Progress is an independent monitor of student academic achievement in the United States. It reports on achievement at specific points in time and trends in achievement over time. NAEP reports to the public, national and state policymakers, and education leaders. It assesses student achievement at grades 4, 8, and 12 in important subjects. NAEP is used to compare performance across states and for 21 urban school districts. NAEP results are reported by gender, race/ethnicity, socioeconomic status, and for students with disabilities and students who are English language learners.

The audiences and the uses of NAEP are well established. They will not change as a result of the added meaning afforded by the inferences proposed in this validity argument. However, providing familiar external referents for performance on 12th grade NAEP will greatly enhance the understanding of NAEP results by its audiences.

Currently, there are either no or very low stakes consequences associated with the use of NAEP results. NAEP is not used as a basis for evaluating or diagnosing individual students, classroom or school performance, the effectiveness of individual teachers or administrators, or for any other accountability purpose. This will not change as a consequence of the inferences proposed in this validity argument.
Although the uses and consequences of NAEP will not change, employing the proposed inferences for NAEP reporting will bring a potential for misinterpretation. NAEP reports should include text explaining the limitations on interpretation and other caveats that were discussed in detail on pages 8-10 above.

**Summary and Conclusion**

The National Assessment Governing Board decided to determine the feasibility of transforming NAEP into a measure of academic preparedness for college. Consequently, the Governing Board made changes to the NAEP 12th grade reading and mathematics frameworks with the explicit purpose of measuring academic preparedness for college. The Governing Board conducted research that established a high degree of overlap between the content of the NAEP 12th grade reading and mathematics assessments and the content of widely used college admissions and placement tests.

Through a partnership with the College Board, performance on 12th grade NAEP was compared with performance on the SAT mathematics and critical reading assessments, with correlations of .91 and .74 respectively. Analyses of these data examined the average NAEP scores and interquartile ranges for students scoring “at” and “at or above” the College Board College Readiness Benchmarks for reading and mathematics. Similar analyses were conducted using data from the 2005 and 2009 NAEP High School Transcript Studies, using the college readiness benchmarks developed by ACT and by the College Board. A longitudinal study was conducted in partnership with the Florida Department of Education, following the 12th grade students in the state NAEP sample into Florida public postsecondary institutions, employing Florida’s longitudinal data base. The average NAEP scores and interquartile ranges were calculated for the Florida students in relation to the ACT or SAT college readiness benchmarks, whether they achieved a first-year GPA of B- or better, and whether they were placed into a remedial course in their first year of college.

The results of these analyses were consistent across studies and across years. They support the conclusions that students in the NAEP 12th grade distribution at or above the Proficient achievement level in reading and at or above 163 on the NAEP score scale for mathematics are

- likely to be academically prepared for entry-level, credit-bearing non-remedial courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions, and
- not likely to need remedial/developmental courses in reading or mathematics in college

That the NAEP sampling, scaling and statistical procedures yield accurate estimates of the percentage of students scoring at or above a selected cut-score (i.e., NAEP achievement level) is well established as a result of numerous validity studies and evaluations.

Thus, the NAEP 12th grade preparedness research results support the inferences that
For reading:

Given the design, content, and characteristics of the NAEP 12\textsuperscript{th} grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

the percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.

For mathematics:

Given the design, content, and characteristics of the NAEP 12\textsuperscript{th} grade mathematics assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

Including these inferences in NAEP 12\textsuperscript{th} grade reports will add meaning to the interpretation of the NAEP 12\textsuperscript{th} grade results. However, steps must be taken to avoid potential misinterpretation. NAEP reports using these inferences must also include the limitations on interpretation and caveats described previously in this validity argument. In addition, the reports should explain the rationale for NAEP reporting on academic preparedness and describe appropriate and inappropriate uses of the results.
References


WestEd ACCUPLACER Reading Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf).

WestEd SAT Mathematics Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pdf).

WestEd SAT Reading Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf).

Setting Achievement Levels on the NAEP 2014 Technology and Engineering Literacy (TEL) Assessment

Status: Information and discussion

Objective: To discuss issues that are being addressed and that should be addressed in TEL scaling analyses.

Attachment: C-1 NCES description of current issues being reviewed in TEL Scaling.

C-2 Overview of the Evidence-Centered Design Method in the article “Evidence-Centered Design for Certification and Licensure” (Williamson, Mislevy, and Almond, 2004)

Background
At the March 1, 2013 meeting, the Committee began discussion on setting achievement levels for the 2014 NAEP TEL assessment. For the May 17, 2013 meeting, an issues paper was developed to support procurement and project planning for developing recommended achievement levels for TEL. In the Committee’s May 2013 discussion, the Committee expressed a need for more information before proceeding with procurement plans, particularly regarding TEL scaling issues that could hinder a strong TEL Achievement Level Setting (ALS) effort.

Timeline
The following timeline provides a preliminary list of key dates and activities related to TEL assessment development and achievement level setting.

<table>
<thead>
<tr>
<th>Date</th>
<th>Activity</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008 - 2010</td>
<td>TEL Framework development</td>
<td>ADC, Board, WestEd (contractor)</td>
</tr>
<tr>
<td>2010 - 2012</td>
<td>Assessment development for 2013 pilot test</td>
<td>NCES, NAEP contractors</td>
</tr>
<tr>
<td>2010 - 2012</td>
<td>Item review for 2013 pilot test</td>
<td>NCES, NAEP contractors, TEL Standing Committee, ADC</td>
</tr>
<tr>
<td>Early 2013</td>
<td>Pilot test – national sample, grade 8</td>
<td>NCES, NAEP contractors</td>
</tr>
<tr>
<td>May 2013</td>
<td>TEL ALS issues paper</td>
<td>COSDAM, consultant</td>
</tr>
<tr>
<td>Late 2013</td>
<td>ALS procurement and contract award</td>
<td>Board staff, COSDAM</td>
</tr>
<tr>
<td>Early 2014</td>
<td>Operational administration – national sample, grade 8</td>
<td>NCES, NAEP contractors</td>
</tr>
<tr>
<td>2014 - 2015</td>
<td>Final phase of ALS process and Board action on TEL</td>
<td>COSDAM, ALS contractor, Board</td>
</tr>
<tr>
<td>2015</td>
<td>Reporting TEL results</td>
<td>Board, NCES, contractors</td>
</tr>
</tbody>
</table>
**TEL Assessment Design**

The 2014 Technology and Engineering Literacy (TEL) assessment is based on the Board-adopted Framework and Specifications (see Abridged TEL Framework in Attachment D-2; complete documents are at [www.nagb.org](http://www.nagb.org), Publications).

The TEL assessment is composed of three major areas:
- Design and Systems
- Information and Communication Technology
- Technology and Society

Another key dimension of the TEL assessment is the three practices, each of which is applicable to the three major areas noted above:
- Understanding Technological Principles
- Developing Solutions and Achieving Goals
- Communicating and Collaborating

The TEL assessment was developed using an evidence-centered design (ECD) approach (see Attachment C-2). From the beginning, all TEL tasks and items were designed using an evidential chain of reasoning that links what is to be measured, the evidence used to make inferences, and the tasks used to collect the desired evidence. In addition to student responses to complex tasks and discrete items, the computer-based TEL assessment allows NAEP to capture a wide array of data on student performance. For example, NAEP will collect information on how students interact with the TEL simulations and experiments. Such data may include the number of experimental trials run and the number and types of variables controlled. These observable data on “strategies and processes” may also contribute to the scoring of student performance.

**TEL Reporting**

Based on the ECD approach, TEL reporting will be expanded beyond the traditional NAEP scores. It is expected that data from the complex performance tasks and discrete items will be reported in a number of ways:
- A composite scale score on which the achievement levels will be set
- Subscores for the content areas (Design and Systems; Information Communication Technology; Technology and Society)
- Reporting on the practices (Understanding Technological Principles; Developing Solutions and Achieving Goals; Communicating and Collaborating)
- Information on students’ processes and strategies, related to the ECD model, captured as observable data from their work on the TEL scenario-based tasks.

**Ongoing Potential Discussion Questions for COSDAM**
- Given the emerging field of setting achievement levels on ECD-based complex performance assessments, what additional background materials are needed to inform the COSDAM/Board decision on an appropriate method for ALS on the TEL assessment?
- To what extent should research studies be built into the TEL ALS project?
Technology and Engineering Literacy Field Trial Analyses

As part of the 2013 NAEP administration, a large field trial (also often referred to as ‘pilot’ in similar contexts) was conducted for the Technology and Engineering Literacy assessment. The field trial was designed to be similar to other field trials in subjects for which an entirely new framework is used. Specifically, the trial was designed to support both a detailed evaluation of the items and tasks individually as well as how they relate to each other (e.g., through scaling and otherwise correlation-based analyses) using a partial balanced incomplete block design. The analysis of the field trial data focuses on three goals:

1. Individual item performance, including response time, to select discrete items and assemble discrete item blocks. Given the short time frame involved in preparation for the 2014 probe assessment, this analysis has been completed and was based on (observed) item responses.

2. Scaling to evaluate to what extent the relations between the items and tasks reflect the various constructs defined and hypothesized in the framework. This involves both the core content domains (Design & Systems, Information & Communication Technology, and Technology & Society) as well as cross-cutting practices (Understanding Technological Systems, Developing Solutions & Achieving Goals, and Communicating & Collaborating). We are approaching this task in largely two ways: scaling of each of the domains and bi-factor modeling of the constellation of domains and practices (which we coined “competencies”).

3. Further development of extended reporting goals. Extended reporting refers to results based on task-use patterns and other process data (e.g., strategies students use to solve a particular problem, consistency and efficiency in running a particular simulation), is exploratory in nature, is generally task-dependent, and could serve to provide additional context to the broader, generalizable scaled results (i.e., domains and competencies). Specifically, these indicators are related to TEL, but not measurements of TEL.

In addition, the field trial data could be used to further an achievement level setting effort associated with the TEL assessment. Particularly, the scaling analyses in combination with item maps can provide an approximate grouping of items and levels of performance. There are a number of subtleties that will be addressed during this presentation, including the extent to which the field trial results and the probe results are sufficiently comparable, how context and positioning effects may play a role, and how administration design differences could affect outcomes. In addition, operationalizing a new construct in NAEP suggests the need for some dimensionality analyses to determine at what level (e.g., overall, by domain) meaningful standards can and ought to be set.

Besides discussing scaling goals and challenges, in this presentation we will also discuss the kinds of data and reports that could be generated based on the field trial data. In principle, the TEL field trial data are very close in design to a regular assessment and should, therefore, provide the similar kinds of summary performance results that have been used in other achievement level setting activities. However, discrete item and survey question selection does result in comparability issues. In addition to item responses, we also collected time-stamped process data and can explore whether there are other types of data (e.g., behavioral data) that could potentially be useful as corroborating information to standard setting efforts.
What is Evidence-Centered Design?
Evidence-Centered Design (ECD) (Almond, Steinberg, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003) is a methodology applied at Educational Testing Service that emphasizes an evidentiary chain of reasoning for assessment design. This approach results in a more complete representation of the design rationale for an assessment, better targeting of the assessment for its intended purpose, and a more substantial basis for a construct-representation validity argument supporting use of the assessment. The approach encourages test developers to design with intent and provides several advantages:

- **Clarity of purpose** – representation of assessment goals and the relevance of design decisions to those goals.
- **Interrelated design** – modeling the interactions of design decisions and how changes in one aspect of design affect other design elements.
- **Evidentiary requirements** – explication of what constitutes relevant evidence of ability and how such evidence bears on assessment-based decision-making.
- **Validity** – a documented chain of reasoning and rationale underlying design decisions and their relevance to the criterion of interest.
- **Innovation** – a guide for developing assessments targeting elusive domain constructs or using emerging technologies and new item types.

The foundations of ECD stem from validity theory (Messick, 1989), psychometrics (Mislevy, 1994), philosophy (Toulmin, 1958), and jurisprudence (Wigmore, 1937). They adapt the evidence-oriented approach to evaluating the degree to which conclusions about people can be made on the basis of collected evidence. The ECD process centers around four key questions:

1. **Claims**: Who is being assessed and what will be declared about them as a result?
2. **Proficiencies**: What proficiencies must be measured to make appropriate decisions?
3. **Evidence**: How will we target, recognize, and interpret evidence of these proficiencies?
4. **Tasks**: Given practical constraints, what situations will elicit the kind of evidence needed?

Addressing these questions results in three fundamental assessment design models, represented here as Figure 1. These ECD models include:

- **Proficiency Model** – defines the claims and constructs of interest for the assessment and their interrelationships.
- **Evidence Models** – define how observations of behavior are considered as evidence of proficiency.
- **Task Models** – describe how assessment tasks must be structured to ensure opportunities to observe behaviors constituting evidence.

These interrelated models comprise a chain of reasoning for an assessment design that connects the design of assessment tasks to evidence of proficiencies targeted by the assessment, which in turn are formally associated with claims made on the basis of assessment results.

**Figure 1: Fundamental Models of Evidence-Centered Design**

The following presents each of these models in turn with some discussion of their implications in the context of certification and licensure testing.

**Proficiency Model**

The proficiency model is really a combination of the formal assessment claims to be made on the basis of assessment and the proficiencies measured by the test. Claims are the specific arguments being made about people on the basis of assessment results. Proficiencies are measured knowledge, skills and abilities of people that provide the basis for making claims.

In order to make such claims or to identify important proficiencies, one must first have a good understanding of the population being served. Therefore, a precursor to claim specification is a definition of the examinee population, the users of the test results, and the intended use of test results in decision-making by these users. In certification and licensure testing the decision being made on the basis of the assessment is typically straightforward: either to issue or withhold the credential in question. Based on this definition, the sole users of test results are the issuing body of the credential. However, since the credential itself represents a measure of the population being served. Therefore, a precursor to claim specification is a definition of the examinee population, the users of the test results, and the intended use of test results in decision-making by these users. In certification and licensure testing the decision being made on the basis of the assessment is typically straightforward: either to issue or withhold the credential in question. Based on this definition, the sole users of test results are the issuing body of the credential. However, since the credential itself represents a claim about the examinee made by the credentialing organization, it is typical to consider the interests of the users of these credentials (e.g., potential employers, the general public selecting their services, state licensure boards, etc.) when establishing claims. The examinee population is typically defined as individuals who have met some educational and/or practice prerequisites and are seeking the credential in question. Implicit in this definition is the perceived value of the credential and how it benefits the personal and professional interests of the examinee.

The understanding of assessment use and population being served drives the specification of claims being made on the basis of assessment results. These claims are represented as stars in the Proficiency Model portion of Figure 1. For example, in licensure testing a common global claim made on the basis of assessment might be something like, “Can engage in professional practice without representing a risk to the health, safety or well-being of the public.” Often, such a global claim about ability is supported by a number of sub-claims that make explicit statements intended to directly support the overall claim of the assessment. Often these are based on elements of the domain of practice that are ultimately reflected in test content. In this way, it is typical for the claims associated with an assessment design to be organized as a claim hierarchy that elaborates the various arguments that a test score represents about individual ability. As such, the specific claims chosen for an assessment design are often directly related to needs of score reporting or delivery of instruction.

The proficiencies of individuals being measured by the assessment follow from the claims. The claims express the goals of assessment design as states of knowledge about aspects of proficiency and represent the declarations that must be supported by test results. In order to support these arguments, certain levels of ability must be demonstrated during the assessment. It is these proficiencies and the levels required to make certain claims that are specified in the proficiency structure. Assume, for example, that for a certification of computer network engineers there is a claim that such persons are adept at troubleshooting technical problems in network connectivity. It might be reasonable to expect that supporting this claim would require declarative knowledge (recall) of computer networking hardware and their technical capabilities and interconnectivity protocols. It might also be reasonable to expect that supporting this claim requires an ability to employ a logical and efficient cognitive strategy to determine the cause of common network problems. Therefore, two proficiencies that might be implied by such a claim could include “hardware connectivity knowledge” and “strategic troubleshooting.” These proficiency variables are inherently latent (not directly observable) and are therefore the target of the inference process of the assessment. These various proficiencies of interest are represented symbolically in Figure 1 by the set of circles and arrows in the Proficiency Model section. The circles represent various proficiency variables of interest and the arrows reflect known relationships between proficiencies (e.g., correlations or prerequisite relationships) and conditional independence relationships between variables.

The specification of claims and the description of proficiencies that one must possess to support these claims are related to traditional approaches to professional domain analysis. Often this is conducted through traditional job analyses, or in the case of assessments emphasizing strategic problem-solving, cognitive task analysis (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999).

**Evidence Models (Conceptual)**

Operationally, the evidence model specifies the manner in which observations during assessments are used to update estimates of ability. However, during the initial phases of assessment design, the evidence models are specified from a purely optimal domain perspective in order to drive task model development. This conceptual specification begins by imagining that there are no constraints or limitations to the ability to observe and track behaviors in naturalistic settings for a domain of interest. The task is to specify the situations and observable behaviors that are most revealing in terms of distinguishing among levels of ability in the proficiency model. The specification of what these crucial

---

1 Some organizations also define the unsuccessful examinee as a user of results when diagnostic information is provided in order to guide further study.
situations are and what important behaviors can be observed will drive both the evidence model development for scoring and the specification of task models (discussed below). We will revisit the Evidence Model from the scoring perspective after the section on Task Models below.

**Task Models**

Task models are detailed descriptions of families of tasks with similar characteristics. These task models establish the framework, or the blueprint, for producing tasks (or items) that address particular targeted areas of the overall test blueprint. The conceptual evidence model helps to specify the characteristics of these task models that best distinguish among levels of ability. The task models, as pictured in Figure 1, consist of several variable elements: task design features (symbolized by the set of drop-down menu variables in the lower portion of the task model figure); presentation material (symbolized by the video screen icon in the upper right portion of the task model figure); and work products (symbolized by the jumble of shapes in the upper left of the task model figure). Task features describe the intent, construction, and associated design elements and options for a task. Presentation material defines what is presented to an examinee as part of a particular task (e.g., any graphics, any text, a question prompt, options to select from among, etc.). The work products are the resultant examinee data captured as a result of the examinee’s interaction with the task, regardless of whether that data is directly used in scoring or not.

As an example of a portion of a task model, assume that for a test of basic math there was a claim of “Can add two integers” and an associated proficiency called “basic addition” (both very fine-grained examples). An item model (one of many) targeting such an ability might have elements such as those that appear as Table 1.

<table>
<thead>
<tr>
<th>Task Model Variable</th>
<th>Possible Values (implication)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability target</td>
<td>basic addition, sum of two integers</td>
</tr>
<tr>
<td>Difficulty factors</td>
<td>• single-digit integers with single-digit outcome (easier)</td>
</tr>
<tr>
<td></td>
<td>• single digit integers with two-digit outcome (moderate difficulty)</td>
</tr>
<tr>
<td></td>
<td>• two-digit integers with two-digit outcome (harder)</td>
</tr>
<tr>
<td>Reading load</td>
<td>• none (easier)</td>
</tr>
<tr>
<td></td>
<td>• few simple words (moderate)</td>
</tr>
<tr>
<td></td>
<td>• word problem (harder)</td>
</tr>
<tr>
<td>Presentation material</td>
<td>• Equation form of a problem (easier)</td>
</tr>
<tr>
<td></td>
<td>• Word problem embedding problem (harder)</td>
</tr>
<tr>
<td>Work product</td>
<td>• multiple-choice (easier)</td>
</tr>
<tr>
<td></td>
<td>• free response (harder)</td>
</tr>
<tr>
<td></td>
<td>• show work (complex scoring)</td>
</tr>
<tr>
<td></td>
<td>• response only (simple scoring)</td>
</tr>
</tbody>
</table>

Note that along with specification of various aspects of the task, it also indicates how different potential specifications can be expected to impact the difficulty of the item. Figure 2 and Figure 3 present the presentation material (in this case assuming paper presentation) for two items, both of which could be produced from the task model excerpted above. Note that while each is consistent with the task model above, each has characteristics that would tend to make it more or less difficult for the examinee, as well as to deliver (particularly assuming computerized presentation material rather than paper) and to score. These design decisions have implications for how a set of tasks discriminates among different levels of ability targeted by the assessment.

**Figure 2**

\[
4 + 5 = ___
\]

a) 1  
b) 5  
c) 9  
d) 45

**Figure 3**

If you place a box that is 12 inches high on top of another box that is 23 inches high, how high are the two boxes together? Show your work and write your answer below.

Another characteristic to note is that for both multiple-choice and the free-response task there is an implication for a need to extend the level of detail of the task model. For the multiple-choice item, this extension would address how the distracters are developed (note, for example, that option (a) in Figure 2 is the answer someone would obtain if they subtracted 4 from 5 instead of adding) and their order in presentation. For a free-response story problem task, any use of word problems operationally would require further specification of the permissible vocabulary, sentence structures, topics, and representation of actors in the text (as well as considering the impact of the potential confound of reading ability with pure math ability on the measurement of proficiency model variables).
Obviously, the work products for these two examples in Figures 2 and 3 differ in that the former consists of an indication of which option is selected, while the latter consists of an indicated response and the calculations the examinee executed to determine the answer. Part of the consideration of such work products includes the medium used to collect the response. The item in Figure 2 is almost equally viable in paper and computerized format, while for the item in Figure 3, it is more difficult to capture the work products in computerized administration than in paper-and-pencil administration.

The set of task models for an assessment design can be organized hierarchically to facilitate the degree to which the test developer must exercise control over the types of tasks used. For example, the math test item of the general form:

\[ \{\text{integer}\} \{\text{operation}\} \{\text{integer}\} = \{\text{integer}\} \]

is a sub-category of the more general form:

\[ \{\text{single-digit integer}\} + \{\text{single-digit integer}\} = \{\text{single-digit integer}\} \]

Depending on the degree of control which must be exercised in test authoring (based on the test blueprint) and the intent of the item usage, a hierarchy of task models can be developed with varying degrees of specificity of the model design. For example, in cases where the prediction of the specific difficulty of an item is important, the test designer may wish to exercise a relatively high degree of control. This is often the case in efforts that use task modeling as the basis for automatic item generation (Embretson, 1998; Embretson, 1999; Williamson, Johnson, Sinharay, & Bejar, 2002; Newstead, Bradon, Handley, Evans, & Dennis, 2002), in which a computer generates items according to a given task model with no human intervention.

**Evidence Models (Scoring)**

Evidence Models specify how the evidence contained in task data informs belief about Proficiency Model variables. Evidence Models for scoring rely on the Proficiency Model as the fixed target for inferences and the Task Models and tasks authored from them, as the mechanism for producing data to be used in scoring. The Evidence Model for scoring, as presented in Figure 1, consists of two subcomponents:

- **evidence rules** – determine what elements of the task performance constitute evidence and summarizes their values
- **statistical model** – aggregates evidence to update estimates of ability in the proficiency model

Evidence rules transform elements of the work product (the record of examinee task performance) into observables; summary representations of work used by the statistical model to update estimates of proficiency. This process is called **evidence identification** in ECD terminology. In Figure 1, this process is represented in the Evidence Rules portion of the Evidence Models diagram. This figure illustrates how work products from the Task Model are parsed to produce observables, symbolized by the three squares (for three observables). With most multiple-choice questions this process seems almost trivial. For each question there is only one observable, the value of which is determined by comparing the response indicated by the examinee with a predetermined key and representing the observable as a simple 1, for correct, when the response is the same as the key, or 0, for incorrect, when the response does not match the key. In other situations the determination of observables requires more effort, such as for the task presented in Figure 3. Some evidence rules would be required to establish both the correctness of the final answer and for representing the degree of adequacy of shown work in computing the answer. Note that the determination of the value of observable variables also implies using some elements of the work product and ignoring other elements. For example, most scoring of multiple-choice items ignores the particular choice the examinee made if the choice was incorrect, while others might infer the nature of misunderstandings examinees may have when they select particular incorrect answers. Also, in computerized testing environments it is common to collect information on the amount of elapsed time an examinee took to respond to a question despite the fact that this is seldom used in the evidence rules for scoring. The representation in Figure 1 implies three observables obtained from this particular work product.

The statistical model portion of the Evidence Model uses the values of observables to update estimates of ability. In number-right scoring this statistical model is a simple summation function in which the prior value plus the value of the observable (1 or 0) equals the new value. In models using item response theory (IRT) this updating is controlled by the parameters associated with the item for which the observable is being used as evidence and by the fundamental statistical relationships for updating ability estimates from observations under the IRT model being applied. In most common applications (e.g., number right, IRT, etc.) there is a single proficiency variable for ability and a single observable variable from each item. In Figure 1, however, we illustrate the case where three observables are produced from an item and these are used to update two proficiency variables. These two proficiency variables, in turn, represent two of the five proficiency variables that make up the Proficiency Model.
Such a representation illustrates the value of such models for more complex assessment designs, such as for computerized simulations that use automated scoring, while still representing the fundamental structure and critical models for design of traditional assessments.

**Summary of ECD Model Interactions**

In review, the ECD process provides a framework for assessment design that emphasizes a systematic consideration of multiple models for design and their interaction. These begin with the fundamentals of assessment purpose (specification of populations being served, decisions being made, known assessment constraints, etc.) from which formal claims are developed. These claims drive the specification of a Proficiency Model. The implications of the Proficiency Model and claims in combination drive the evidential needs of the assessment, formally represented as the Evidence Model. These needs are actualized in the design of assessment tasks, the blueprints for which are expressed as Task Models.

Once tasks from these models are developed and fielded, the scoring process is essentially a reversal of the development process. The administered tasks result in work products with pre-established properties. These work products are parsed according to the evidence rules of the Evidence Model to produce observables. The statistical model of the Evidence Models is applied to draw inferences about proficiencies on the basis of these observables. Finally, the ultimate values of proficiency variables establish what assessment claims can be supported on the basis of the assessment. These reported claims, in turn, are used by the consumers of score reports to make informed decisions.

**Conclusion**

This work has presented the basic concepts of ECD and made an argument for the relevance and value of such an approach for any assessment design process, whether for a paper-and-pencil assessment using multiple-choice tasks or a computerized assessment using complex simulations and automated scoring. It is hoped that through wide adoption of such a process, the process of assessment design can be improved, both by formalizing processes that good assessment designers perform implicitly, and by encouraging consideration of issues not previously addressed in formal assessment design. It is also hoped that such resultant design rationales strengthen the quality and the validity arguments for use of such measures for their intended purpose.

**References**


Update on Evaluation of NAEP Achievement Levels Procurement

Objective
To receive a brief informational update from NCES on the current status of the procurement being planned to evaluate NAEP achievement levels. Ongoing updates will be provided at each COSDAM meeting.

Background
The NAEP legislation states:

The achievement levels shall be used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public.

In providing further detail, the aforementioned subsection (f) outlines:

(1) REVIEW-

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

B. ISSUES ADDRESSED- Such continuing review shall address--

(i) whether any authorized assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);

(ii) whether student achievement levels are reasonable, valid, reliable, and informative to the public;

(iii) whether any authorized assessment is being administered as a random sample and is reporting the trends in academic achievement in a valid and reliable manner in the subject areas being assessed;

(iv) whether any of the test questions are biased, as described in section 302(e)(4); and

(v) whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.

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

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

Responsively, a procurement has been planned to administer an evaluation of NAEP achievement levels. The last update COSDAM reviewed on this topic was in May 2013.

In this brief written update, NCES provides the Committee with a summary of the status of this procurement.
Evaluation of NAEP Achievement Levels

The National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES) will administer the Evaluation of the NAEP Achievement Levels. The Department’s Contracts and Acquisitions Management office posted a Request for Information (RFI) on FedBizOpps.gov on May 22, 2013. We anticipate that the Department will issue a Request for Proposals (RFP) this summer, with an award announced later this fall.
NAEP 12th Grade Preparedness Research

Based on the Program of Preparedness Research adopted by the Governing Board in March 2009, four categories of research studies were conducted to produce evidence to develop and support the validity of statements for NAEP reporting on the academic preparedness in reading and mathematics of 12th grade students for college and job training.

- content alignment studies;
- statistical relationship studies;
- judgmental standard setting studies; and
- surveys

Additionally, the Texas Commissioner of Higher Education offered the opportunity to conduct a benchmarking study with Texas higher education institutions, and a pilot study to examine the feasibility was conducted.

The research studies completed to date are available in an online technical report. In addition, the NAEP 12th Grade Preparedness Commission conducted a symposium in Washington, DC on July 9, 2013 focused on the Board’s preparedness research results and the Phase 2 research plans.

The following informational attachments are provided:

- Updates related to the Board’s Course Content Analysis Research:
  - College Course Content Analysis Progress Update (Attachment E-1) .......... Page E2
  - Job Training Program Status Update (Attachment E-2) ......................... Page E15

Additionally, the following attachments are provided for reference:

- Proposed research projects for phase 2 of the Board’s preparedness research program (Attachment E-3) ................................................................. Page E16
  - National and State Partnerships
  - Research with Frameworks
- Background materials describing each study category (Attachment E-4) ...... Page E18
In September 2012, the Governing Board awarded a contract to the Education Policy Improvement Center (EPIC) to conduct research on entry level non-remedial college course content in order to (1) identify the prerequisite knowledge and skills in reading and mathematics for entry-level college courses and (2) determine the extent to which there is a match with the content of grade 12 NAEP reading and mathematics assessments. This project addresses academic preparedness for college only—a separate parallel research project addresses preparedness for job training (described below).

In this project, EPIC will determine the entry-level (introductory) credit-bearing courses most frequently taken by entering students that are reflective of college-level reading and mathematics demands and that satisfy general education requirements. These introductory courses should have no college-level prerequisite course requirements, and only non-remedial courses that satisfy general education requirements should be included in the analysis. Further, in cases where multiple versions of a course are offered for majors and non-majors, only the course for non-majors should be included.

Using course artifacts for a generally representative sample of institutions, EPIC will analyze the introductory course artifacts for commonalities and differences in the reading and mathematics prerequisites needed to qualify for placement into the course. From these analyses, EPIC will develop descriptions of the knowledge, skills, and abilities (i.e., the prerequisite KSAs) needed for students to qualify for placement into the introductory course, based on an analysis of the course artifacts. And as part of a set of comparative analyses, EPIC will then use these descriptions to review:

- the description of minimal requirements for placement into college-level coursework as developed in the NAEP preparedness judgmental standard setting (JSS) research
- KSAs represented by 2009 grade 12 items that map to the NAEP scale with a response probability of .67 and fall within the range of cut scores set by the two replicate panels in the JSS research
- 2009 and 2013 grade 12 NAEP items
- the KSAs represented by 2009 items that map in the range of the NAEP score scale from the the Basic level through the Proficient level; and
- the NAEP achievement level descriptions.

A new progress report is attached with more details on the project and a description of work completed to date.
INTRODUCTION AND BACKGROUND

The College Course Content Analysis (CCCA) study is one of a series of studies contributing to National Assessment of Educational Progress’ (NAEP) Program of 12th Grade Preparedness Research conducted by the National Assessment Governing Board (NAGB). The purpose of the CCCA study is to identify a comprehensive list of the reading and mathematics knowledge, skills, and abilities (KSAs) that are pre-requisite to entry-level college mathematics courses and courses that require college level reading based on information from a representative sample of U.S. colleges. The Educational Policy Improvement Center (EPIC) is the contractor working for the Board to conduct this study.

Another goal of the CCCA study is to extend the work of the two previous preparedness studies—the Judgmental Standards Setting (JSS)\(^1\) study, implemented in 2011 and the Job Training Program Curriculum (JTPC) study, implemented in 2012. The CCCA study is designed so the results can be compared to the JSS and JTPC studies, reporting on how this new information confirms or extends interpretations of those earlier studies. The design of the CCCA study is based on the JTPC study but with modifications based on the lessons learned.

The CCCA study will answer four core research questions.

1. What are the prerequisite KSAs in reading and mathematics to qualify for entry-level, credit-bearing courses that satisfy general education requirements?
2. How do these prerequisite KSAs compare with the 2009 and 2013 NAEP reading and mathematics frameworks and item pools?
3. How do these prerequisite KSAs compare with previous NAEP preparedness research (i.e., the descriptions of minimal academic preparedness requirements produced in the JSS research)?
4. How can these prerequisites inform future NAEP preparedness research (i.e., planning and analysis efforts relative to the 2013 grade 12 NAEP reading and mathematics assessments)?

The final report is due May 2014, and until then COSDAM will receive detailed reports at each Board meeting.

METHODOLOGY
The Design Document for the CCCA study is complete. It provides guidance for the study by describing:

- Criteria for collecting courses and artifacts;
- A sampling plan to comprise a representative sample of institutions;
- Review and rating processes, including a training plan and process for ensuring reviewer effectiveness and consistency; and
- The process for ensuring reliability across reviewers providing artifact analysis.

This study comprises three primary phases:

1. Identification and collection of course artifacts,
2. Review of course artifacts by Review Teams, and
3. Analysis and reporting.

The first phase of the study is complete. The course artifacts have been identified, all artifacts have been collected, review packets have been created from those artifacts, and all of the data collection surveys are programmed and ready for use. The second and third phases of the study have begun.

Most notably NAEP Advisory Panels, in both reading and mathematics, were conducted in June of 2013 and the guidance from those panels is being integrated into the implementation of the next phases of the study. Content reviewer training sessions were conducted in early July and the independent content reviews will occur during July 2013 and August 2013. Preparation for data analysis and final reporting has also begun.

OVERVIEW OF ACTIVITIES BY PHASE

Phase 1: Identification and collection of course artifacts

In the CCCA study, a course artifact is defined as a syllabus, a non-textbook based assignment or assessment, and textbook excerpt. In mathematics, there are some instances where the only specifically identified assignments were listed in the syllabus and were from the textbook. In those cases, a textbook based assignment or assessment was allowed. The CCCA sample of artifacts is derived from extant artifacts and combined with newly gathered course artifacts. Extant artifacts contributing to the CCCA sample were extracted from EPIC’s repository of extant artifacts compiled during previous research on entry-level curricula at postsecondary educational institutions. Project staff has solicited new course artifacts as needed to create a complete and representative sample.
EPIC identified a set of inclusion criteria that courses must meet to be included in the CCCA study as well as a set of institutional characteristics of which the final CCCA Artifact Bank must be representative. The final CCCA Artifact Bank will comprise a set of courses and artifacts that will be used as the basis for the content review.

At the conclusion of artifact collection, the CCCA Artifact Bank will include all relevant artifacts compiled into course packets to be reviewed by mathematics and reading content review teams in the second phase of the study.

Phase 1 preparatory work also included the convening of NAEP Advisory Panels, for reading and mathematics respectively, to obtain content-based guidance and recommendations. In these meetings, preliminary coding schemas, training materials and decision rules were reviewed. NAEP advisors also reviewed all of the course packets that will serve in validation data analyses, training sessions, and determining sufficient reviewer competence (qualifying). Guidance from this NAEP Advisory Panels is being integrated into the implementation of the study.

**Phase 2: Review of course artifacts by Review Teams**

In Phase 2, content reviewers are recruited and training materials are developed. Content reviewers will first be trained to review the course packets from a holistic perspective and identify prerequisite mathematics and reading KSAs. In the second independent review training, the NAEP frameworks for grade 12 reading and mathematics will be used as a basis for coding the packets. If additional KSAs are identified during either review sessions, the new KSAs will be documented and included in all successive reviews, comparisons and data analyses. The overarching goal of the CCCA study is to identify all prerequisite KSAs, not just those KSAs associated with the NAEP frameworks.

The CCCA design has embedded validity checks within the process to evaluate the reliability of the review team coding. Two validation packets were created for each of the four course titles in reading and mathematics. The validation packets look like any other course packet and will be mixed in with the others during the independent and group reviews. The content reviewers will not know which packets are the validation packets. The NAEP experts have coded the validation packets and their coding will serve as a reference for determining how well the content reviewers are coding. The percent agreement between the four review teams’ group consensus coding on the validation packets and the reference coding as reliability evidence will be calculated. Project staff will also report the agreement of group consensus coding by the four review teams within each course title. The agreement statistic will be calculated using the same method.

The CCCA study’s methodology combines independent individual judgments with panel consensus processes. The first independent review is focused on applying the conceptual
understanding of the mathematics or reading knowledge and skills required in entry-level college courses by content reviewers with experience in teaching these types of courses and training and experience in the EPIC methodology of coding artifacts. The goal of the second, or group, review is to focus on adjudicating differences in coding of the packets completed during the independent review and the confirming the identification of exclusions in the NAEP framework objective statements. An additional focus is to review all KSAs that were identified in the packets but were not found in the NAEP frameworks.

The final result of this two-part review process will be a comprehensive list of prerequisite KSAs, answering the Board’s research question: what are the prerequisite KSAs in reading and mathematics to qualify for entry-level, credit-bearing courses that satisfy general education requirements.

Finally, a review is conducted by NAEP content experts to address the remaining research questions.

**Phase 3: Analysis and reporting**

Phase 3 includes processing and analyzing the judgments collected during the review of course artifacts by review teams, and preparing the data to be reported in ways that are directly responsive to research questions in accordance with the analysis plan specified within the Design Document. Standard statistical methods and metrics necessary will be employed to monitor and demonstrate validity and reliability, and both conceptual (information processing/document analysis) and technical (quantitative) analyses will be conducted. The CCCA study is structured to provide a fully crossed, three factor design to ensure that results can be reviewed in statistical generalizability analyses, which will allow us to evaluate the reliability of the study design.

Final results will include narrative summaries of the prerequisite knowledge, skills, and abilities in mathematics and reading. Summary analyses will also address all aspects of the CCCA study design (see *Illustration 1*). As project elements are completed, the appropriate sections of *Illustration 1* are shaded in dark gray. Project elements that have begun and are in progress are shaded in a lighter gray. Those project elements that have just begun have no shading in the diagram.
Illustration 2 displays an updated schedule of the CCCA study. As a result of feedback from the NAEP Advisory Panels, the schedule for content reviewer training has been changed to accommodate two sessions of training: an orientation session focusing on a holistic review of the packets, and a second training session after the reviewers are familiar with the packets. That second training will address the coding scheme, decision rules and use of the NAEP frameworks. Completed events are shaded black. Upcoming events and changes to the schedule are shaded.
PROGRESS UPDATE

Identification and Collection of Course Artifacts (Phase 1)

Table 1 contains the finalized list of entry-level courses to be included in the CCCA study.

Table 1: Course Titles Included in the CCCA Study

<table>
<thead>
<tr>
<th>Mathematics</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>College algebra</td>
<td>English literature</td>
</tr>
<tr>
<td>Finite math</td>
<td>Introduction to psychology</td>
</tr>
<tr>
<td>Introduction to calculus/precalculus</td>
<td>U.S. government/Introduction to political science</td>
</tr>
<tr>
<td>Statistics</td>
<td>U.S. history</td>
</tr>
</tbody>
</table>

The criteria for the course titles is as specified in the Design Document are:
• Entry-level
• Credit-bearing
• Frequently taken
• No college level prerequisites
• Not honors level
• Not remedial
• Not for majors

For each course in the study, a course packet will be considered complete if it includes the following:
• Syllabus
• Textbook excerpt
• Textbook table of content
• Non-textbook based assignment or exam from the first third of the course

Collecting mathematics course artifacts has been challenging, particularly identifying calculus courses that do not require prerequisites. In order to meet the requirement of collecting artifacts for 24 course packets per course type, we have slightly relaxed some of our representativeness criteria. One end result is that sufficient mathematics packets will be collected but larger institutions will slightly overrepresented.

Another factor identified in the artifact collection effort is that mathematics courses often use assignments from the textbook and do not create assignments outside of the textbook. To address this issue, the criteria for a complete course packet has been relaxed to allow an assignment to be textbook-based when the assignment is specifically identified in the syllabus. This change was also supported by the NAEP content advisors and has resulted in an improvement in the overall math packet quality.

Collecting artifacts for English Literature packets has also been challenging than expected due to the common college requirement that student take a writing composition class prior to taking English classes.

Tables 2 and 3 are summaries of the characteristics of representative institutions where courses have been submitted as candidates for packet creation. All courses met the criteria for inclusion in the study and the packets are sufficiently data-rich. These percentages should be considered preliminary, as the set of packets to be used in the study has not been completely finalized. More packets than are needed for the study have been collected in order to have sufficient overage and be able to make substitutions, if necessary. Note the number of completed packets in the “N-count” in the headers of the tables.
Table 2: Updated Institutional Characteristics of Sample for Mathematics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>College algebra (N 20)</th>
<th>Finite math (N 16)</th>
<th>Introduction to calculus (N 20)</th>
<th>Statistics (N 20)</th>
<th>Mathematics Overall (N 76)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Year</td>
<td>40%</td>
<td>25%</td>
<td>15%</td>
<td>25%</td>
<td>26%</td>
</tr>
<tr>
<td>4-Year</td>
<td>60%</td>
<td>75%</td>
<td>85%</td>
<td>75%</td>
<td>74%</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>75%</td>
<td>44%</td>
<td>40%</td>
<td>20%</td>
<td>54%</td>
</tr>
<tr>
<td>Medium</td>
<td>10%</td>
<td>13%</td>
<td>20%</td>
<td>25%</td>
<td>17%</td>
</tr>
<tr>
<td>Large</td>
<td>15%</td>
<td>44%</td>
<td>40%</td>
<td>55%</td>
<td>29%</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>60%</td>
<td>56%</td>
<td>65%</td>
<td>50%</td>
<td>58%</td>
</tr>
<tr>
<td>Private not-for-profit</td>
<td>40%</td>
<td>44%</td>
<td>35%</td>
<td>50%</td>
<td>42%</td>
</tr>
<tr>
<td><strong>Geographic Region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>West</td>
<td>10%</td>
<td>6%</td>
<td>15%</td>
<td>15%</td>
<td>12%</td>
</tr>
<tr>
<td>Midwest</td>
<td>30%</td>
<td>13%</td>
<td>20%</td>
<td>30%</td>
<td>24%</td>
</tr>
<tr>
<td>East</td>
<td>10%</td>
<td>19%</td>
<td>10%</td>
<td>25%</td>
<td>16%</td>
</tr>
<tr>
<td>Southeast</td>
<td>30%</td>
<td>25%</td>
<td>30%</td>
<td>20%</td>
<td>26%</td>
</tr>
<tr>
<td>Southwest</td>
<td>20%</td>
<td>38%</td>
<td>25%</td>
<td>10%</td>
<td>22%</td>
</tr>
</tbody>
</table>
Table 3: Updated Institutional Characteristics of Sample for Reading

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>English literature (N = 17)</th>
<th>Introduction to psychology (N = 20)</th>
<th>U.S. government (N = 20)</th>
<th>U.S. history (N = 20)</th>
<th>Reading Overall (N = 80)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Program type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Year</td>
<td>25%</td>
<td>35%</td>
<td>15%</td>
<td>15%</td>
<td>23%</td>
</tr>
<tr>
<td>4-Year</td>
<td>75%</td>
<td>65%</td>
<td>85%</td>
<td>85%</td>
<td>78%</td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>55%</td>
<td>60%</td>
<td>50%</td>
<td>55%</td>
<td>55%</td>
</tr>
<tr>
<td>Medium</td>
<td>15%</td>
<td>20%</td>
<td>25%</td>
<td>25%</td>
<td>21%</td>
</tr>
<tr>
<td>Large</td>
<td>30%</td>
<td>20%</td>
<td>25%</td>
<td>20%</td>
<td>24%</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public</td>
<td>60%</td>
<td>60%</td>
<td>50%</td>
<td>50%</td>
<td>55%</td>
</tr>
<tr>
<td>Private not-for-profit</td>
<td>40%</td>
<td>40%</td>
<td>50%</td>
<td>50%</td>
<td>45%</td>
</tr>
<tr>
<td><strong>Geographic Region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>West</td>
<td>0%</td>
<td>20%</td>
<td>20%</td>
<td>15%</td>
<td>21%</td>
</tr>
<tr>
<td>Midwest</td>
<td>40%</td>
<td>25%</td>
<td>30%</td>
<td>25%</td>
<td>30%</td>
</tr>
<tr>
<td>East</td>
<td>15%</td>
<td>20%</td>
<td>20%</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td>Southeast</td>
<td>15%</td>
<td>30%</td>
<td>20%</td>
<td>25%</td>
<td>23%</td>
</tr>
<tr>
<td>Southwest</td>
<td>30%</td>
<td>5%</td>
<td>10%</td>
<td>10%</td>
<td>6%</td>
</tr>
</tbody>
</table>

**NAEP Advisory Panel Review of Course Artifacts (Phase 1)**

A preliminary review of packets was completed at two NAEP Advisory Panels, one in reading and one in mathematics, both conducted in June of 2013. The table below provides an overview of the type and number of packets to be reviewed at each CCCA review session.
Table 2: Allocation of Packets Across CCCA Events

<table>
<thead>
<tr>
<th>Total number of course packets and purpose</th>
<th>NAEP Advisory Panel Pre coded by NAEP experts</th>
<th>Coded by content reviewers and alternate reviewers</th>
<th>Independent Review Coded by content reviewers and alternate reviewers</th>
<th>Group Content Review Reviewed in content teams</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Training packets</td>
<td>4 Total, 2 for math and 2 for reading</td>
<td>4 Total, 2 for math and 2 for reading</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4 Qualifying packets</td>
<td>4 Total, 2 for math and 2 for reading</td>
<td>4 Total, 2 for math and 2 for reading</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>16 Validation packets</td>
<td>16 Total, 8 for math and 8 for reading</td>
<td>Not coded in training</td>
<td>16 Total, 8 for math and 8 for reading</td>
<td>Depends on the number of packets that need to be reviewed during the group review process</td>
</tr>
<tr>
<td>160 Operational packets</td>
<td>Not pre-coded by NAEP experts</td>
<td>Not coded in training</td>
<td>160 Operational packets</td>
<td>Depends on the number of packets that need to be reviewed during the group review process</td>
</tr>
<tr>
<td>TOTAL PACKETS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>184 Total</td>
<td>24 Total</td>
<td>8 Total</td>
<td>176 Total</td>
<td></td>
</tr>
<tr>
<td>92 for math</td>
<td>12 for math</td>
<td>4 for math</td>
<td>88 for math</td>
<td></td>
</tr>
<tr>
<td>92 for reading</td>
<td>12 for reading</td>
<td>4 for reading</td>
<td>88 for reading</td>
<td></td>
</tr>
</tbody>
</table>

The primary goals of the NAEP Advisory Panel meetings were: (1) to obtain comprehensive guidance regarding the process of content review training; and (2) to establish benchmarks for coding a subset of the packets. Twelve (12) packets were reviewed in the NAEP Advisory meeting – 8 validation packets, 2 training packets, and 2 qualifying packets for both reading and mathematics. The NAEP experts were briefed on the proposed training process and provided with an initial set of training materials that included a coding schema, decision rules and a reference sheet. The request was that they begin coding the 12 packets using the existing guidance. As they coded, they provided feedback and suggested new or different approaches regarding the process and decision rules.

The outcomes from each of the Advisory Panel meetings were:

1. Guidance for making improvements to the training process
   a. Training in two separate sessions – the first for a holistic review of the entire packet without reference to the NAEP frameworks and then a second session to train on how to review using the NAEP frameworks
   b. Training in small groups of four instead of a large group
   c. Deeper understanding of the complexity of the task and advice on how to make it easier for the reviewers
2. Updated, and simplified, coding schema
3. Updated, and simplified, decision rules
4. Updated, and simplified, reference sheets
5. Guidance for establishing the criteria for sufficiency of a packet
6. Guidance for establishing a review procedure for course packets
7. Guidance for identifying the criteria for qualities of good training examples
8. Guidance for estimating time to complete review tasks independently and in group
9. Benchmarks for 12 packets and recommended usage for one of three possible uses – validation, training and/or qualifying.
Most of the guidance from the NAEP experts was integrated into the development of the training materials and the planned implementation of the CCCA study. In order for guidance to be accepted, it had to be feasible, not be in conflict with other design factors in the Design Document and support comparability with the JSS and JTPCS study findings.

Coding Schema and Review of the Course Artifacts (Phase 2)

Using the Design Document as the roadmap, phase 2 activities are well underway.

Development of the content reviewer coding instruments for both mathematics and reading, based on the coding schema and decision rules in the design document specifications, was completed. Both the coding schema and decision rules were thoroughly reviewed by the NAEP Advisory Panel experts, in conjunction with the review of the sample course title packets. As a result of that work, the recommendation was to simplify the coding schema from six levels to three levels. The coding instrument has been updated to reflect the three level coding schemas. The change from six levels of coding to three levels will not impair comparability with JSS or JTPCS findings.

Table 4: Simplified Coding Schema

<table>
<thead>
<tr>
<th>Applicability and Importance</th>
<th>Design Document Coding Schema</th>
<th>Post NAEP Advisory Panel Coding Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>1—KSA is NOT applicable to this course</td>
<td>2—KSA is NEW content taught in this course</td>
<td>1—KSA is NOT PREREQUISITE to this course.</td>
</tr>
<tr>
<td>3—KSA is PREREQUISITE for this course and is NOT IMPORTANT. Although a prerequisite, possessing this KSA will make little difference on course outcomes.</td>
<td>4—KSA is PREREQUISITE for this course and is MINIMALLY IMPORTANT. This KSA is a prerequisite, which if possessed, is likely to result in better course outcomes.</td>
<td>2—KSA is PREREQUISITE for this course.</td>
</tr>
<tr>
<td>5—KSA is PREREQUISITE for this course and is IMPORTANT. Without this KSA, students will struggle with the course.</td>
<td>6—KSA is PREREQUISITE for this course and is VERY IMPORTANT. Without this KSA students are not prepared for and will be unlikely to complete this course.</td>
<td>3—KSA is PREREQUISITE for this course and is IMPORTANT.</td>
</tr>
</tbody>
</table>
While the task of reviewing mathematics packets is very different from the reading task, it was clear to both panels that the task of content review needed to be simplified to assure manageability of the task and a reasonable time commitment for review activities. Changes are not expected to have any negative effect on the content maps that will eventually be created for review by the NAEP experts at the NAEP comparison meetings in Fall 2013. Both panels endorsed the practice of reviewing the entire packet in a holistic manner to get familiar with the packet, the task, and the overall variation among the packets with the goal of identifying the most relevant prerequisite KSAs in a summary manner. This approach ensures that the content reviewers have the big picture in mind and that the potential risk of bias toward only identifying NAEP KSAs is mitigated.

The overarching guidance from both of the NAEP Advisory Panels was that the content reviewer task of reviewing 28 packets was challenging because of complexity and time-consuming because of the number of packets to be reviewed. The advice was to review the entire training and review process with the goal of simplifying wherever possible.

Several additional changes, and improvements are being considered. The estimated time needed to review each individual packet has increased (informed by the NAEP content advisors, and so an increase in stipend and an increase in the length of the group review session by half a day are under discussion.

*Design Document and Analysis and Reporting (Phase 3)*

Preparation for analysis and the final reporting have begun with the majority of the effort in data management. Staff are working with sample data and testing to ensure that accurate data collection protocols and routines of effective quality control, data cleaning procedures and data storage/security protocols are in place and use.

The final report is also underway. The table of contents has been established and preliminary table shells have been drafted. The final report will be written in sections and reviewed throughout the rest of the study.
Attachment E-2

Job Training Program Content Analysis Final Report

In October 2011, the Governing Board began work with WestEd and its subcontractor, the Education Policy Improvement Center (EPIC), to conduct follow-up research relative to the NAEP preparedness judgmental standard setting (JSS) research, wherein panelists reviewed NAEP questions and made judgments about the content knowledge needed by minimally prepared students. The research results from this project are intended to supplement the JSS research findings by providing a clearer understanding of the knowledge and skills required for entry- and exit-level coursework in designated occupational programs. By reviewing course artifacts such as syllabi, text books, and assignments, this study will help to determine if the knowledge, skills, and abilities (KSAs) required of students in the training programs are appropriately represented by the borderline preparedness descriptions (developed in the JSS research), by all the items on the 2009 NAEP, and by the 2009 NAEP items in the scale score ranges identified by panelists in the JSS research project.

The executive summary for the final report was included with the May 2013 COSDAM materials. The full report is now available online at:

http://www.nagb.org/what-we-do/preparedness-research/types-of-research/jss.html
Attachment E-3
Phase 2 Academic Preparedness Research Plans

Continued research plans call for NAEP-SAT, NAEP-ACT, and NAEP-EXPLORE statistical linking studies, more research partnerships with states, analysis of course content prerequisites for job training programs and freshman college courses, and efforts to partner with experts in military occupational training. A summary of each proposed research study follows. At the November 2012 Board meeting, COSDAM began discussion on these research plans.

National and State Statistical Linking Studies with the SAT and with the ACT

In 2013, the Governing Board will partner again with the College Board, as it did in 2009, to conduct a statistical linking study at the national level between NAEP and the SAT in reading and mathematics. Through a procedure that protects student confidentiality, the SAT records of 12th grade NAEP test takers in 2013 will be matched, and through this match, the linking will be performed. A similar study at the national level is planned in partnership with ACT, Inc.

In addition, the state-level studies, begun in 2009 with Florida, will be expanded in 2013. Again using a procedure that protects student confidentiality, the postsecondary activities of NAEP 12th grade test takers in the state samples in partner states will be followed for up to five years using the state longitudinal data bases. Five states will be partners in these studies: Florida, Illinois, Massachusetts, Michigan, and Tennessee. These studies will examine the relationship between 12th grade NAEP scores and GPA, placement into remedial versus credit-bearing courses, and scores on admissions and placement tests. Data sharing agreements are in development for each state partner.

August 2013 Update: No updates at this time.

In 2013, linking studies between 8th grade NAEP in reading and mathematics and 8th grade EXPLORE, a test developed by ACT, Inc. that is linked to performance on the ACT, are planned with partners in two states, KY and TN. The objective is to determine the feasibility of identifying the point on the NAEP scales that indicate students are “on track” for being academically prepared for college and job training by 12th grade. As a foundation for the linking study, content alignment studies between 8th grade NAEP reading and mathematics and 8th grade EXPLORE would also be conducted as a part of the planned partnership with Act, Inc.

August 2013 Update: No updates at this time.

The Governing Board is conducting a procurement (1) to design a comprehensive and multi-method evaluation of the grade 12 NAEP frameworks and item pools in both reading and mathematics as measures of academic preparedness for college and job training; and (2)

E16
based on the evaluation, to produce specific recommendations for changes that may be needed to further refine 12th grade NAEP in reading and mathematics as a measure of academic preparedness for college and to determine the extent to which changes would be needed to make 12th grade NAEP in reading and mathematics a valid measure of academic preparedness for entry into job training programs that require at least three months of post-secondary training, but not a bachelor’s degree in college.

The review of the 12th grade reading and mathematics frameworks by Achieve, Inc. in 2005 and 2006 led to changes in the frameworks for the 2009 assessments intended to measure 12th grade academic preparedness for college and job training. The content alignment studies between 12th grade NAEP reading and mathematics and the SAT and ACT college admissions tests in reading and mathematics tests found a high degree of overlap in content widely recognized as representing academic preparedness for college. The content alignment study with WorkKeys, as well as the Judgmental Standard Setting studies for job training, surfaced questions about the capacity of the current 12th grade NAEP to measure academic preparedness for job training. The planned evaluation is part of the continuing program of preparedness validity research.

In this procurement, the Board seeks innovative, practicable design proposals for evaluations that will provide the foundation needed to make valid statements about academic preparedness.

**August 2013 Update:** The contract has been awarded to HumRRO. A kickoff meeting has been conducted. The project is now just getting underway.

Reporting on academic preparedness for college and job training is a challenging and important new direction for NAEP. Hence, the Governing Board is also conducting a procurement to seek proposals for research designs and studies that are feasible. The objective of the research is to advance the Governing Board’s efforts to identify locations on the 12th grade NAEP reading and mathematics scales that represent the knowledge and skills to qualify for training in various occupations.

**August 2013 Update:** The procurement process did not result in a contract award.
Attachment E-4
Overview of the Types of NAEP Preparedness Research

As part of the ongoing updates to COSDAM, the following is a summary of each research study category from phase 1 of the Board’s program of research for reporting academic preparedness.

### Content Alignment Studies

Content alignment studies are a foundation for the trail of evidence needed for establishing the validity of preparedness reporting, and are, therefore, considered a high priority in the Governing Board’s Program of Preparedness Research. The alignment studies will inform the interpretations of preparedness research findings from statistical relationship studies and help to shape the statements that can be made about preparedness. Content alignment studies were recommended to evaluate the extent to which NAEP content overlaps with that of the other assessments to be used as indicators of preparedness in the research.

A design document was developed by Dr. Norman Webb for the NAEP preparedness research alignment studies, and this design was implemented for the studies of the 2009 NAEP with the SAT and ACUPLACER in reading and mathematics. This design, with minor modifications, has also been used for the alignment of the 2009 NAEP with WorkKeys tests in these subject areas.

Content alignment studies for the first phase of the Board’s Program of Preparedness Research have been completed for NAEP in reading and in mathematics with WorkKeys, the SAT, and ACCUPLACER. In addition, a content alignment study was designed and conducted by ACT for the ACT and NAEP in reading and mathematics before the content alignment design document was developed.

### Studies to Establish Statistical Relationships

Highest priority has generally been placed on these studies. Currently, two main sets of studies have been conducted under this heading. One set addresses statistical linking of NAEP with other assessments, and the other set examines longitudinal data for NAEP examinees.

**For statistical linking,** there has been a study to relate SAT scores in reading and in mathematics to the national sample of NAEP scores for grade 12. The objective was to provide a statistical linking of SAT and NAEP scores for all students in the 2009 grade 12 NAEP who had taken the SAT by June 2009. ETS staff reported that the match rate of approximately 33% of NAEP scores to SAT scores compares favorably to the national SAT participation rate of approximately 36% of public school students. The final sample used for linking the NAEP reading and SAT critical reading included approximately 16,200 students. For NAEP and SAT mathematics, the linking sample included approximately 15,300 students.

**For longitudinal data,** a series of analyses were conducted to examine statistical relationships for Florida’s NAEP examinees. NAEP’s 2009 state-representative sample of Florida 12th graders was used to match NAEP scores for reading and mathematics to student scores on several tests.
collected by the Florida Department of Education (FLDOE). The data sharing agreement with FLDOE provides access to scores for the SAT, ACCUPLACER, and WorkKeys. Additionally, ACT, Inc. has given permission to the Florida Department of Education to share ACT scores with the Governing Board for purposes of conducting the grade 12 preparedness research. A plan to allow for electronic transfer of data was developed to keep secure the identity of students, consistent with the NAEP legislation, FLDOE requirements, and requirements of each assessment program.

Records for roughly half of the Florida grade 12 NAEP examinees in 2009 could be matched to an ACT score and half to an SAT score. This match rate is consistent with other data for Florida students. The match of WorkKeys scores to the total 2009 state NAEP sample of 12th graders was only about 6%. FLDOE reported that around 89,300 Florida 12th graders were enrolled in vocational-technical programs in school year 2008-09. The match of WorkKeys examinees to NAEP examinees was not sufficient to warrant additional analyses for the 2009 cycle. The state of Florida has only recently implemented the testing of high school students in vocational programs with the WorkKeys exam, and we anticipate that the number of examinees will increase in subsequent years.

### Judgmental Standard Setting Studies

A series of judgmental standard setting studies was planned to produce preparedness reference points on the NAEP scale for entry into job training programs and for placement in college credit-bearing courses. Within this category of studies, the Technical Panel for 12th Grade Preparedness Research placed highest priority on the judgmental studies related to preparedness for job training programs in 5-7 exemplar jobs. This priority is largely related to the paucity of national data available for statistical studies in these areas. The Governing Board has not assumed that academic preparedness for college and for job training are the same. Rather, our studies are aimed at determining the level of performance on NAEP that represents the reading and mathematics knowledge and skills needed to qualify for job training programs for each of the occupations included in our research studies and for placement in credit-bearing college courses that fulfill general education requirements for a bachelor’s degree.

In order to maximize the standardization of judgmental standard setting (JSS) studies within and across post-secondary areas, a design document was developed to specify the number of panelists, the eligibility criteria for panelists, the procedures for drafting and finalizing borderline performance descriptions, the methodology to be implemented, feedback to be provided, key aspects to be evaluated, and reports to be produced. The methodology and basic procedures specified for the design of these studies were those implemented for the achievement levels-setting process for the 2006 grade 12 economics NAEP and for the 2009 science NAEP for grades 4, 8, and 12.

The five exemplar jobs approved by COSDAM for inclusion in these studies are as follows:

1. automotive master technicians
2. computer support specialists
3. heating, ventilation, and air conditioning technicians
4. licensed practical nurses
5. pharmacy technicians

A pair of replicate panels with 10 panelists each was convened for each subject and post-secondary area for a total of 24 operational panels.

**Higher Education Survey**

A survey of two-year and four-year post-secondary institutions was conducted in Fall 2011 to gather information regarding (1) the placement tests used and (2) the cut scores on those tests in reading and mathematics below which need was indicated for remedial/developmental courses in reading and mathematics, and at or above which placement in credit-bearing entry level courses was indicated. The sample of accredited postsecondary education institutions was nationally representative. A weighted response rate of 81% was achieved.

**Benchmarking Studies**

Benchmarking studies in the preparedness research context are studies in which NAEP is administered to groups of interest, e.g., college freshmen enrolled in credit-bearing college level courses that fulfill general education requirements for a four-year degree without the need for remediation. Determining the average NAEP performance of this group would then provide a “benchmark” score that can be considered as one of the reference points on the NAEP scale. A benchmarking study in combination with reference points from other studies in the Program of Preparedness Research can assist the Board in determining the areas of the NAEP scale that indicate preparedness. A benchmarking study of Texas college freshmen was planned, and it had the support of the Texas Commissioner of Higher Education and the cooperation of nine Texas higher education institutions. A small scale pilot study to evaluate the feasibility of the study design was implemented.

The Governing Board and the National Center for Education Statistics (NCES) collaborated on the implementation of this small scale pilot study, which was carried out by Westat, the NAEP sampling and administration contractor to NCES. The data collection phase for the pilot ended on October 15, 2010. Of the eligible sample of 1,234 students, 255 actually attended a NAEP session, for an overall response rate of 20.7 percent. As announced at the November 2010 meeting of COSDAM, NCES, Westat, and Governing Board staff met to discuss alternatives. Board staff decided that we will not proceed to the operational phase of this study due to low participation rates and the lack of feasible alternatives to increase participation.

No additional benchmarking studies are planned for the 2009 NAEP preparedness research.

**OVERVIEW OF REFERENCED ASSESSMENTS**

For additional background information, the following list presents a brief description of the assessments that the Technical Panel on 12th Grade Preparedness Research recommended for analysis in NAEP preparedness research. Many of these assessments are the primary focus of the proposed content alignment studies and statistical relationship studies. In each case, only the
mathematics and reading portions of the assessments are the targets for analysis, although analyses with the composite scores may be conducted.

- **ACCUPLACER** – ACCUPLACER is a computer adaptive test used for college course placement decisions in two-year and four-year institutions. It is produced by the College Board and includes assessments of sentence skills, reading comprehension, arithmetic, elementary algebra, college level math, and written essays.

- **ACT** – The ACT assessment is a college admissions test used by colleges and universities to determine the level of knowledge and skills in applicant pools, including reading, English, and mathematics tests. ACT has *College Readiness Standards* that connect reading or mathematics knowledge and skills and probabilities of a college course grade of “C” or higher (75%) or “B” or higher (50%) with particular score ranges on the ACT assessment.

- **ACT WorkKeys** – WorkKeys is a workplace focused set of tests that assess knowledge and skills in communication (business writing, listening, reading for information, writing) as well as problem solving (applied technology, applied mathematics, locating information, observation). There is also an interpersonal skills section of WorkKeys.

- **COMPASS** – ACT Compass is a computer-adaptive college placement test. It is produced by ACT and includes assessments of Reading, Writing Skills, Writing Essay, Math, and English as a Second Language.

- **SAT** – The SAT reasoning test is a college admissions test produced by the College Board. It is used by colleges and universities to evaluate the knowledge and skills of applicant pools in critical reading, mathematics, and writing. The College Board has provided SAT score data to be used in research studies to establish a statistical relationship between the SAT and NAEP.
# National Assessment Governing Board

## Reporting and Dissemination Committee

**August 2, 2013**  
10:00 a.m.-12:30 p.m.

## AGENDA

<table>
<thead>
<tr>
<th>Time</th>
<th>Item</th>
<th>Attached to</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00 – 10:40 am</td>
<td>Review of Board Policy and Guidelines on Reporting, Release, and Dissemination of NAEP Results <em>Stephaan Harris and Larry Feinberg, NAGB Staff</em></td>
<td>Attachment A</td>
</tr>
</tbody>
</table>
| 10:40 – 11:15 am| Parent Outreach Activities  
  - Update on Parent Summit  
  - **ACTION:** Parent Outreach Plan  
    *Stephaan Harris, NAGB Staff*  
    *Amy Buckley, Reingold Communications* | Attachment B |
| 11:15 – 11:45 am| Possible New Formats for NAEP Reporting  
  *Arnold Goldstein, NCES Staff* | Attachment C |
| 11:45 am – 12:00 noon | **ACTION:** Release Plan for NAEP 2013 Reading and Mathematics Report Cards  
  *Stephaan Harris, NAGB Staff* | Attachment D |
| 12:00 – 12:20 pm| Plans for Focused NAEP Reports  
  *Grady Wilburn, NCES Staff*  
  *Larry Feinberg, NAGB Staff* | Attachment E |
| 12:20 – 12:30 pm| **Information Items:**  
  - Review of Recent NAEP Release: Long-Term Trends  
  - Projected Schedule for Future NAEP Reports | Attachment F, G |
Review of Board Policy and Guidelines on Reporting, Release, and Dissemination of NAEP Results

The Reporting and Dissemination Committee is continuing an ongoing discussion on its role in the reporting, release, and dissemination of National Assessment of Educational Progress (NAEP) results. Specifically, the Committee has expressed interest in examining how their role might change, while preserving the distinct and legal roles and responsibilities of the Governing Board, which sets policy for NAEP, and the National Center for Education Statistics (NCES), which assesses the students, analyzes the data, and uses the findings to draft NAEP reports. The Governing Board’s NAEP reporting, release, and dissemination policy (in full below), adopted in 2006, was used as a starting point for this discussion.

The Committee desires more influence and input into NAEP report development and content. Committee Chair Alonso has asked members to propose strategies for how the Board can provide input at the beginning, or conception, phase of report development, rather than solely providing feedback on a late-stage draft report. The goal is to have input at a “big picture” level rather than to provide edits to text, graphics, and pictures on the final drafts of NAEP reports.

Being mindful of a changing and competitive media landscape and the need to make NAEP relevant and meaningful to a diverse group of audiences, the Committee is also exploring how NAEP data can best be featured and distributed via Report Cards and electronic tools.

NCES has outlined the review process and timeline of some major NAEP Report Cards to inform the discussion (this document is in full below). Governing Board staff proposes a few updated discussion questions to help guide the Committee’s ongoing conversation:

1. Given time constraints posed by the development of NAEP Reading and Mathematics reports in particular, what are effective ways the Board can provide early feedback to allow for necessary changes and modifications while keeping on a tight schedule?

2. How best can NAEP findings be displayed online and in reports to present rich and technical data in easily understandable ways? Are there supplemental materials or resources that could be developed in tandem to engage audiences unfamiliar with NAEP?

3. What are some new or improved strategies to employ with NAEP release events, which are now primarily webinars, to better reach various media and stakeholder audiences?
Overview of NAEP Six-Month Reporting:
Development and Review Process

The Assessment Division of the National Center for Education Statistics (NCES) reports and disseminates results for the National Assessment of Educational Progress (NAEP). An extensive report planning and review process is put in place to ensure that NAEP achievement data are of the highest quality and reported on a timely basis. Results for reading and mathematics at grades 4 and 8 must be transmitted to the Governing Board within six months of the end of data collection. For six-month reporting in particular, the timeline for moving the reports through development to Governing Board acceptance is very tight. Moving these reports through the major planning and review phases must be accomplished within ten weeks.

In order to meet the mandated six-month reporting deadline for the 2013 Reading and Mathematics report cards, it is important for all stakeholders involved in the NAEP planning and review process to provide substantial feedback at report development/planning phases and to review the developed report in an expedited fashion.

At the upcoming Governing Board meeting, a reporting timeline will be shared and points for Governing Board input on the report content will be discussed.
The Nation’s Report Card™ informs the public about the academic achievement of elementary and secondary students in the United States. Report cards communicate the findings of the National Assessment of Educational Progress (NAEP), the only continuing and nationally representative measure of achievement in various subjects over time. The Nation’s Report Card compares performance among states, urban districts, public and private schools, and student demographic groups.

Introduction

NAEP collects data through representative-sample surveys and reports fair and accurate information on academic achievement to the American public. By law (P.L. 107-110, as amended by P.L. 107-279), NAEP is administered by the Commissioner of the National Center for Education Statistics (NCES) under policy set by the National Assessment Governing Board (“the Governing Board”), a bipartisan, independent policymaking body.

According to the statute, the Governing Board shall exercise “independent judgment, free from inappropriate influences and special interests” and in the exercise of its responsibilities, “shall be independent of the Secretary and the other offices and officers of the Department [of Education].” Among the responsibilities specifically delegated to the Governing Board are: (1) “develop guidelines for reporting and disseminating [NAEP] results”; (2) “take appropriate actions needed to improve the form, content, use, and reporting of [NAEP] results”; and (3) “plan and execute the initial public release of [NAEP] reports.”

To carry out these responsibilities, the Governing Board hereby adopts policy principles and guidelines for the reporting, release, and dissemination of The Nation’s Report Card.

As outlined in the appendix, this policy defines The Nation’s Report Card as, and applies to, the initial reporting of NAEP results from national, state, and trial urban district assessments (TUDA), and to other special reports or studies authorized by the National Assessment Governing Board, including printed reports and the initial release Web site.
Delineation of NAEP Reporting, Release, and Dissemination Responsibilities

The NCES Commissioner, under Governing Board policy guidance, is responsible for administering the assessment, ensuring the technical soundness and accuracy of all released data, preparing NAEP reports, and presenting NAEP results.

In addition to setting policy, Governing Board is responsible for ensuring policy compliance of Governing Board-authorized NAEP reports, determining their respective dates of release, and planning and executing the initial public release of NAEP results.

Part I: Report Preparation and Content

Policy Principles


2. The primary audience for The Nation’s Report Card is the American public.
   a. All reports shall be written in language appropriate for an audience of the interested general public, the majority of whom are unlikely to have a technical understanding of education statistics or assessment.

3. The Nation’s Report Card shall report data objectively, accurately, clearly, and fairly, in accordance with NCES data quality standards. Results shall be insulated from ideological and other special interests.
   a. The Nation’s Report Card shall include straightforward presentations of data. Reports may suggest correlations, but should not conclude cause-and-effect relationships. Any interpretation of results must be strongly supported by NAEP data.
   b. The Nation’s Report Card and its Web site may include references and links to the National Assessment Governing Board Web site, NCES Web site, and the NAEP Validity Studies Panel. Non-NAEP materials and links to non-NAEP resources shall not be included in initial release documents, with the exception of relevant federal and state government information, such as NCES surveys and other district, state, national, or international testing programs.
   c. To improve public understanding of results, The Nation’s Report Card should contain information about Governing Board-approved NAEP contextual variables and subject-specific background information—as outlined in the
Background Information Framework for the National Assessment of Educational Progress (adopted by the National Assessment Governing Board, 8/1/03)—when available and reliable. Reports may also contain other contextual information from trustworthy sources outside of the NAEP program, such as expenditures per pupil, student/teacher ratios, and student enrollment.

4. In accordance with the law, The Nation’s Report Card shall include results for the nation; states and school districts, when collected in conjunction with specific NAEP programs, respectively; and school types, disaggregated by subgroup whenever reliable. Subgroup results shall be prominently positioned to facilitate public review but shall not be used to adjust findings.
   a. Disaggregated subgroup data should be accompanied by information about demographic changes in the student population assessed.
   b. Results for states and school districts may be presented in alphabetical or rank order, accompanied by appropriate language to make the public aware of any data comparison limitations.
   c. Data shall be publicly released on inclusion and accommodation rates for all NAEP samples, including national, state, district, and school type. Results for students with disabilities and English language learners shall be presented separately.

5. The Nation’s Report Card shall report results by Governing Board-adopted achievement levels, average scale scores, and percentile distributions. Trend information shall be an important part of reports unless comparable and reliable data are not available.
   a. Reports shall contain clear explanations of achievement levels, including item maps and sample test questions and answers to illustrate what students in each grade assessed should know and be able to do at each achievement level.

6. All NAEP data determined by the NCES Commissioner to be valid and reliable shall be made available on the World Wide Web at the time of initial public release, except for data from limited special purpose samples and pilot studies. A separate, dedicated Web site aimed at a broad public audience – http://nationsreportcard.gov – shall be utilized for initial public releases.
   a. All released NAEP data shall be subject to NCES quality control procedures to ensure accuracy and completeness.
   b. At least one block of released NAEP questions shall be posted on the World Wide Web for each subject and grade for which results have been collected.
c. Concise information on test content, methodology, performance standards, and scoring shall be included in all NAEP reports. More extensive material on these topics should be readily accessible on the World Wide Web.

7. Results of special studies authorized by the Governing Board will be reported after careful review of information quality and statistical validity. These shall be treated as initial public releases of The Nation’s Report Card, and shall be subject to NCES quality control procedures and Governing Board policies.

8. The Governing Board shall adopt general guidelines to inform the development of The Nation’s Report Card and its Web site, and may set additional specifications for particular reports.

9. The Governing Board shall review the format and content of initial releases, including Web pages, to ensure compliance with Governing Board policy.

a. The Nation’s Report Card shall contain a description of the policymaking roles and responsibilities of the National Assessment Governing Board, including a list of current Governing Board members, their affiliations, and regional locations.

Part II: Public Release of NAEP Results

Policy Principles

1. Release activities shall be planned and executed by the National Assessment Governing Board. The Governing Board shall determine the release date, time, embargo policies, and manner of release for The Nation’s Report Card, as covered by this policy.

a. After the Governing Board has approved the final draft of The Nation’s Report Card, including the pages that will be made available through the initial release Web site, the Chairman of the Reporting and Dissemination Committee, on behalf of the Governing Board, shall determine the date of the initial public release, in consultation with the Chairman and Executive Director of the National Assessment Governing Board and the NCES Commissioner.

b. The initial release shall be completed within 30 days of approval of the final draft of The Nation’s Report Card. In setting that release date, attention will be paid to balancing the priorities of an expeditious release with provision for adequate planning time, given the scheduling circumstances of the various parties involved.
c. Prior to the initial public release, NAEP results may be provided on an embargoed basis to federal, state, and TUDA-district officials and members of the press.

2. The Governing Board shall be responsible for organizing and conducting the release event and related activities.

   a. A release plan shall be adopted by the Governing Board for each report. Elements of the plan may include issuance of a press release, a press conference and/or Web-based announcement, distribution of summary findings and graphics, time period for the initial public release phase of http://nationsreportcard.gov, and other related activities.

   b. The official press release announcing NAEP results shall be issued by the Governing Board. Accompanying statements from the Governing Board’s Executive Director or Governing Board members may also be issued.

   c. At the press conference or other event for release of NAEP results, the NCES Commissioner or his/her designee shall present major data findings, accompanied by a written statement. The National Assessment Governing Board shall select members to provide individual commentary on the meaning of results. In addition, the Governing Board may invite other officials or experts to comment on the significance of the results in accordance with the approved release plan.

   d. At press conferences, questions from the audience shall be limited to accredited members of the media. At other public release events, the Governing Board shall determine who may attend and ask questions or comment.

3. The Nation’s Report Card shall seek to encourage wide public attention to NAEP results and clear understanding of their meaning and significance.

   a. Video materials may be prepared to accompany the release. These shall be clearly identified as having been provided by the Governing Board or NCES of the U.S. Department of Education. The video materials may only contain sound bites, background footage, and other information for journalists to develop their own stories.

4. Release procedures shall underscore the credibility of The Nation’s Report Card and encourage the participation of schools, school districts, and states in NAEP.

   a. NAEP data in statements distributed at The Nation’s Report Card initial public release events shall be checked for accuracy by NCES.
5. *The Nation’s Report Card* releases shall be clearly separated from any ideological or other special interests.

   a. Activities related to the initial public release of *The Nation’s Report Card* shall not be used to disseminate any materials unrelated to NAEP. No materials of any kind may be distributed at an initial release event without the prior approval of the Governing Board.

6. The National Assessment Governing Board will cooperate with the NCES Commissioner in the release of technical reports, working papers, and secondary analyses not covered by the policy.

7. The Governing Board will develop a reporting schedule each year for upcoming NAEP assessments based on data review and report production plans that are provided and updated by NCES.

**Part III: Dissemination and Outreach**

**Policy Principles**

1. Information from *The Nation’s Report Card* shall be disseminated through the media, the World Wide Web, and special publications and materials. Efforts shall be made to develop widespread public awareness of NAEP data and their meaning and of the value of *The Nation’s Report Card* to the nation and participating jurisdictions.

   a. NAEP results shall be available in both printed and electronic form, including on *The Nation's Report Card* Web site, at the scheduled time of release and in the permanent record.

   b. To build public awareness of *The Nation’s Report Card*, the home page of the initial release Web site shall remain on-line and include links to previous releases. This homepage shall link to respective pages found on the NAEP Web site.

2. To build understanding of *The Nation’s Report Card* and the data it reports, other information about NAEP may be disseminated at the time of the initial release and on a continuing basis.

   a. Informational materials accompanying results shall explain the mission and value of *The Nation’s Report Card* in clear and compelling terms.

3. *The Nation’s Report Card* and supplementary NAEP materials shall be made available through a wide network of education, business, labor, civic, and other interested groups and to policy makers and practitioners at all levels of education and government.
a. The Nation’s Report Card shall be distributed promptly to governors and chief state school officers, as well as to superintendents of TUDA districts. The reports shall be posted on the World Wide Web immediately at the time of initial release, with printed copies available to the public upon request.

b. Notification of upcoming releases shall be widely disseminated. Schools and school districts participating in NAEP samples shall be provided with information on how to access reports electronically and obtain printed copies upon release.

c. NCES and Governing Board staff shall encourage national and state organizations that are interested in education to disseminate NAEP results to their members.

d. The NCES Commissioner and staff, Governing Board members and staff, and NAEP State Coordinators are encouraged to increase awareness and understanding of NAEP among the public, educators, and government officials. They are encouraged to speak about the NAEP program to a variety of audiences; at meetings and conferences of national, state, and local organizations; on radio and television; and to writers for magazines and newspapers and other members of the media.

e. Talking points on key data findings shall be developed for each release and distributed to Governing Board members.

4. A variety of materials shall be developed, appropriate to various audiences, to carry out NAEP dissemination. Key audiences for these materials shall include the interested general public, policymakers, teachers, administrators, and parents.

5. Detailed data on cognitive results, Governing Board-approved contextual variables, and subject-specific background information (as outlined in Part I, Policy Principle 3, Item C) shall be made readily available through the World Wide Web to all those wishing to analyze NAEP findings, subject to privacy restrictions. Additional restricted data shall be available for scholarly research, subject to NCES licensing procedures.

a. The limitations on interpretations, conclusions, and recommendations in official NAEP reports (as outlined in Part I, Policy Principle 3) shall apply fully to any materials disseminated as part of the NAEP program by NCES and the Governing Board.

b. Researchers receiving secondary analysis grants from NCES may analyze data and provide commentary. Their reports may be disseminated by NCES if they meet NCES standards.
Appendix

NAEP Initial Release Reporting Covered by this Policy

The Nation’s Report Card™

The primary means for the initial public release of NAEP results shall be a summary report in each subject, known as The Nation’s Report Card™ and intended for the interested general public. The reports shall be made available in both print and electronic (Web-based) form. These reports shall present key findings and composite and disaggregated results. The printed reports shall be relatively brief, and written in a clear, jargon-free style with charts, tables, and graphics that are understandable and attractive. Data tables may be included in an appendix, either bound into the report or printed separately. This format shall be used to report key results for the nation and the states and of NAEP Trial Urban District Assessments.

A separate, dedicated Web site for the initial release of NAEP results shall be focused on a broad public audience, including less sophisticated users of the technology. The URL – http://nationsreportcard.gov – should be readily located via Internet search engines. Key NAEP findings will be available, clearly organized and prioritized. World Wide Web pages shall provide key findings, including composite and disaggregated results, as well as access to more extensive data sets.

Individual State and School District Reports

Relatively brief reports of key results shall be prepared for individual states, as well as for TUDA-participating school districts. All reports shall contain composite and disaggregated data, and may include an appendix with data tables.

Special Studies and Reports

Special studies and reports authorized by the National Assessment Governing Board and based on NAEP data collections will focus on specific topics of public interest and educational significance. They are aimed at policymakers and interested members of the public. They may include newly released data as well as data previously released that are analyzed to address issues identified by the Governing Board.
PARENT LEADER ENGAGEMENT OUTREACH STRATEGY
DEVELOPED BY REINGOLD
JULY 2013

INTRODUCTION

Reingold, the Governing Board’s communications contractor, has worked with Board staff and members of the Reporting and Dissemination Committee to develop and refine a parent leader engagement outreach plan for implementation in collaboration with National Center for Education Statistics (NCES) activities. Below is the outreach plan offered for approval by the Committee and then the full Board at the August 2013 meeting. The plan entails recommended strategies to reach parent leaders, including a suggested timeline as well as examples of potential outcomes and suggested metrics to measure the effectiveness of each strategy. The plan also includes overall goals of the parent outreach endeavor and important targets for the Board’s efforts in this arena.

GOALS

The Governing Board’s parent engagement plan seeks to promote the important role the National Assessment of Educational Progress (NAEP) plays in assessing and improving education in America, and instill a concern among parent leaders for increasing the achievement of all children. Parent leader outreach efforts should clearly convey how the Board believes parent leaders can use NAEP, and inspire parent leaders to:

1. Learn about NAEP and the data and resources available.
2. Understand NAEP’s applicability to their organization and mission.
3. Access and use NAEP tools to inform their work.
4. Inform and empower parents in their networks to learn about, understand, and use NAEP data and resources.
5. Have discussions and ask questions about improving student achievement and narrowing achievement gaps.

AUDIENCE AND PRIORITY OUTREACH TARGETS

Parent leaders are defined as organizations and individuals whose work and interests involve education and parents, and who see the connection between system performance and the potential for impact on individual students.
The parent leader audience has been segmented into these five subgroups: general education parent leaders; K-12 education parent leaders; minority and underrepresented population parent leaders; community parent leaders; and parent-focused media and online influencers.

Initial outreach efforts will focus on 50 priority parent leader groups across the parent leader subgroups. It will be important to create a targeted strategy for engaging these 50 groups with customized approaches, recognizing that they have varying levels of knowledge of NAEP.

Steps for selecting the parent leader target audience include:

- Reviewing the Board’s current stakeholder database to make sure that relevant individuals and groups within the subgroups are included.
- Determining the 50 parent leader groups that will be the focus of initial efforts.
- Analyzing the 50 groups and leaders to identify how familiar they are with NAEP, what communications assets they have, and what channels and activities they use to communicate to their networks.
- Developing a relationship map that identifies connections of Board members, Board alumni, and other NAEP champions to the 50 groups.

OUTREACH STRATEGIES

Below are the recommended strategies to engage the parent leader audience. This integrated approach uses traditional channels, such as in-person events and media relations, as well as outreach through new channels, including online media and social media.

I. Develop a Parent Leader Toolkit and Supporting Materials

Relevant, user-friendly materials will be fundamental to the success of the outreach plan, especially materials that have greatest use and applicability across all parent leader audiences and allow parent leaders to speak knowledgeably about NAEP. These three items will be the primary components of the parent leader toolkit:

- **NAEP 101 video.** This will be an introductory video to NAEP. It has become clear through Board outreach events and meetings with education groups that most leaders in education and the community do not know enough about NAEP to allow them to connect their efforts with its data and resources.
  
  ➢ **Examples of outcomes:** Parent leader groups embed the video on their websites for their audiences to see and use, or link to it on social media channels; parent leader groups show the video at their major education conferences
  
  ➢ **Possible metrics:** Number of video views; number of video engagements (shares, comments); increase in traffic from YouTube to the website

- **Parent presentation.** A PowerPoint presentation has been used occasionally at Governing Board events and conferences. This important tool must be updated to include the Board’s core messages for parent leaders and illustrate how NAEP materials can help parent leaders
engage their networks and advance their goals.

- **Examples of outcomes:** Parent leader groups use the presentation at conferences or events; parent leader group asks for a Governing Board member to give the presentation to its membership

- **Possible metrics:** Number of presentation downloads; number of email (or other outreach) requests for the presentation; number of live presentations given

- **Parent leader discussion guide.** Complementing the NAEP 101 video and the presentation, the discussion guide will assist parent leaders in their conversations about improving student achievement for all children. Discussion points will support their efforts with policymakers and administrators to understand how their school system or state compares with others nationwide, and to discuss what is being done to increase academic rigor and achievement for all students.

  - **Examples of outcomes:** Parent leader groups host workshops with parents walking through how to use the discussion guide; parent leaders distribute the discussion guide to their local or state school administrations

  - **Possible metrics:** Number of discussion guide downloads; number of email (or other outreach) requests for the discussion guide; number of printed guides or distribution outlets

- **Specialty Materials.** As the outreach effort grows, more materials will be developed to better demonstrate NAEP’s relevance and usefulness for each parent leader audience. Materials will be customizable and/or downloadable as needed and include:

  - **State and district profiles.** These will be parent leader-friendly versions of the NAEP state and Trial Urban District Assessment district profiles, with a focus on achievement-level data and key background variable findings. They also will include brief explanations of what the data show, including trend lines.

  - **Data infographics.** NAEP data will be shaped into infographics that are visually appealing and engaging to parent leaders.

  - **Parent leader testimonials.** Stories from parent leaders who have used NAEP data as resources to address education issues will be made into a video or a PDF document for print distribution.

  - **Background variables one-pager.** This will include information on the wealth of background variables collected with each NAEP assessment, and how parent leaders can access and use these data in their work.

  - **NAEP and the Common Core FAQ.** Most parent leaders may be more familiar with the Common Core State Standards initiative than with NAEP and have questions about the role of each. The NAEP 101 video may address this, but it will help to also address the differences in a frequently asked questions (FAQ) format available for parent leaders.
Examples of outcomes: Parent leader groups use materials at events or conferences; parent leaders distribute the materials to their local or state school administrations; parent leader groups share the materials on their websites and/or on social media channels

Possible metrics: Number of downloads of the materials; number of links back to the parent Web pages from the materials; number of printed materials or distribution outlets

II. Expand Integrated Web Presence and Online Engagement With Parent Leaders

Effective websites are a combination of strong content, strategic design, and online outreach. The outreach strategies will make the Governing Board’s website a primary destination for parent leaders, who may also visit it through search engines, word of mouth, or other channels, and so it is critical that the Web pages are user-friendly and provide relevant materials. The easier it is for content to be consumed and shared, the more online reach and visibility the parent engagement effort will have. Steps to optimize the parent leader online presence include:

- **Prioritizing content on parent Web pages.** Revisit the design and structure of the parent pages to make them easy to use, conveying key information and reinforcing messages tailored for this audience.
  - Examples of outcomes: Increased traffic (and returning visitors) to parent Web pages; increased downloads of materials; Visitors sharing the Web page or specific pieces of content from the Web page to their network or posting on their social media sites
  - Possible metrics: Number of visitors to parent Web pages; time spent on Web pages; number of conversions on established Web pages goals or desired actions such as downloading materials, signing up for an event, or watching a video

- **Performing search engine optimization to capitalize on search terms parent leaders use.** Determine priority keywords the Governing Board can use to make its parent pages appear higher and more often in search engine results, and create or refine website content to help raise the website’s ranking in search engine results.
  - Examples of outcomes: Increased traffic to parent Web pages; new visitors come to the website via search and then sign up for the latest NAEP release event; increased awareness of NAEP among new audiences unfamiliar with NAEP but searching for education information
  - Possible metrics: Number of parent Web page visitors; numbers of referrals to Web pages from search engines; shifts in Web pages’ rankings on search engines over time

- **Sharing NAEP digital content with targeted parent leader groups.** Provide timely and relevant NAEP content to the 50 priority parent leader groups in a variety of formats, such as
social media posts, a website paragraph, a newsletter blurb, infographics, or graphs from the state or district profiles.

- **Examples of outcomes:** Speakers start tweeting about NAEP/data during a high-profile national education conference or summit; celebrity spokesperson for education sees tweets and starts retweeting to his or her audience

- **Possible metrics:** Digital shares or engagements, including views for all video content; increase in traffic to the event Web page; increase in registrations

- **Initiating topics on discussion forums where parent leaders share best practices.** Provide content to parent-focused sites, and work with the site managers to promote topics, questions, or conversations on some of the many other popular parent sites.

- **Examples of outcomes:** Portal hosts a banner advertisement or application that links through to the Governing Board or NAEP website; parent leader uses a conversation thread as fodder for his or her next presentation to his group

- **Possible metrics:** Number of post views, replies, and quality of engagement of the thread; increased traffic back to the website; shared NAEP links and resources on the thread for users to click through

- **Expanding and promoting the NAEP Results app.** The NAEP Results app was published on the iTunes store in 2012, allowing mobile users to dive into NAEP data via mobile device. In addition to promoting the app, the Board can consider working with NCES to integrate content and functionality that is specifically tailored to parent leaders.

- **Examples of outcomes:** Parent leader groups host a series of Web-based meetings each month to walk through different NAEP release results via the app; parent leaders use the app to walk school leadership through state-level data

- **Possible metrics:** Number of app downloads; deeper analytic data such as total users, time spent on app, engagements, downloads, and other user actions

### III. Expand Thought Leadership Through Partnerships, Events, and the Media

The Governing Board can raise awareness of NAEP and the Board’s role in education through consistent efforts to engage key influencers. The Board has successfully established relationships with nationally recognized parent-focused organizations, including the National PTA, and can continue to use new and existing partnerships and publications to influence new audiences in a strategic way.

- **Speak at education-related conferences.** Representatives of the Board can present at gatherings such as conferences of parent, education, policy, business, and civil rights organizations.
Examples of outcomes: Conference participant asks to have Governing Board member speak at another upcoming conference; host organizations upload NAEP materials to their websites for others to download following the conference

Possible metrics: Number of conference participants; number of requests for additional materials; number of requests for additional speaking engagements; number of new relationships created with participants and organizations

- Co-sponsor panels, forums, or workshops. The Board can work with groups like Achieve or Council of the Great City Schools to host conversations about NAEP data releases and other NAEP efforts of interest to parent leaders.

Examples of outcomes: Parent leader group includes a panel on NAEP tools at its next conference to educate its network of parents; parent leader group uses the Governing Board panel as a springboard for developing a series of monthly sessions for parents on using NAEP data

Possible metrics: Number of total attendees; number of new attendees not in parent leader database; number of requests for follow-up

- Develop editorial pieces for parent leaders, such as articles to appear in a newsletter or blog for parent leaders. Engage parent leaders with regularly updated, timely communications that tie together the day’s headlines about education with NAEP findings, with links back to the parent pages of the website.

Examples of outcomes: Parent leader group places the article in its monthly newsletter and causes a spike in the number of parent leaders registering for a report card release; a reader shares the newsletter item with several new parent leader contacts that follow up with the Board’s email address; the content is shared over social media

Possible metrics: Numbers of articles or blogs placed; number of impressions; number of click-throughs to the website

- Pitch parent-focused articles or newsletters to education journalists or publications. Use the ongoing relationships the Board has developed with the media to distribute targeted, parent-focused messages and encourage them to publish, post, and share content tailored for the parent leader audience.

Examples of outcomes: Parent leader group shares an article with its organization, and then calls a meeting to discuss it at an upcoming education panel; webinar parent leader participant cites a media article as his or her source for the introduction to NAEP and the Board’s work

Possible metrics: Number of articles placed; number of impressions; number of links back to the website
## EXECUTION TIMELINE

### Governing Board Parent Leader Engagement Timeline

<table>
<thead>
<tr>
<th>Task</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Audience and Priority Outreach Targets</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review and finalize stakeholder database</td>
<td>8/5/13</td>
<td>8/23/13</td>
</tr>
<tr>
<td>Determine 50 priority groups</td>
<td>8/26/13</td>
<td>9/6/13</td>
</tr>
<tr>
<td>Map top 50 groups to awareness/resources</td>
<td>9/9/13</td>
<td>9/20/13</td>
</tr>
<tr>
<td>Develop relationship map</td>
<td>9/9/13</td>
<td>9/20/13</td>
</tr>
<tr>
<td><strong>Materials Development</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parent Leader Toolkit</td>
<td>8/26/13</td>
<td>10/5/13</td>
</tr>
<tr>
<td>Subsequent tailored materials</td>
<td>10/8/13</td>
<td>7/31/14</td>
</tr>
<tr>
<td><strong>Website and Online Engagement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Update nagb.org parent pages</td>
<td>8/12/13</td>
<td>10/5/13</td>
</tr>
<tr>
<td>Perform search engine/keyword optimization</td>
<td>9/9/13</td>
<td>10/5/13</td>
</tr>
<tr>
<td>Digital content outreach and engagement</td>
<td>10/8/13</td>
<td>7/31/14</td>
</tr>
<tr>
<td><strong>Thought Leadership: Partnerships, Events, and Media</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking engagements</td>
<td>10/7/13</td>
<td>7/31/14</td>
</tr>
<tr>
<td>Co-sponsor workshops, panels, forums, etc.</td>
<td>10/7/13</td>
<td>7/31/14</td>
</tr>
<tr>
<td>Develop editorial pieces - newsletters, blogs</td>
<td>8/12/13</td>
<td>7/31/14</td>
</tr>
<tr>
<td>Develop and pitch parent-focused stories</td>
<td>8/12/13</td>
<td>7/31/14</td>
</tr>
</tbody>
</table>

* Dates are for illustrative purposes only. Final deliverable due dates will be determined upon confirmation of scope of outreach.
POSSIBLE NEW FORMATS FOR NAEP REPORTING

The 2013 National Assessment of Educational Progress (NAEP) Mathematics and Reading Report Cards will be released in October 2013, and reflect the beginning of changes in how NAEP results will be presented to the public in terms of the actual report and online data, tools, resources, and other materials. Reporting of the NAEP 2013 results will be primarily through the NAEP web site, accompanied by a printed report summary that will be shorter than previous Report Cards. The National Center for Education Statistics (NCES) is planning to create a series of charts and graphs, accompanied by brief text, that convey the important results and interesting findings. The goal is to find ways to streamline NAEP data and make findings more consumable for the general public.

At the May 2013 Board meeting, NCES Commissioner Jack Buckley presented a short video to detail some of the strategies NCES and its contractors were formulating to update NAEP reports and the web site in an effort to make findings more easily understandable. NCES also shared with Board staff an initial draft template of what future NAEP Report Cards could look like. Both the video and the report template will be shared with Reporting and Dissemination Committee members at the August meeting.

The Committee will hear an update from NCES on new and updated formats considered for NAEP reporting and how content could be presented differently. The Committee will also have the opportunity to ask questions and offer feedback in this crucial area.
The 2013 NAEP Mathematics and Reading Report Cards will be released together to the general public during October 2013 in one event, as approved by the Board at the May 2013 meeting. Following a review and approval of the report’s results, the release will be arranged as an online webinar. The release event will include a data presentation by the Commissioner of Education Statistics, with moderation and comments by at least one member of the National Assessment Governing Board and at least one additional panelist with a background in mathematics and/or reading education or assessment. Full accompanying data will be posted on the Internet at the scheduled time of release.

The 2013 NAEP Report Cards in mathematics and reading will present findings from a representative sample of about 320,000 4th-graders and 315,000 8th-graders nationwide. These samples included about 6,000 private school students at each grade, and the rest were public school students. About half the students took the math assessment, and half took the reading assessment. Results released will be for the nation and the states (including the District of Columbia and the Department of Defense Education Activity schools). Data will be presented for all students and by race/ethnicity, gender, type of school (public, all private, and Catholic) and eligibility for the National School Lunch Program. Contextual information (i.e., student, teacher, and school survey data) with findings of interest will also be reported.

DATE AND LOCATION

The release event for the media and the public will occur in October 2013. The release date will be determined by the Chair of the Reporting and Dissemination Committee, in accordance with Governing Board policy, following acceptance of the final report.
EVENT FORMAT

- Introductions and opening statement by a National Assessment Governing Board representative
- Data presentation by the Commissioner of Education Statistics
- Comments by at least one Governing Board member
- Comments by at least one expert in the field of reading and mathematics assessment or education
- Questions from the webinar audience
- Program will last approximately 75-80 minutes
- Event will be broadcast live over the Internet, and viewers will be able to submit questions electronically for panelists. An archived version of the webinar, with closed captioning, will be posted on the Governing Board website at www.nagb.org.

EMBARGOED ACTIVITIES BEFORE RELEASE

In the days preceding the release, the Governing Board and NCES will offer access to embargoed data via a special website to approved U.S. Congressional staff in Washington, DC; approved senior representatives of the National Governors Association and the Council of Chief State School Officers; and appropriate media as defined by the Governing Board’s Embargo Policy. A conference call for journalists who signed embargo agreements will be held to give a brief overview of findings and data and to answer questions from the media.

REPORT RELEASE

The Commissioner of Education Statistics will publicly release the report at the NAEP website—http://nationsreportcard.gov—at the scheduled time of the release event. An online copy of the report, along with data tools, questions, and other resources, will also be available at the time of release on the NAEP site. An interactive version of the release with panelists’ statements, a Governing Board press release, subject frameworks, and related materials will be posted on the Board’s web site at www.nagb.org. The site will also feature links to social networking sites and audio and/or video material related to the event.

ACTIVITIES AFTER THE RELEASE

The Governing Board’s communications contractor, Reingold, will work with Board staff to coordinate two separate post-event webinars or other communications efforts—one targeted for the larger mathematics community, and one targeted for the larger reading community. The goal of these activities is to extend the life of the results and provide value and relevance to stakeholders with an interest in student achievement and assessment in these subject areas.
Update on NAEP Focused Reports

As has been previously reported to the National Assessment Governing Board, NCES is in the process of revitalizing a series of reports that focus on different aspects of NAEP. “Focus on NAEP,” as this series is titled, will consist of short reports geared towards our stakeholders and designed to inform them on topical areas on NAEP that are not a part of the regular NAEP discussion. Initial topics include a summary of results across the 2010 assessments (civics, geography, and U.S. history), an in-depth discussion of the performance of English language learners, how NAEP sampling is conducted, and a look at the gender gap across several academic subjects.

These reports will primarily be web-based, but designed so that a paper version could be handed out to our audiences at various events, such as report releases, conferences, and meetings. Producing the Focus on NAEP series online will allow us to utilize the various online tools available to us. This includes the ability to link to the NAEP Data Explorer and the NAEP Questions Tool with just a click of the mouse; the incorporation of graphs, tables, and other graphics that allow a deeper look by embedding links and paths to other documents on the NAEP website and elsewhere; and the ability to house this series of reports in one area online, making it easy for the user to see other Focus on NAEP topics they might be interested in learning about. The transition to online presentation will allow these reports to represent the data in more creative ways than in printed form.

Currently, the content for several reports has been developed. The next step, currently under way, is to design the web versions.
NOTE TO Reporting and Dissemination Committee on Focused NAEP Reports

The Policy Statement on NAEP Background Questions and the Use of Contextual Data in NAEP Reporting, adopted by the Governing Board in August 2012, contains two implementation guidelines that deal with special or focused NAEP reports. These would be extracted from the vast NAEP data sets and report in some detail on particular topics of interest.

- **Guideline 9:** Special focused reports with data through the 2013 assessment will be issued on the following topics: private schools, charter schools, gender gaps, and black male students. Reports shall include significant contextual information as well as cognitive results. Advisory committees, composed of a range of knowledgeable persons, may be appointed to provide input on reporting issues.

- **Guideline 10:** Exploratory analyses will be carried out to determine if existing background questions may form the basis for additional focused reports. Such reports may be issued by the Governing Board as well as by the National Center for Education Statistics.

In addition to the gender gaps report mentioned in the NCES update, the report on black male students (in grade 8 only) is expected to be ready for release in January 2014, according to the schedule in Attachment F for this committee meeting. The NCES special reports on private schools and charter schools will be developed later.

As part of the exploratory analyses authorized by the Board, four reports have been prepared:

- **Who Attends Charter Schools and How are Those Students Doing** by Naomi Chudowsky
- **Time for Learning (national report)** by Alan Ginsburg
- **Time for Learning: States and Districts** by Alan Ginsburg
- **Monitoring What Matters about Context and Instruction in Science Education** by Alan Friedman and Alan Ginsburg

These reports have been shared with the Ad Hoc Committee on NAEP Background Information and posted on the Governing Board web site. They have received some press coverage in Education Week. Over the next few years the Board may wish to have consultants prepare additional reports using available data. Also, as provided for in the 2012 resolution, the Board may choose topics for one-time special studies with specifically-developed modules of background questions that would be the basis for reports giving descriptive information on issues of current policy interest, such as technology use or out-of-school learning activities.
EVENT DEBRIEF
July 11, 2013

Overview
The National Assessment Governing Board’s webinar to release The Nation’s Report Card: Trends in Academic Progress 2012 took place at 11:30 a.m. EDT on June 27, 2013. Reingold executed activities related to the event in cooperation with staff and contractors of the Governing Board and National Center for Education Statistics.

Panelists included:
- Jack Buckley, Commissioner, National Center for Education Statistics
- Brent Houston, Principal, Shawnee Middle School, Shawnee, Okla.; Member, National Assessment Governing Board
- Kati Haycock, President, The Education Trust
- Cornelia Orr, Executive Director, National Assessment Governing Board (moderator)

Webinar Attendance
This release event had 270 attendees from 206 organizations. With 425 people registered for the event, the attendance rate was 64 percent. (Internal staff and contractors are not included in these totals.)

- This event was particularly appealing to education groups, which made up 36 percent of the audience—the largest percentage of the total attendance.
- There was high representation from state education departments, which made up 26 percent of participants.
- About 16 percent of attendees were from an institution of higher education.
- The miscellaneous group, totaling 15 percent of participants, included individuals and representatives from government agencies, business organizations, and consultancies.
Traditional Media Coverage

Media coverage highlights include:

- 44 reporters gained access to embargoed report card data.
- Within 24 hours of the release event, the following news organizations published 22 original stories about the long-term trend report:
  - Associated Press
  - The Atlanta Journal-Constitution
  - Bloomberg
  - Choice Media
  - CNN, Schools of Thought
  - The Daily Caller
  - Education Week (two stories)
  - Gawker
  - KRMG
  - The Hechinger Report
  - The Huffington Post (two stories)
  - Latino Ed Beat
  - Milwaukee-Wisconsin Journal Sentinel
  - Pittsburgh Post-Gazette
  - Reuters
  - The Kansas City Star
  - The Wall Street Journal
  - The Washington Post
  - USA Today
  - U.S. News & World Report
- Within one week of the release event, nine additional original stories were published.

Social Media Coverage

On June 27, the day of the release event, there were 693 on-topic social media mentions. Social media mentions of the release event or data posted within 24 hours after the event included representation from numerous organizations including 50CAN, the Center on Reinventing Public Education at the University of Washington, The Education Trust, the Foundation for Excellence in Education, Knowledge Alliance, the Mid-continent Regional Educational Laboratory, and the National Association of State Boards of Education.
The Nation’s Report Card: 
Trends in Academic Progress 2012

Selected articles from news media 24-hour coverage

**High school seniors fare no better than in 1970s**
The Associated Press—Philip Elliott

**5 ways students changed in the last 40 years**
CNN, Schools of Thought—Jamie Gumbrecht

**NAEP report: A closer look at trends in the achievement gap**
Education Week—Erik Robelen

**Report: High school students have made no progress in 40 years**
US News and World Report—Allie Bidwell

**Minorities’ learning gap shrinks, report says**
USA Today—Greg Toppo

**Decades-long study shows gains in U.S. education**
The Kansas City Star—Joe Robertson

**Education spending: High schoolers not any smarter than in the 1970s**
KRMG—Rick Couri
WASHINGTON (AP) — Students preparing to leave high school are faring no better in reading or math than students did four decades ago, the government said in a report Thursday that was certain to renew concerns about U.S. schools.

Test scores for 17-year-olds have changed little since the early 1970s, while students ages 9 and 13 improved their performances during the same period, according to the government review popularly called the nation's report card.

Black and Hispanic students achieved the greatest gain in reading and math scores since the 1970s and the performance gap between white and minority students narrowed. "In some ways, the findings are full of hope. Today's children ages 9 and 13 are scoring better overall than students at those ages in the early '70s," said Brent Houston, principal of the Shawnee Middle School in Oklahoma and a member of the National Assessment Governing Board, which administers the tests.

But he also noted challenges for older students.

"There is a disturbing lack of improvement among 17-year-olds. Since the early 1970s, the average scores of 17-year-olds in both reading and mathematics have remained stagnant," he said.

The report says that in reading, today's 9- and 13-year-olds are outperforming students tested in 1971, when that skill was first tracked. They also did better in math, compared with students in 1973, the initial measurement.

Officials suggest the results for 17-year-old students reflect fewer low-performing students dropping out.

For instance, Hispanic students had a 32 percent dropout rate in 1990 and that number fell to 15 percent in 2010, said Peggy Carr, an associate commissioner with the National Center for Education Statistics.

"These students are generally scoring at the lower end of the distribution but it's a good thing that they're staying in schools," Carr said.

Black and Hispanic students at all ages narrowed the performance gap with white students, according to the National Assessment of Educational Progress. Among 17-year-old students, the gaps between black and white students and between Hispanic and white students were cut by half.

In math, 9-year-old black and Hispanic students today are performing at a level where black and Hispanic 13-year-olds were in the early 1970s.

"Black and Hispanic children have racked up some of the biggest gains of all," said Kati Haycock, president of the Education Trust, an advocacy organization. "These results very clearly put to rest any notion our schools are getting worse. In fact, our schools are getting better for every group of students that they serve."
The overall composition of classrooms is changing as well.

Among 13-year-old students, 80 percent were white in 1978. By 2012, that number fell to 56 percent. The number of Hispanics roughly tripled from 6 percent in 1978 to 21 percent in 2012.

"Over a 40-year period, an awful lot changes in our education system," said Jack Buckley, the chief of the National Center for Education Statistics.

While most groups of students saw their scores climb since 1971, the same cannot be said when comparing 2008 results with 2012. The 9-year-old and 17-year-old students saw no changes and only Hispanic and female 13-year-olds showed improvement in reading and math.

The 2012 results were based on 26,000 students in public and private schools. The tests took roughly one hour and were not significantly different than when they were first administered in the early 1970s.

Unlike high-stakes tests that are included in some teachers' evaluations, these tests are a more accurate measurement because "these are not exams that teachers are not teaching to," Haycock said.

"Nobody teaches to the NAEP exam, which is why it's such as useful measure to what our kids can actually do," she said.
The National Assessment of Educational Progress releases a short-term snapshot of how students fare in science, civics or other subjects.

But it doesn't quite answer the big question: How are students really doing?

That's the job of a report released Thursday, "The Nation's Report Card: Trends in Academic Progress 2012." It's an assessment released every four years that tracks U.S. students' performance in reading and math since the 1970s. The 2012 assessment included more than 50,000 students from public and private schools. It tracks them at ages 9, 13 and 17, regardless of grade level, and compares their performance using tests that take about an hour and features mostly multiple-choice questions.

Here are five things to know about academic progress since the 1970s, according to the 2012 report.

9-year-olds and 13-year-olds outscore 1970s counterparts: Indeed, those kids scored higher in reading and math. In reading, 9-year-olds and 13-year-olds improved at every level, so even the lowest-performing kids now are ahead of the lowest-performing kids then. In fact, kids in the low and middle range showed the greatest gains.

17-year-olds? Not so much: Seventeen-year-old students aren't scoring better in reading and math, but their scores aren't falling, either. In reading, the lowest-performing 17-year-olds made gains since the 1970s, as did lower- and middle-performing 17-year-olds in math. But scores overall are about the same as in the early 1970s – and that might not be all bad. In a conference call with reporters, Peggy Carr, associate commissioner of the National Center for Education Statistics' assessment division, pointed out that there are far fewer dropouts than in the 1970s, but even with more kids in school, performance has remained steady.

Gender gaps are shrinking: Just as in the 1970s, girls perform better in reading, and boys perform better in math.

But female students are narrowing the math gap, or even eliminating it. "In 2012, there were no significant gender gaps in mathematics at age 9 and 13," the report says. "At age 17, male students scored higher in mathematics than female students. The gender gap in 2012 at age 17, however, was narrower than in 1973 due to the increase in the average score for female students."

Meanwhile, male students are squeezing the gap in reading by showing significant improvement at age 9.

Black and Hispanic students are making gains: Consider just how much students' demographics have changed: In 1978, 80% of U.S. students were white, 13% were black, 6% were Hispanic and 1% were Asian. In 2012, 56% of students were white, 15% were black, 21% were Hispanic and 6% were Asian.

White students still perform better than black and Hispanic students in reading, but the gaps between white and black and white and Hispanic are narrower for all ages. It's particularly noticeable among 9-year-olds: "The average score for black students was 36
points higher in 2012 than in 1971 ... and the score for white students was 15 points higher," the report says. "The average score for Hispanic students increased 25 points from 1975, and the score for white students increased 12 points."

In math, white students performed better overall, but black and Hispanic students made larger gains than white students since 1973.

**Take another look at that summer reading list:** At age 9, 53% of students say they read for fun at home almost every day. By age 13, it's 27%. At 17, it's down to 19%. The percentages for 9-year-olds have remained the same since 1984, when the question was first asked, but it has decreased over time for 13- and 17-year-olds. Why does it matter?

"At all three ages, students who reported reading for fun almost daily or once or twice a week scored higher than did students who reported reading for fun a few times a year or less," the report says.
A new report from "the nation's report card" (and my own Education Week story yesterday) emphasizes progress in closing achievement gaps for black and Hispanic students between the early and mid-1970s and today.

While this is mostly true (except for one category, the Hispanic-white gap for 9-year-olds in math), it doesn't tell the whole story. (Thanks to Bob Rothman from the Alliance for Excellent Education, who posted a comment on my story that got me thinking.)

Much of that narrowing of the achievement gap was actually accomplished by the mid to late 1980s, the data indicate. It's ebbed and flowed a bit since then, but in most cases, the gaps are no smaller today than they were two decades ago. In fact, they're sometimes larger, though not by amounts deemed statistically significant.

For example, the black-white achievement gap for 13-year-olds in reading reached its narrowest point in 1988, at 18 points, compared with 23 points in 2012.

In math, the black-white achievement gap for 9-year-olds was 25 points in 1986, the exact same figure as for 2012.

The only instance I could find where the numerical achievement gap was smallest today was the black-white gap for 9-year-olds in reading. It reached 23 points in 2012. In 1988, the figure was 29 points (though this difference is not considered statistically significant).

Of course, achievement gaps are not the only thing to be concerned with. The hope is that all students will make progress over time. The good news here is that whites, blacks, and Hispanics all have seen increases in their average scores since results were first available on the long-term trends report. But the results get more complex when comparing the results from the mid- to late-1980s to today. In reading, average scores for 9-year-old blacks are improved today over 1988, but for 13-year-olds and 17-year-olds the difference was not statistically significant.

The new data come from the National Assessment of Educational Progress long-term trends report. Keep in mind that this assessment is different from the main NAEP in reading and math administered every two years. You can learn more about the differences here. I should also note that the NAEP study does not report out data for Asian/Pacific Islander students or for American Indian (including Alaska Native) students because of limits on the data available. They were included in the national samples, however, and some data for them can be found on the NAEP Data Explorer website.

Given how much data we're talking about with the NAEP long-term trends report, it's hard to identify perfect trend lines. Any analysis is complicated by the multiple factors involved: We're talking about:

- Two subjects, reading and math;
- Two gaps, black-white and black-Hispanic;
- Three age levels (9, 13, and 17); and
- Up to 13 different assessment years.
Of course, any talk of progress in closing gaps can quickly introduce an element of politics, as people may wish to use the data to argue that a particular evolution in education policy explains the changes, such as the push for tougher accountability measures. So don't be surprised in coming days if these data are used to defend a number of different agendas. I'll stay out of that debate here, but will do my best to shed a little more light on what the data show.

Now, let's do the numbers!

Below I've reproduced several graphics from the NAEP report that provide a detailed look at changes in the gaps over time. Take a look and draw your own conclusions about what it all means. But be sure to keep an eye out for those small asterisks. They signal years in which the results are considered different by a statistically significant margin.

In the chart below, you'll notice the gap was smallest in 1988.

The Hispanic-white gap in 1988 was exactly the same as in 2012, this next chart shows.

In math, this next chart shows the black-white achievement gap was the same amount, 25 points, in 1986, 1994, 1996, and 2012.

As this final chart shows, the Hispanic-white gap in 2012 was 17 points. This is NOT considered a statistically significant change from 1975, though it is measurably larger than for several points in time, including the 1999 results.

I'll close by briefly tackling one other complication raised by the NAEP data. I was a little puzzled about how it could be that average national scores for 17-year-olds were about the same in 2012 as they were back in the early 1970s, even as whites, blacks, and Hispanics all saw progress. For help in making sense of this, take a trip over to the Change the Equation blog.

"The reason for this apparent impossibility?" the blog post says: "Black and Hispanic students, who unfortunately lag behind their white peers, make up a much bigger share of the population now than they did in 1973. That brings down the total score."

Anyway, as always there's lots to mine in the new NAEP report. But it's complicated stuff, and there are plenty of caveats, so take your time!
While today's elementary and middle school students are scoring higher in reading and mathematics than 40 years ago, and scores show that race and gender achievement gaps may be narrowing, there is a "disturbing" lack of improvement among the nation's high school students, according to a report from the National Assessment of Educational Progress.

The NAEP's long-term trend assessment measures the basic reading and mathematics skills of 9-, 13- and 17-year-olds in American public and private schools every four years to show how students' performance has changed over time. In 2012, black and Hispanic students of all ages scored significantly higher in both mathematics and reading than students did in the early 1970s, when the assessments were first given. Gender gaps have also narrowed since the 1970s, with female students of all ages scoring better in math, and male students, who typically score lower in reading, narrowing the gap at age 9.

But since 2008, only one achievement gap – the White-Hispanic reading gap for 13-year-olds – has narrowed, according to the report.

"If we have a crisis in American education, it is this: That we aren't yet moving fast enough to educate the 'minorities' who will soon comprise a 'new majority' of our children nearly as well as we educate the old majority," said Kati Haycock, president of The Education Trust, an organization that promotes closing achievement gaps.

"At best, students of color are just now performing at the level of white students a generation ago," she added in a released statement.

At first glance, it appears that all groups have shown some gains since the early 1970s. But the one group that has remained stagnant is 17-year-old students. Taken as a whole, that group has not made an improvement in either subject over the last 40 years.

Brent Houston, who serves on the National Assessment Governing Board that conducted the study, said in a released statement that the data collected for the report also include parents' level of education, which shows an increasing number of whom have graduated from college. This emphasis on education, he said, should translate into better performance for their children. But still, the average scores of 17-year-olds have stayed flat.

"If parents are achieving more, you'd think that older students in particular would be achieving at higher levels," he said in the statement.

But the report also revealed that children who more frequently read for fun are scoring higher in reading than those students who do so less frequently. In 2012, 53 percent of the 9-year-olds tested said they read for fun almost every day, and 23 percent said they do so once or twice a week. Those two groups scored more than 10 points higher than 9-year-olds who said they read for fun only a few times a year.

Overall, there are "considerable bright spots" in the report, said NAEP Governing Board Chair David Driscoll, in a released statement.

"Assessing students at particular ages over the decades provides a unique perspective on learning and achievement and a way to take a step back to see overall achievement trends and just how far we've come," he added.
Minorities’ learning gap shrinks, report says
USA Today—Greg Toppo
Published June 27, 2013, 12:40 p.m.

Over the past two generations, African-American 9-year-olds have nearly cut in half the reading skills gap between them and their white peers, new findings from the federal government show. Hispanic kids have cut the reading gap by more than one-third.

Data out Thursday from the U.S. Department of Education show that young African-American and Hispanic students' reading skills, while still lagging those of white peers, grew faster than white students' skills from 1971 to 2012. The results are from the long-term National Assessment of Educational Progress (NAEP), often called "The Nation's Report Card."

Since 1971, when the department first started tracking reading, African-American 9-year-olds have narrowed the reading gap with whites from 44 points to 23 points, scoring 206 on a 500-point scale in 2012. White students, on average, scored 229. Hispanic 9-year-olds, meanwhile, have shrunk the gap from 34 points to 21 points since 1975, scoring 208. NAEP results didn't break out Hispanic scores until 1975.

The data suggest that black and Hispanic students still have a long way to go — they are now reading nearly as well as white 9-year-olds did in 1971.

Peggy Carr, who heads testing for the education Department's National Center for Education Statistics, says it's "impressive" that overall scores have risen, even as the agency began including English-language learners and disabled students in the testing pool. Also, since the 1970s, the percentage of Hispanic students has steadily grown, from 6% to 21%.

Among the few disappointments in the new data: 17-year-olds' performance. The lowest-performing have improved their reading since 1971, but overall scores are flat.

The Washington, D.C.-based Education Trust, an advocacy group for low-income and minority students, said the new findings "put to rest any notion that our schools are getting worse." But the group said that American education isn't moving fast enough to improve skills — especially for Hispanic students, who already make up the majority of students in Texas public schools, according to the Texas Education Agency. By the end of the decade, the U.S. Census Bureau estimates that more than half of Americans under 18 will belong to a racial or ethnic minority group.

Daria Hall, the Education Trust's K-12 policy director, notes that much improvement came in the past 15 years as states held schools more accountable for academic results. The recent slowing of improvement suggests that schools have "gotten a lift out of accountability, about providing the public and teachers information about where kids are, but we've got to pick up the pace."

She says schools must invest in well-prepared teachers and a rich curriculum — and they need to confront inequities in school discipline such as suspension and expulsion rates that disproportionately affect minority students. "We've got to do a full-court press on all of these things," she says.
Decades-long study shows gains in U.S. education

The Kansas City Star—Joe Robertson
Published June 27, 2013

Amid all the worry over the state of American education, a long look back can actually feel good.

The latest returns in an ongoing, four-decade study by the National Assessment of Educational Progress released today served as a reminder that — within the concerns of America’s international standing — we have gotten better.

The report, in what is known as The Nation’s Report Card, found that U.S. students are getting stronger in reading and math. The achievement gaps between white children and minority children have been narrowing.

The study involving more than 50,000 students across the nation noted there has been significant growth in performance among 9-year-olds and 13-year-olds.

Even with 17-year-olds, where the growth was mostly flat, there was good news, said Peggy Carr, associate commissioner of the National Center for Education Statistics.

Just 20 years ago, dropout rates among 17-year-olds were twice what they are now, especially among Hispanic students. Hispanic dropout rates have dropped from 32 percent to 15 percent, she said.

“A lot more students are staying in school and (the scores) are not going down,” she said. Under those circumstances, she said, “flat is good.”

Nine-year-olds overall have seen 13 points in growth in reading since 1971, and 13-year-olds improved by eight points — both considered significant.

In math, 9-year-olds have improved by 25 points, 13-year-olds by 19 points, also significant.

The performance gaps between white students and black students and between white students and Hispanic students have narrowed at all of the age groups.

Closing the gaps is growing in importance, not just as a social justice issue, but an economic issue. The study also showed the changing face of American students. In the 1970s, 80 percent of the students assessed were white, 13 percent were black and 6 percent were Hispanic. In 2012, 56 percent were white, 15 percent were black and 21 percent were Hispanic.

The question remains if the improvement is moving fast enough. The gains made since the last long-term study report, in 2008, showed only small growth that was not statistically significant.

The U.S. has been decidedly mediocre or slightly above average in many tests that make international comparisons.

American 15-year-old students scored average in the latest reading tests by the Program for International Student Assessment, ranking somewhere between seventh and 20th among 33 participating nations. But the U.S. came in below average in math, ranking between 17th and 28th.

So the work has to go on to improve U.S. schools, said Kati Haycock, president of the
Education Trust, but the long-term study should give the work encouragement.

“This should get us beyond the finger pointing over whether our children are in crisis or not,” she said. “Our schools are getting better — for everyone. If there is a crisis, it’s whether we’re moving fast enough to serve the new majority as well as we serve the present majority.”
Education spending: High schoolers not any smarter than in the 1970s

Published June 28, 2013, 6:25 a.m.

Study covers nearly 30,000 students

The biggest disappointment comes from the fact the students are not performing better despite increases in education spending.

The findings come from the National Assessment of Educational Progress more commonly called the Nation's Report Card.

Officials point at lower-performing students remaining in school as one of the biggest reasons for the numbers.

Even so, some are saying the numbers are not as bad as it seems.

Shawnee Oklahoma Middle School principal Brent Houston is a member of the National Assessment Governing Board.

"In some ways, the findings are full of hope. Today's children ages 9 and 13 are scoring better overall than students at those ages in the early '70s," he said.

The results are from 2012 and are based on 26,000 students in both public and private schools.
### Upcoming NAEP Reports as of August 2013

<table>
<thead>
<tr>
<th>Report</th>
<th>Expected Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial NAEP Releases</strong></td>
<td></td>
</tr>
<tr>
<td>2013 Mathematics Grades 4 and 8</td>
<td>October 2013</td>
</tr>
<tr>
<td>2013 Reading Grades 4 and 8</td>
<td>October 2013</td>
</tr>
<tr>
<td>2013 Mathematics TUDA Grades 4, 8</td>
<td>December 2013</td>
</tr>
<tr>
<td>2013 Reading TUDA Grades 4, 8</td>
<td>December 2013</td>
</tr>
<tr>
<td>2011 State Mapping</td>
<td>December 2013</td>
</tr>
<tr>
<td>2013 Reading and Mathematics, Grade 12</td>
<td>April 2014</td>
</tr>
<tr>
<td><strong>Other NAEP Reports</strong></td>
<td></td>
</tr>
<tr>
<td>Linking NAEP and TIMSS 2011 Mathematics and Science Results for the 8th Grade- (Highlights Report)</td>
<td>September 2013</td>
</tr>
<tr>
<td>Linking NAEP and TIMSS 2011 Mathematics and Science Results for the 8th Grade- (Technical Report)</td>
<td>December 2013</td>
</tr>
<tr>
<td>Performance of Grade 8 Black Male Students on the National Assessment of Educational Progress</td>
<td>January 2014</td>
</tr>
<tr>
<td><strong>Other Related Reports from NCES</strong></td>
<td></td>
</tr>
<tr>
<td>2008-09 Baccalaureate and Beyond Longitudinal Study *</td>
<td>August 2013</td>
</tr>
<tr>
<td>Degrees of Debt, Student Borrowing and Loan Repayment of Bachelor’s Degree of Recipients 1 Year After Graduating *</td>
<td>August 2013</td>
</tr>
<tr>
<td>Digest of Education Statistics, 2012</td>
<td>August 2013</td>
</tr>
<tr>
<td>Public School Teacher Classroom Autonomy and Mobility **</td>
<td>August 2013</td>
</tr>
<tr>
<td>STEM Attrition: College Students’ Paths Into and Out of STEM Fields ***</td>
<td>August 2013</td>
</tr>
</tbody>
</table>

*From the Baccalaureate and Beyond Survey
**From the 2008 Schools and Staffing Survey
***From the Beginning Postsecondary Survey
Releases in 2013

- Linking NAEP and TIMSS 2011 Mathematics and Science Results for the 8th Grade (Highlights)
- 2013 Reading Report Card: Grades 4 and 8
- 2013 Mathematics Report Card: Grades 4 and 8
- 2013 Reading Report Card: Trial Urban Districts (TUDA): Grades 4 and 8
- 2013 Mathematics Report Card: Trial Urban Districts (TUDA): Grades 4 and 8
- Linking NAEP and TIMSS 2011 Mathematics and Science Results for the 8th Grade (Technical Report)
- 2011 State Mapping Report
Releases in 2014

- Performance of Grade 8 Black Male Students on the National Assessment of Educational Progress
- 2013 Reading and Mathematics Report Card: Grade 12

Assessment Data Collection Schedule 2014

- U.S. History: Grade 8
- Civics: Grade 8
- Geography: Grade 8
- Technology and Engineering Literacy: Grade 8
Common Core State Standards Assessment Consortia

Introduction

At this meeting, representatives of the PARCC and SBAC Common Core State Standards (CCSS) Assessment Consortia will discuss several topics on which they are working that are of interest to the Governing Board. This discussion will be in a moderated question and answer format. After brief introductory remarks, the questions provided below will be responded to by Joe Willhoft, representing the SMARTER Balanced Assessment Consortium (SBAC), and Jeff Nellhaus, representing the Partnership for Assessment of Readiness for College and Careers (PARCC). Once Mr. Nellhaus and Mr. Willhoft have provided initial responses to the questions for each topic area, Board members may ask follow-up questions before going on to the next topic area.

DISCUSSION TOPICS/QUESTIONS

1. **Inclusion/Accommodation Policies**: What are the current decisions about the inclusion of students with disabilities in the consortia assessments and the accommodations that would be permitted for students, especially in reading? How have you determined which accommodations provide the most valid assessment of the CCSS constructs?

2. **Technology-Based Assessments**: How is the planning progressing for the use of computers and other devices to administer the CCSS assessments (lessons learned)? How ready will your member states/districts/schools be for technology-based testing in 2013-14? 2014-15? Since the assessments being developed are utilizing technology-enhanced items, how will you be addressing the construct differences that may occur when paper-based tests are used?

3. **Survey Questions**: What kind of consideration, if any, has been given to asking students and others "non-cognitive" questions? For example, questions that might provide insight about background experiences or characteristics, affective skills, education-related experiences in and out of the classroom, and/or the implementation of the CCSS?

ATTACHMENTS

**Attachment A** provides background information on the establishment of the two Common Core State Standards Assessment Consortia as well as the two consortia addressing alternate assessments for students with disabilities. It also includes information on the work being done on English language learner assessments via the ASSETS consortium, but does not include summary information on the ELDA21 consortium, both of which were addressed at our May Board meeting in L.A.

**Attachment B** includes a brief summary about PARCC and the bio for the presenter, Jeff Nellhaus.

**Attachment C** includes a brief summary about SBAC and the bio for the presenter, Joe Willhoft.
Background

The Race to the Top Assessment Program, authorized under the American Recovery and Reinvestment Act of 2009 (ARRA), provided funding to consortia of States to develop assessments that are valid, support and inform instruction, provide accurate information about what students know and can do, and measure student achievement against standards designed to ensure that all students gain the knowledge and skills needed to succeed in college and the workplace. These assessments are intended to play a critical role in educational systems; provide administrators, educators, parents, and students with the data and information needed to continuously improve teaching and learning; and help meet the President's goal of restoring, by 2020, the nation's position as the world leader in college graduates.

In September of 2010, the U.S. Department of Education awarded two Comprehensive Assessment Systems grants to the Partnership for Assessment of Readiness for College and Careers (PARCC) Consortium and the Smarter Balanced Assessment Consortia (SBAC). The consortia are to develop and implement assessment systems by the 2014-2015 school year. In addition, PARCC and SBAC were each provided a supplemental grant award to support the work in their approved application and to successfully transition to the new standards and assessments. Each received a supplemental grant award to include activities that focused on:

- Developing gap analyses between current and new standards, curriculum analysis tools, professional development related to the new standards and assessments including support for educators to better understand the content of the new standards, state and local assessment audits to determine what assessments will no longer be needed.
- Enhancing technology to be used in the assessments systems, including assessment delivery.
- Supporting educator understanding and use of assessment results, and other steps needed to build the professional capacity to implement more rigorous common standards.

On January 7, 2011, PARCC and SBAC each entered into a Cooperative Agreement with the Department regarding these grants. The agreement is intended to provide for substantial communication, coordination, and involvement between the Department and the grantee to ensure the success of the grant.

Also in 2010 the Office of Special Education Programs, U.S. Department of Education offered competitive grants to spur the development of a new generation of alternate assessments for students with disabilities to be jointly developed and used by groups of states. Grants were awarded to two consortia — the Dynamic Learning Maps Alternate Assessment Consortium (DLM) and the National Center and State Collaborative (NCSC). These new alternate assessments will be aligned to the Common Core State Standards (CCSS) and are expected to fit cohesively within the comprehensive assessment systems under development PARCC and SBAC. Both DLM and NCSC are to be ready for use by the 2014-15 school year, the same year in which the comprehensive assessment systems will be operational.

- The Dynamic Learning Maps Alternate Assessment Consortium (DLM) received a $22 million award. Thirteen (13) member states include Iowa, Kansas, Michigan, Mississippi, Missouri, New Jersey, North Carolina, Oklahoma, Utah, Virginia, Washington, West Virginia, and Wisconsin serving approximately 60,000 students who require an alternate assessment. The University of Kansas Center for Educational Testing and Evaluation (CETE) serves as the host, fiscal agent, and project management lead in partnership with member states and three additional partner organizations: the University of North Carolina at Chapel Hill on professional development and support materials;
Edvantia, Inc., on alternate standards definitions and project evaluation; and The Arc on the reporting system and dissemination.

- The National Center and State Collaborative (NCSC) received a $45 million award. Nineteen (19) member states/jurisdictions include Alaska, Arizona, Connecticut, District of Columbia, Florida, Georgia, Indiana, Louisiana, Massachusetts, Nevada, New York, North Dakota, Pacific Assessment Consortium (PAC-6), Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, and Wyoming serving approximately 90,000 students who participate in an alternate assessment based on alternate achievement standards. The National Center for Educational Outcomes at the University of Minnesota is the host fiscal agent and leads the Project Management Team. Four additional organizations also provide leadership: UKY on professional development; NCIEA on assessment design; UNCC on curriculum and instruction; and edCount, LLC, on evaluation.

To support the development of English proficiency assessments, the U.S. Department of Education’s 2011 competitive Enhanced Assessment Grant provided funding for the development of new assessments by a state consortium. In September 2011 the sole award of $10.5 million was given to the Wisconsin Department of Public Instruction in collaboration with the World-Class Instructional Design and Assessment (WIDA) Consortium. The assessment system under development, called Assessment Services Supporting ELs through Technology Systems (ASSETS) is to be ready for use by the 2015-16 school year. Twenty-nine (29) member states/jurisdictions include Alabama, Delaware, the District of Columbia, Idaho, Illinois, Maine, Maryland, Massachusetts, Minnesota, Mississippi, Missouri, Montana, Nevada, New Hampshire, New Jersey, New Mexico, North Carolina, North Dakota, Oklahoma, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Utah, Vermont, Virginia, Wisconsin, and Wyoming. WIDA at the Wisconsin Center for Education Research serves as the project management partner. Other organizations have major responsibilities. They include: the Center for Applied Linguistics for item and test development; WestEd for accommodations, validation, and interoperability; the University of California, Los Angeles (UCLA) for language learning progressions development and validation research; Data Recognition Corporation for field testing; and MetriTech for scoring.
The Partnership for Assessment of Readiness for College and Careers (PARCC) is a consortium of 21 states and the District of Columbia working together to develop a common set of K-12 assessments in English and math anchored in what it takes to be ready for college and careers. These new K-12 assessments will build a pathway to college and career readiness by the end of high school, mark students’ progress toward this goal from 3rd grade up, and provide teachers with timely information to inform instruction and provide student support. The PARCC assessments will be ready for states to administer during the 2014-15 school year. PARCC received a $186 million grant through the U.S. Department of Education’s Race to the Top assessment competition to support the development and design of the next-generation assessment system.

The PARCC Vision

PARCC states have committed to building a K-12 assessment system that:

- Builds a pathway to college and career readiness for all students,
- Creates high-quality assessments that measure the full range of the Common Core State Standards,
- Supports educators in the classroom,
- Makes better use of technology in assessments, and
- Advances accountability at all levels.

PARCC States

Collectively the 21 states and District of Columbia in PARCC educate about 24 million students and include 16 of the 22 Race to the Top winners. The PARCC states include: Alabama, Arizona, Arkansas, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Louisiana, Maryland, Massachusetts, Mississippi, New Jersey, New Mexico, New York, North Dakota, Ohio, Oklahoma, Pennsylvania, Rhode Island, and Tennessee.
**PARCC 2013-2013 Progress**

In the past year, PARCC has made significant progress to build its assessment system by contracting with vendors for the following:

- Creating the architecture to build PARCC assessment technology infrastructure
- Development of assessment items and tasks
- Facilitation and management of Educator Leader Cadres
- Item tryouts, field test & operational forms construction
- Project Management

**Biography of Jeffrey Nellhaus**

Jeffrey Nellhaus joined Achieve as Director of PARCC Assessment. PARCC (Partnership for the Assessment of Readiness for College and Careers) is one of two state consortia that received Race to the Top Assessment Grants to design and develop next generation student assessment systems based on the Common Core State Standards in English Language Arts & Literacy and Mathematics.

Before joining Achieve, Mr. Nellhaus spent nearly 25 years with the Massachusetts Department of Elementary and Secondary Education where held the positions of Deputy Commissioner, Acting Commissioner, and Associate Commissioner for Curriculum and Assessment. While at the MA DESE, Mr. Nellhaus directed the design, development and implementation of the Massachusetts Comprehensive Assessment System (MCAS), and the development of the Massachusetts Curriculum Frameworks, which include the Common Core State Standards. For his work on MCAS he was awarded the Manuel Carballo Governor’s Award for Excellence in Public Service.

Mr. Nellhaus has served on the National Validity Studies Panel to National Assessment of Education Progress (NAEP) and on Technical Advisory Committees for the states of Maine, Kentucky and Rhode Island. He has also served on the Technical Advisory Committee on Standard Setting for NAEP and on the Growth Model Peer Review Panel for the U.S. Department of Education.

Prior to joining the Massachusetts Department of Education, Mr. Nellhaus was a Peace Corps Volunteer in India, taught chemistry and mathematics in a public high school in Vermont, and directed a federally-funded educational program in Thailand for Cambodian and Laotian refugees preparing to resettle in the U.S.

Mr. Nellhaus holds a B.S. in Chemistry from the University of Massachusetts, a M.S. in Science Teaching from Antioch Graduate School of Education, and an Ed.M. in Administration, Policy and Planning from Harvard Graduate School of Education.
The SMARTER Balanced Assessment Consortium (SBAC) is a national consortium of states that have been working collaboratively since December 2009 to develop a student assessment system aligned to a common core of academic content standards to apply for a Race-to-the-Top Assessment grant. On Sept. 2, 2010, the SBAC was awarded a four-year $176 million Race to the Top assessment grant by the US Department of Education (USED) to develop a student assessment system aligned to a common core of academic standards.

SBAC will create state-of-the-art adaptive online exams, using “open source” technology. The online system will provide accurate assessment information to teachers and others on the progress of all students, including those with disabilities, English language learners and low- and high-performing students. The system will include:

1. the required summative exams (offered twice each school year);
2. optional formative, or benchmark, exams; and
3. a variety of tools, processes and practices that teachers may use in planning and implementing informal, ongoing assessment. This will assist teachers in understanding what students are and are not learning on a daily basis so they can adjust instruction accordingly.


Smarter Balanced member states educate more than 19 million of the nation’s public K to 12 students. These states share a commitment to developing a next-generation assessment system aligned to the Common Core State Standards that provide educators with meaningful feedback and actionable data.

Biography of Joe Wilhoft

Mr. Wilhoft is the Executive Director of the SMARTER Balanced Assessment Consortium, one of two federally-funded consortia that are developing a new generation of state assessment systems aligned to the Common Core State Standards. Prior to this appointment he was the assistant superintendent for assessment and student information for the state of Washington. His responsibilities included design and implementation of Washington’s assessment program and collection and reporting of student information for the state’s longitudinal data system. Before working at the state level, Joe directed assessment and evaluation activities at the district level for more than twenty years, primarily in Tacoma Public Schools in Washington and in Maryland.

Joe earned his doctorate in educational measurement and statistics from the University of Maryland. He is past president of the Maryland Assessment Group, the Washington Educational Research Association, and the American Educational Research Association Classroom Assessment Special Interest Group. He
has been involved in multiple collaborative data and assessment efforts, including having served on technical advisory committees in several states and the Technical Work Group for the most-recent congressionally-mandated evaluation of the National Assessment of Educational Progress (NAEP). He currently co-chairs the NAEP Policy Advisory Task Force, a collaborative effort of the National Assessment Governing Board and the Council of Chief State School Officers.
Interpreting NAEP Results Using Preparedness Research Findings

At the August 2013 meeting, the Governing Board will discuss the way forward on reporting the NAEP 12th grade results from the 2013 reading and mathematics assessments. As background for the discussion, included in this tab are:

- a draft of a prototype chapter for the report (Attachment 1; new document)
- the independent technical reviews of the preparedness validity argument by Gregory Cizek and Mark Reckase (Attachments 2 and 3; new documents)
- the preparedness validity argument (Attachment 4; included in the May 2013 COSDAM briefing materials, but changes were made to the proposed inferences as described below and indicated in highlighting on pages 8 and 41)

The draft prototype chapter was prepared as an example of what NAEP reporting on academic preparedness for college would look like in the report of the 2013 12th grade assessment results.

As previously reported to the Governing Board, the Board staff and NCES staff have been working collaboratively since March 2013 to develop options for reporting NAEP 12th grade results based upon the preparedness research findings. The options ranged from merely providing information about the 12th grade preparedness research and findings to reporting 12th grade results using statements (inferences) about 12th grade students’ academic preparedness.

After the May 2013 Board meeting, at which the Board reviewed the draft validity argument, the two staffs met and agreed that the next step should be to use the guidance from the Board discussion on the validity argument and prepare a prototype chapter for the report. This would provide something specific and concrete as a basis for further Board discussion.

The Board staff drew two main conclusions from the Board discussion in May about the validity argument:

- While finding the validity argument supportive, the Board wanted to consider the independent technical reviews that were to be presented at the August 2013 meeting to inform its decision making.
- The Board found the inferences that were being proposed to be “not quite there yet.”

The inference proposed in May was of the form “12th grade students scoring at or above Proficient are likely to be academically prepared…” Because “likely” was not quantitatively defined, the Board found this formulation ambiguous and potentially confusing to the public. During the discussion, Board member Andrew Ho said he was proposing a solution that he would share with staff. Mr. Ho proposed an inference of the general form as follows:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness, the percentage of students scoring at or above Proficient on Grade 12 NAEP is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities that would make them academically prepared for college.
Mr. Ho’s formulation for the preparedness inference was shared with Michael Kane, who is advising Board staff on the validity argument. Mr. Kane supported using this formulation in place of the one originally proposed and suggested adding “or reasonable” after “plausible.” Board staff revised the validity argument accordingly and it was this formulation that was considered by the independent technical reviewers of the validity argument.

**Question for Board Consideration:**

With the understanding that additional work will be required in collaboration with NCES, along with additional guidance from the Board, is the general approach exemplified in the prototype chapter (Attachment C1) an acceptable basis for moving forward with reporting on academic preparedness for college as a part of the reporting of the NAEP 12th grade reading and mathematics assessment results for 2013?
Towards NAEP as an Indicator of Academic Preparedness for College and Job Training
Ray Fields July 18, 2013

For over a decade, the National Assessment Governing Board has been conducting research to enable 12th grade NAEP to serve as an indicator of academic preparedness for college and job training. This chapter provides the rationale for pursuing this goal; the research results from studies conducted in connection with the 2009 administration of 12th grade NAEP; and the implications for NAEP 12th grade reporting.

INTRODUCTION
Indicators of many kinds are used to monitor critical aspects of national life and inform public policy. These include economic indicators (e.g., gross domestic product), health indicators (e.g., cancer rates), and demographic indicators (e.g., population trends by race/ethnicity and gender).

NAEP serves the public as a national and state indicator of education achievement at the elementary and secondary levels. NAEP monitors student achievement at key points in the elementary/secondary progression: grades 4, 8, and 12.

According to the National Assessment Governing Board, the 4th grade is the point at which the foundations for further learning are expected to be in place (e.g., when “learning to read” becomes “reading to learn”).

The 8th grade is the typical transition point to high school.

The 12th grade is the end of the K-12 education experience, the transition point for most students to postsecondary education, training, the military, and the work force. (Draft Policy Statement on NAEP).

NAEP is the only source of nationally representative 12th grade student achievement results. State tests of academic achievement are usually administered before 12th grade and are quite different across the country. Likewise, college admission tests like the ACT and SAT are generally taken before 12th grade by a self-selected sample and therefore, are not representative of all 12th graders.

Consequently, NAEP is uniquely positioned to serve as an indicator of academic preparedness for college and job training at grade 12—the point that represents the end of mandatory schooling for most students and the start of postsecondary education and training for adult pursuits.

A wide array of state and national leaders has embraced the goal that 12th grade students graduate “college and career ready.” These include the leadership and members of the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), the Business Roundtable (BRT), the U.S. Chamber of Commerce (the Chamber), a task force on education reform of the Council on Foreign Relations, and state and national political leaders. (Fields and Parsad).

NAEP and ACADEMIC PREPAREDNESS
The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would afford an invaluable public service: providing an indicator of the human capital potential of today’s and future generations of the nation’s population.
The Board began this initiative in 2004, after receiving recommendations from a distinguished blue-ribbon panel that had examined whether NAEP should continue assessing at the 12th grade.

The panel stated that “America needs to know how well prepared its high school seniors are... [only NAEP] can provide this information...and it is necessary for our nation’s well-being that it be provided.” The panel recommended that NAEP continue to assess at grade 12 and that the 12th grade assessment be transformed to measure preparedness for college, job training, and the military. (National Commission on NAEP 12th Grade Assessment and Reporting; p. 2.)

To transform 12th grade NAEP into an indicator of academic preparedness, the Governing Board took several significant steps.

1. The Board determined that measuring academic preparedness for college and job training should be an intended purpose of 12th grade NAEP.

2. The Board contracted with Achieve, Inc., in 2005 to review the NAEP 12th grade reading and mathematics assessment frameworks and identify where changes, if any, would be needed. Modest changes were recommended.

3. Accordingly, the Board made changes to the frameworks to be used for the administrations of the 12th grade assessments, scheduled for 2009 and 2013.

4. In 2006, the Governing Board assembled a team of noted psychometricians, industrial/organizational psychologists, and K-12 and postsecondary researchers to serve as a technical panel, advising on validity research to conduct.

5. In 2008, the technical panel recommended a comprehensive program of research. The validity of statements about academic preparedness in NAEP reports would be affected by the degree to which the results were mutually confirming.

Figure 1. presents a model of the research program, with five types of research displayed, the interrelationships that would be examined, and the potential meaning of the research results in terms of the NAEP score scale.

Figure 1 about here (see page 8)

6. The Governing Board began contracting for the research studies in 2008, in connection with the 2009 administration of the 12th grade reading and mathematics assessments. More than 30 research studies were completed during the period 2009-2012.

The Research Findings
The research findings were consistent across studies and across years. For example, the content of the 12th grade NAEP reading and mathematics assessments was found to be similar to widely recognized tests used for college admission and placement (see http://www.nagb.org/what-we-do/preparedness-research/types-of-research/content-alignment.html).

Performance by the same students on NAEP and the SAT mathematics and reading tests was correlated at 0.91 and 0.74, respectively.
Statistical linking studies examining performance on NAEP and the college admission tests found that the college readiness benchmarks set for the ACT and SAT reading and mathematics were in a range around the Proficient achievement levels on the 12th grade NAEP reading and mathematics assessments. For example, the average NAEP reading score of students scoring at the SAT benchmark was 301, not significantly different from the cut-score for Proficient of 302 (see Fig. 2 and 3).

A longitudinal study followed a representative sample of Florida 12th grade NAEP test-takers into the state’s public colleges (see Fig. 2 and 3). The longitudinal study permitted an analysis of performance on NAEP and actual student outcomes. In the first year of this study, an analysis was conducted of performance on NAEP and (1) enrollment in regular versus remedial courses, and (2) first year overall college grade point average (GPA). As with the other statistical studies, the average NAEP score of the students who were not placed into remedial courses or who had a first year college GPA of B- or better was in a range around the 12th grade reading and mathematics Proficient achievement levels.

Results from the more than 30 studies were used to develop a validity argument to support proposed inferences (claims) about academic preparedness for college in relation to student performance on 12th grade NAEP. The validity argument was reviewed by two independent technical reviewers. The technical reviewers concluded that the validity argument supports the proposed inferences.

The complete research reports and the validity argument, along with the two independent technical reviews, can be found at http://www.nagb.org/what-we-do/preparedness-research.html.

Although the research results support inferences about NAEP performance and academic preparedness for college, the research results to date do not support inferences about NAEP performance and academic preparedness for job training.

A second phase of NAEP preparedness research began in 2013 and is expected to be completed in time for reporting 12th grade results in 2015. The second phase of research results will be examined to determine the degree to which they confirm existing results.

A TRANSITION TO REPORTING ON ACADEMIC PREPAREDNESS
The reporting of the 12th grade results for 2013 represents a transition point for NAEP.

The interpretations of the 2013 NAEP 12th grade reading and mathematics results related to academic preparedness for college set forth in this report are considered foundational and subject to adjustment in the future.

These interpretations are included in this report because the independent technical reviewers found them to be technically defensible, but more importantly, to promote public discussion about their meaningfulness and utility.

The Context for Academic Preparedness for College
In the United States in 2013, there is no single, agreed upon definition of “academic preparedness for college” used by colleges for admission and placement (Fields and Parsad). Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.
In this complex mix are 2-year institutions, 4-year public and private institutions with a wide range of selectivity, and proprietary schools. Institutions range from highly selective (i.e., with admission criteria including very high grade point averages, successful completion of rigorous high school coursework and very high SAT and/or ACT scores) to open admission (i.e., all applicants are admitted).

Even within institutions, requirements may vary across majors or programs of study. For example, the mathematics and science high school coursework and academic achievement needed for acceptance into an engineering program in a postsecondary institution may be more rigorous than the general requirements for admission to the institution or for a degree in elementary education in that institution.

**Defining Academic Preparedness for College**

Given the diversity of postsecondary education institutions, it is essential to provide a reasonable definition of academic preparedness for NAEP reporting. The definition should be relevant to NAEP’s purpose of providing group estimates of achievement. (It is important to note that NAEP does not provide individual student results.) The definition should be meaningful to NAEP’s primary audiences: the general public and national and state policymakers.

The definition proposed in this report is intended to apply to the typical degree-seeking entry-level student at the typical college. For NAEP reporting, “academically prepared for college” refers to the reading and mathematics knowledge and skills needed for placement into entry-level, credit bearing, non-remedial courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

It is important to note the focus on “placement” rather than “admission.” This distinction is made because students who need remedial courses in reading, mathematics or writing may be admitted to college, but not placed into regular, credit-bearing courses. The criterion of importance is qualifying for regular credit-bearing courses, not admission.

The definition is not intended to reflect
- academic requirements for highly selective postsecondary institutions;
- the additional academic requirements for specific majors or pre-professional programs, such as mathematics, engineering, or medicine; or
- academic requirements applicable to entry into certificate or diploma programs for job training or professional development in postsecondary institutions.

The definition is focused on the first year of college; it does not address college persistence beyond the first year or completion of a degree. The definition will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.

**Proposed Inferences for NAEP Reporting**

The NAEP preparedness research does not affect the NAEP results in any way. The distribution of student achievement is unchanged. That is, the average scores, the percentiles, and the achievement level results are not impacted by the NAEP preparedness research.
The independent technical reviewers confirmed that the research findings support inferences about performance on NAEP 12th grade results in reading and mathematics in relation to academic preparedness for college.

**Proposed Inferences**

In the NAEP/SAT linking study for reading (Figure 2), the average NAEP score for 12th grade students scoring at the SAT college readiness benchmark for critical reading is 301, not significantly different from the Proficient cut-score of 302. The results from the Florida longitudinal study are confirmatory.

These data, together with the content analyses that found NAEP reading content to be similar to college admission and placement tests, support the inference for reading that Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

**the percentage of students scoring at or above a score of 302 (Proficient) on Grade 12 NAEP in reading is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.**

In 2013, XX% of 12th graders nationally scored at or above 302 (Proficient) in reading.

The study results support these inferences. However, there will be students scoring at or above Proficient who are not academically prepared and students scoring below Proficient who are academically prepared (i.e., there will be false positives and false negatives). This will be true for any assessment program that sets cut-scores for a similar purpose.

In the NAEP/SAT linking study for mathematics (Figure 3), the average NAEP score for 12th grade students scoring at the SAT college readiness benchmark for mathematics is 163, lower than and significantly different from the Proficient cut-score of 176. The results from the High School Transcript Study and the Florida longitudinal study are confirmatory.

These data, together with the content analyses that found NAEP mathematics content to be similar to college admission and placement tests, support the inference for reading that

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships...
between NAEP scores and NAEP content to other relevant measures of college academic preparedness, the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

In 2013, XX% of 12th graders nationally scored at or above 163 in mathematics.

To consider the plausibility of these estimates, comparisons can be made with the percentages of students who met the ACT or SAT college readiness benchmarks.

Information is available about students who were seniors in 2009 (ACT) and in 2010 (SAT). Thus, the ACT data are for the same student cohort as the NAEP data, but the SAT data are for a cohort that followed one year later.

It also must be noted that, unlike the NAEP results, neither the ACT nor the SAT results represent all 12th graders. Further, there is overlap among ACT and SAT test-takers, with about 20% estimated to take both tests.

Assuming that a substantial portion of students who do not take either test are not academically prepared for college, it is not inconsistent that the NAEP percentages are lower than those for the respective college readiness benchmarks.

<table>
<thead>
<tr>
<th>Percentages* Scoring at/above ACT and SAT College Readiness Benchmarks and at/above Proficient in Reading on NAEP and at/above 163 in Mathematics on NAEP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
</tr>
<tr>
<td>ACT (2009)</td>
</tr>
<tr>
<td>SAT (2010)</td>
</tr>
<tr>
<td>NAEP (2009)</td>
</tr>
</tbody>
</table>

* About 48% of 12th graders took the ACT or SAT. NAEP represents 100% of 12th graders.

Limitations on Interpretation and Other Caveats

False Negatives and False Positives
Some proportion of 12th grade students scoring below Proficient on the 12th grade NAEP Reading or below a score of 163 on the Mathematics Assessment are
- likely to be academically prepared for college
- not likely to need remedial/developmental courses in reading or mathematics in college,

but with a lower probability than those at or above Proficient in reading or 163 in mathematics.

In addition, some proportion of 12th grade students scoring at or above Proficient on the 12th grade NAEP Reading or 163 on the Mathematics Assessment may not
- be academically prepared for college
- need remedial/developmental courses in reading or mathematics in college.

Not a Preparedness Standard
The proposed inferences are not intended to represent or be used as standards for minimal academic preparedness for college. The proposed inferences are intended solely to add meaning to interpretations of the 12th
grade NAEP reading and mathematics results in NAEP reports.

GPA of B- or Better
The variable “first-year GPA of B- or better” was selected because of its use as a research-based criterion in defining college readiness benchmarks developed for the SAT by the College Board. The College Board had agreed to partner with the Governing Board in a study linking performance on 12th grade NAEP with the SAT. Another leader in college testing programs, ACT, Inc. has developed similar benchmarks for its college admission assessments using a similar criterion and similar methodology. Because they are based on credible research related to college outcomes, and because performance on the respective tests could be linked to performance on NAEP, the college readiness benchmarks used by these testing programs were relevant, useful points of reference for the NAEP preparedness research.

The College Board has set a score of 500 on the SAT Mathematics and Critical Reading tests as its college readiness benchmarks in those areas. Based on its research, the College Board has determined that the score of 500 predicts, with a probability of .65, attainment of a first-year overall GPA of B- or higher. Similarly, the ACT college readiness benchmarks are based on research indicating a .50 probability of attaining first-year grades in relevant courses (e.g., college algebra and courses requiring college level reading) of B or better and .75 probability of C or better.

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

Data Limitations
The NAEP preparedness research studies are comprehensive and the results consistent and mutually confirming, but, for reading the statistical studies are limited to one year for data at the national level and to one state-based longitudinal study. For mathematics, there are two separate years of data at the national level and one state-based longitudinal study. Therefore, more evidence exists to support the plausibility of inferences related to mathematics than to reading.

Preparedness for Job Training
The completed research with respect to academic preparedness for job training does not support conclusions relative to the NAEP scale. Plans for future research will be reviewed by the Governing Board.

Conclusion
The independent technical reviewers found the Governing Board’s preparedness research to be methodical, rigorous, and comprehensive. They concluded that the research findings support the use of the proposed inferences in NAEP reports about 12th graders’ academic preparedness for college.

The interpretations of NAEP results in relation to academic preparedness for college are being reported on a preliminary basis. They are provided to help foster public
understanding and policy discussions about defining, measuring, validating and reporting on academic preparedness for college by NAEP and more broadly.

Including these inferences in NAEP 12th grade reports is intended to add meaning to the interpretation of the NAEP 12th grade results. However, the potential for misinterpretation exists. For these reasons, the section above on limitations on interpretation and other caveats is included in this chapter.

The Governing Board will monitor the use of these inferences as well as unintended consequences arising from their use as a part of the next phase of the preparedness research.

The next phase of the preparedness research is being conducted in connection with the NAEP reading and mathematics assessments administered in 2013. The research results will be used as additional validity evidence in relation to NAEP reporting on 12th grade academic preparedness.

Figure 1. Model of the Preparedness Research Program
Figure 2.

**NAEP 12th-Grade Preparedness Research: Reading**

Average Scores and Inter-quartile Ranges for Selected Variables, SAT and ACT College Readiness Benchmarks From the 2009 NAEP/SAT Linking Study and 2009 Florida Longitudinal Study

Figure 3.

**NAEP 12th-Grade Preparedness Research: Mathematics**

Average Scores and Inter-quartile Ranges for Selected Variables, SAT and ACT College Readiness Benchmarks From the 2009 NAEP/SAT Linking Study, 2006 High School Transcript Study, 2009 High School Transcript Study, and 2009 Florida Longitudinal Study
References to be added.
Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College

Prepared for:
National Assessment Governing Board

Prepared by:
Gregory J. Cizek, PhD
Professor of Educational Measurement and Evaluation
University of North Carolina at Chapel Hill
cizek@unc.edu

June 30, 2013
Review and Comment on

Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College

Introduction

The National Assessment Governing Board (NAGB) sought input on the constellation of logical and empirical evidence it has amassed in support of certain claims centering on how scores on the 12th Grade National Assessment of Educational Progress (NAEP) might be interpreted with respect to college preparedness. The logic underlying those claims and the logical and empirical support for the claims can be referred to as the validity argument.

According to Kane (2013):

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (p. 1)

The remainder of this paper presents the preparedness score interpretation claims proposed for the 12th grade NAEP scores and an overall an evaluation of the plausibility of those claims.

To produce this evaluation, I relied primarily on two documents that presented the NAEP preparedness validity argument and evidence (Fields, 2013a, 2013b). A draft response to Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College (Fields,
2013a) was submitted to the National Assessment Governing Board on May 29, 2013 (Cizek, 2013). This paper is a response to a revision of *Validity Argument for NAEP Reporting on 12th Grade Academic Preparedness for College* (Fields, 2013b)

**The Proposed Interpretations and Claims**

The proposed score interpretations related to college preparedness for NAEP Reading and Mathematics are the following:

**READING** – "The percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college."

**MATHEMATICS** – "The percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college." (Fields, 2013b, p. 8)

The proposed interpretations are grounded in four claims (taken from Fields, 2013b):

1. The 12th grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

2. Performance on 12th grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

**Evaluation of Validity Evidence in Support of the Proposed Interpretations**

Overall, my review and analysis leads me to conclude that the logical and empirical evidence amassed provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear and coherent. The proposed interpretations are warranted in two ways: 1) by the accumulation of confirming evidence that is uniformly in the direction that would be hypothesized by the proposed interpretations; and 2) by the paucity of disconfirming evidence. On this point, it is noteworthy that the present validation effort appeared to be searching, objective, and contemplated the potential for disconfirming evidence.

It is my opinion, based on the evidence provided, that future NAEP reporting can provide reasonably confident and accurate indications of college preparedness in Reading and Mathematics.

It should be recognized, of course, that validation efforts typically should not be considered final or complete at any given juncture (see Cizek, 2012). Additional data can be gathered; additional experience with the test is gained; theory related to (in this case) college preparedness evolves; and new relationships among variables can be explored. The following three recommendations suggest additional validation strategies or evidential sources that may have the potential to strengthen warrants for the intended preparedness score interpretations.

1) To enhance the clarity of the proposed interpretations, I offer the following recommendation: *NAGB should consider making the score interpretations parallel by specifying the NAEP scale score associated with preparedness in Reading.*
As currently worded, a defensible and specific scale score associated with preparedness is offered for NAEP Mathematics score interpretations; however, the interpretation for Reading is phrased as an achievement level: “The percentage of students in the 12th grade NAEP distribution at or above (Proficient for reading and a score of 163 for mathematics) is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in (reading or mathematics) that would make them academically prepared for college.”

The lack of parallelism in construction seems awkward, unnecessary, and potentially confusing to readers and users of this information. I recommend expressing both the Reading and Mathematics interpretations as NAEP scale scores, with elaboration as achievement levels if desired. An example of a slightly reworded interpretation along these lines would be:

“The percentage of students in the 12th grade NAEP distribution at or above a scaled score of XXX (Proficient) in Reading and a score of 163 in Mathematics is a plausible estimate of the percentage of students who possess the knowledge, skills, and abilities in those subjects that would make them academically prepared for college.”

2) To enhance the coherence of the proposed interpretations, I offer the following recommendation: \textit{NAGB should consider conducting additional research into the content coverage of the NAEP and the alignment of NAEP with traditional college admissions measures.}
In its present form, it is argued that, in essence, the content of NAEP assessments in Reading and Mathematics covers everything that traditional college admissions measures (e.g., ACT, SAT, etc.) do, but also more. It is claimed that NAEP content coverage is "broader." The Venn diagram below illustrates this claim:

![Venn Diagram](image)

**Figure 1**

*Hypothetical content coverage between NAEP Assessment and College Admissions Assessment*

Figure 1 illustrates (ignoring the relative size of the circles) the claim that NAEP is somewhat of an umbrella assessment in terms of content coverage compared to the traditional college admissions measures on which alignment research has already been conducted. However, it is not clear that the fact that an umbrella relationship exists unequivocally supports the claim that NAEP assessments capture the same things about college preparedness as the college admissions tests or, importantly, that conclusions based on such alignment can unambiguously be made with respect to preparedness. For example, it would be theoretically possible for an examinee could score "Proficient" on

---

1 The Venn diagram and the reference to the SAT are presented only illustrate content relationships between content coverage on assessments. The diagram is not intended to represent the actual proportional content coverage between NAEP and college admissions assessments, nor that of the SAT in particular.
NAEP Reading (and be deemed prepared for college) by getting very little of the "SAT-like" content correct on NAEP (that is, content deemed necessary for college success) and getting a lot of the “other” NAEP content correct (that is, the additional/broader content that may or may not necessarily be relevant to college preparedness).

3) To enhance the comprehensiveness of the proposed interpretations, I offer the following recommendation: NAGB should consider conducting additional research into the predictive validity of the NAEP with respect to college success.

Perhaps the most important variable assessed in the validation of traditional college admissions assessments is the ultimate criterion of college success—typically operationalized as first year GPA, persistence, or some other variable. Although the validity evidence gathered so far links NAEP scores to scores on other measures that are, in turn, linked to college success, the present validity case for NAEP preparedness does not do so directly. For the future, independent evaluations of direct evidence regarding the extent to which NAEP preparedness scores are associated with college criterion outcomes would substantially bolster the evidence in support of the intended score interpretations.

Conclusion

The logical and empirical evidence gathered to date provides strong support for the proposed 12th Grade NAEP Reading and Mathematics score interpretations related to academic preparedness for college. The case for the validity of the interpretations is clear, coherent, and comprehensive. Recommendations were presented for future strategies to strengthen the validity
case. Nonetheless, based on the empirical evidence and logical rationales to date, there appear to be strong warrants for the intended interpretations regarding NAEP reporting and indications of college preparedness in Reading and Mathematics.

References


Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use, Psychological Methods, 17(1), 31-43.


Beginning in March, 2004, the National Assessment Governing Board (NAGB) began work to support the use of the 12th grade National Assessment of Educational Progress (NAEP) as a measure of “preparedness” of students for academic work at the college level. There are many challenges to this work, but one of the most important is to show that there is validity evidence to support the inference that students who are estimated to be above a specified level on the NAEP reporting score scale have the skills and knowledge to profit from credit-bearing, first-year college level coursework.

During the nine year period of this effort, the thinking about the way that validity evidence is collected and reported has had some significant changes. Particularly over the last few years, the work of Michael Kane (e.g., Kane, 2013) has provided guidance about how to present validity evidence for the interpretation of the results of an academic test in the form of what is now called a “validity argument.” The document that I reviewed was one of the first that I have seen that takes this approach to heart and makes a highly credible effort to apply this perspective on validation. In one sense, this is not surprising because work on NAEP has tended to be at the forefront of innovative psychometrics, be it on the use of item response theory procedures or standard setting. In another sense, it is surprising that NAGB has adopted this approach because there are few practical models for the creation of a validity argument. Even though there may have been some risk in being among the first to report support for an inference using the validity argument, this document is quite successful at providing a well supported validity argument. It gives other testing programs a very nice model for future reports on the validation of inferences from test scores.

My general view is that this document presents solid support for the inference that the proportion of the examinee population that is estimated to be above the specified cut score on the NAEP reporting score scale meets the definition of “preparedness for credit-bearing, first-year college coursework.” The evidence that was collected to support the inference is quite extensive and the connection of the evidence to the argument is logical and compelling. There are also appropriate cautions about over interpretation of results. It is very nice to see the areas of weakness in the supporting documents as well as the strengths. This adds credibility to the conclusions from the validity argument. This is not to say that the argument could not be tightened and elaborated, but this is an impressive example of a validity argument for a complex inference from a complex assessment.
A More Detailed Analysis

Although I have a very positive reaction to the report, it is important to probe the specifics of the argument and the claims being made. This may be interpreted as a desire for even more detail than is given in the report, but there is always a need for balance between detail and clear communication. The report is already long and detailed. I am reluctant to suggest adding more to it. But I do want to highlight some specific issues about some of the assumptions and claims in the argument.

The following statement is the basic inference that is the focus of the argument.

“The percentage of students in the NAEP distribution at or above a particular score level in reading or mathematics on 12th grade NAEP is a plausible, or reasonable, estimate of the percentage of 12th grade students who are academically prepared for college.” (P. 6)

This statement is very rich in meaning. To fully understand it, some background information is assumed to be known by the reader. Some of this background is listed here, but the list may not be comprehensive.
1. NAEP produces an accurate representation of the distribution of achievement of students in the areas of reading and mathematics.
2. The estimate of the proportion of students above a cut score on the NAEP reporting score scale is fairly accurate.
3. Students who are estimated to be above the specified cut score are likely to have high school grades and college admissions test scores that will make them eligible for admission to college.
4. Those students who are eligible for admission attend college and enroll in entry-level, credit-bearing courses.
5. The skills and knowledge in reading and mathematics are prerequisite to learning the content presented in the entry-level, credit-bearing courses.

The first two entries in the list are well supported by the technical documentation for NAEP. There are many years of research studies and analyses that show the technical quality of the assessment program. The last three of the entries in the list are more difficult to support because NAEP does not provide accurate student level scores and the individual students who participate are usually not identified so their academic history following the NAEP administration cannot be recorded. It is here that the special studies and data collections that have been done by NAGB are important to fill in links of the validity argument.

A Slight Variation on the Validity Argument

During the process of reviewing the report on the validity argument, I took notes on component parts of the argument. In some cases, the purpose of the notes was to highlight assumptions that were not explicitly stated. In other cases, the purpose was to elaborate on a step in the validity argument. A summary of these notes in the form of a slightly different validity argument than the one given in the report is given below. This is not meant to imply a problem.
with the validity argument in the NAGB report, but rather to add some commentary on that argument.

1. There is a body of knowledge and skills that is taught at the secondary school level that is prerequisite to gaining admission into entry-level, credit-bearing courses at colleges and universities.
   a. There seems to be strong evidence for this from the America Diploma Project and the analysis of the admission and placement tests.
   b. It might be helpful to think of this in terms of a Venn diagram that shows the intersection and union of the content descriptions from all of these different sources. The argument should be made that NAEP is based on a reasonable sampling of content from the intersection or the union.

2. College admissions test scores and high school transcripts provide information about the prerequisite knowledge and skills and these are used to make decisions about admissions to the entry-level courses.
   a. This is easy to document, but it is not explicitly stated in the argument. Of course, different institutions use the information in different ways.

3. The knowledge and skills reflected in college admissions tests and high school transcripts that are prerequisite to the entry-level college courses can be described in some detail to allow the design of a test to assess the knowledge and skills.
   a. This is clearly supported by the information from the studies.
   b. It would be useful to have a summary description of the common components from all of the parts.

4. NAEP assessments provide information about student acquisition of the knowledge and skills described above.
   a. This is the main thrust of all of the content analysis.
   b. The argument is compelling, but it would be helpful to have a general content description that is the result of all of the content analysis.

5. There is a threshold value for the knowledge and skills defined above. If students do not meet this threshold, they will not be ready to take the entry level courses.
   a. The comparative data make a good argument for the existence of the cut score.

6. A cut score on NAEP is consistent with the threshold.
   a. There is a good process for identifying a reasonable cut score on NAEP to correspond to #5.
   b. The combination of information from different tests results in strong support for parts of the argument.
7. The proportion of students estimated to be above the cut score on NAEP gives a good estimate of the proportion who exceed the threshold for admission into entry level courses.
   a. This is well supported by the statistical analysis procedures if the argument for an appropriate cut score is supported. In this case, there is reasonable support for the cut score from the connection to placement and admissions tests.

From this argument, I believe that the following inference from NAEP reported results is supported: The proportion of students estimated to be above the specified cut score on the NAEP reporting score scale is a reasonable estimate of the proportion of students who have the prerequisite knowledge and skills in mathematics and reading to profit from entry-level, credit-bearing college courses.

Reference

Rationale for NAEP Reporting on 12th Grade Academic Preparedness

The National Assessment Governing Board is conducting a program of research to determine the feasibility of the National Assessment of Educational Progress (NAEP) reporting on the academic preparedness of U.S. 12th grade students, in reading and mathematics, for college and job training.

Since 1969, NAEP has reported to the public on the status and progress of student achievement in a wide range of key subjects at grades 4, 8, and 12. NAEP provides national and state-representative results, results for twenty-one urban districts, and results by subgroups of students (e.g., by race/ethnicity, gender, and for students with disabilities and English language learners). NAEP, by law, does not provide individual student results.

The Governing Board’s initiative on 12th grade academic preparedness began in March 2004, with the report of a blue-ribbon panel. The panel was composed of K-12 education leaders—the “producers” of high school graduates—and leaders in business, postsecondary education, and the military—the “consumers” of high school graduates.

The panel members recognized the importance of 12th grade as the gateway to postsecondary education and training, and viewed NAEP as a “truth teller” about student achievement. These distinguished state and national leaders recommended unanimously that “NAEP should report 12th grade students’ readiness for college-credit coursework, training for employment, and entrance into the military.” (National Commission on NAEP 12th Grade Assessment and Reporting; p. 6.). They stated that “America needs to know how well prepared its high school seniors are… [only NAEP] can provide this information…and it is necessary for our nation’s well-being that it be provided.” (Ibid. p. 2.).

The Governing Board approved this recommendation, with a minor modification. The term “readiness” was changed to “academic preparedness” and “entrance into the military” was subsumed by “job training.”

“Readiness” was changed to “academic preparedness” because “readiness” is broadly understood to include both academic preparedness and other characteristics needed for success in postsecondary education and training, such as habits of mind, time management, and persistence (Conley). NAEP does not purport to measure such characteristics. Rather, NAEP is designed to measure academic knowledge and skills.

1 The blue-ribbon panel was known officially as the National Commission on NAEP 12th Grade Assessment and Reporting.
“Entrance into the military” was subsumed by “job training” with the intention of identifying occupations with civilian and military counterparts and utilizing the military’s experience as the world’s largest occupational training organization and its extensive research on the relationship between performance on the Armed Service Vocational Aptitude Battery (ASVAB) and job training outcomes.

The Governing Board approved the 12th grade academic preparedness initiative because it believes that the academic preparation of high school students for postsecondary education and training is important to the nation’s economic well-being, national security, and democratic foundations (see Governing Board resolution of May 21, 2005 at http://www.nagb.org/content/nagb/assets/documents/policies/resolution-on-preparedness.pdf).

Indicators of many kinds are used to monitor critical aspects of national life and inform public policy. These include economic indicators (e.g., gross domestic product), health indicators (e.g., cancer rates), and demographic indicators (e.g., population trends by race/ethnicity and gender). The Governing Board believes that NAEP reporting on the academic preparedness of 12th grade students would serve as a valuable indicator of the human capital potential of rising generations of citizens, a nation’s greatest resource.

The Governing Board is not alone in recognizing the importance of 12th grade academic preparedness for the nation. A wide array of state and national leaders has embraced the goal that 12th grade students graduate “college and career ready.” These include the leadership and members of the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), the Business Roundtable (BRT), the U.S. Chamber of Commerce (the Chamber), the Council on Foreign Relations, and the Obama Administration. The reason for this attention to 12th grade academic preparedness is well summarized by a statement of the Business Coalition for Student Achievement, an organization coordinated by BRT and the Chamber:

“Ensuring that all students graduate academically prepared for college, citizenship and the 21st century workplace...is necessary to provide a strong foundation for both U.S. competitiveness and for individuals to succeed in our rapidly changing world.”

The NGA and CCSSO have collaborated to develop Common Core State Standards (CCSS) for mathematics and English language arts. These standards are aimed at fostering college and career readiness by the end of high school. The CCSS have been adopted formally by 45 states, several territories and the Department of Defense Education Activity. Viewing the need for rigor in education standards and outcomes through the lens of national security, a similar conclusion was made in the report of the Independent Task Force on U.S. Education Reform and National Security of the Council on Foreign Relations. The Task Force was co-chaired by former New York City School Chancellor Joel Klein and Former Secretary of State Condoleezza Rice. The Obama administration has stated that “educating every American student to graduate from high school prepared for college and for a career is a national imperative.” (Fields and Parsad; pp. 3-4).

Twelfth grade is the end of mandatory schooling for most students and represents the transition point to adult postsecondary pursuits. If it is essential for students to graduate from high school
academically prepared for college and job training, it is essential for the public and policymakers to know the degree to which this is occurring.

A trusted indicator is needed for reporting to the public and policymakers on the status of 12th grade academic preparedness in the U.S., but no such indicator exists. State tests at the high school level are typically administered at 10th and 11th grade. College admission tests, like the SAT and ACT, are administered before the 12th grade, generally to self-selected samples of students.

State tests and college admission tests do not provide a measure of what students know and can do at the very end of K-12 education. Even if these state tests and college admission tests were administered at the 12th grade, they could not be combined to produce nationally representative results.

NAEP is the only source of national and state-representative student achievement data at the 12th grade. As such, NAEP is uniquely positioned to serve as an indicator of 12th grade academic preparedness.

**Defining Academic Preparedness for College**

In the United States in 2013, there is no single, agreed upon definition of “academic preparedness for college” used by colleges for admission and placement. Postsecondary education in the U.S. is a complex mix of institutions, public and private, that have different admission requirements and different procedures and criteria for placing individual students into education programs.

In this complex mix are 2-year institutions, 4-year public and private institutions with a wide range of selectivity, and proprietary schools. Institutions range from highly selective (i.e., with admission criteria including very high grade point averages, successful completion of rigorous high school coursework and very high SAT and/or ACT scores) to open admission (i.e., all applicants are admitted).

Even within institutions, requirements may vary across majors or programs of study. For example, the mathematics and science high school coursework and academic achievement needed for acceptance into an engineering program in a postsecondary institution may be more rigorous than the general requirements for admission to the institution or for a degree in elementary education in the institution.

In order to design the NAEP 12th grade preparedness research, a working definition of preparedness was needed. The Governing Board’s Technical Panel on 12th Grade Preparedness Research recommended use of the following working definition, which defines academic preparedness for college as

… the academic knowledge and skill levels in reading and mathematics necessary to be qualified for placement…into a credit-bearing entry-level general education course that fulfills requirements toward a two-year transfer degree or four-year undergraduate degree
at a postsecondary institution [without the need for remedial coursework in those subjects]. (National Assessment Governing Board, 2009; p.3.)

This definition was intended to apply to the “typical” college, not to highly selective institutions, and thus, to the vast majority of prospective students, or about 80% of the college freshmen who enrolled in 2-year and 4-year institutions within 2 years following high school graduation (Ross, Kena, Rathbun, KewalRamani, Zhang, Kristapovich, and Manning, p 175). To make this clear, the definition is further elaborated as follows.

Academic preparedness for college refers to the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements (ECNRG) in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

This is consistent with the approach used by the College Board and ACT, Inc. in developing their respective college readiness benchmarks, which are used as external referents in the NAEP 12th grade preparedness research. The ACT benchmarks “represent predictive indicators of success for typical students at typical colleges (Allen and Sconing).” The SAT benchmarks are “an indication of college readiness at a typical college (College Board).”

Domain Definition for Academic Preparedness for College in Reading and Mathematics

The working definition described above set the stage for designing the preparedness research studies, but begged a basic question—What are the reading and mathematics knowledge and skills needed to qualify for placement into ECNRG and are they measured by NAEP? This question would be addressed by examining the degree of content match between NAEP and multiple widely accepted external sources that had developed domain definitions for academic preparedness for college in mathematics and reading.

A perfect match between two different sources could not be expected, but a sufficient content match between NAEP and each of a multiple of relevant widely accepted external sources would, collectively, support the inference that the needed knowledge and skills are measured by NAEP. Consequently, the Governing Board identified the following external sources for content comparison with NAEP: The American Diploma Project (ADP) benchmarks for mathematics and English, the ACT College Readiness Standards for Mathematics and Reading, and the ACT, SAT, and ACCUPLACER assessments for reading and mathematics. The results of the content comparison studies between NAEP and these other sources are described in the validity argument below.

The Central Issue: Validity

Having made the decision to determine the feasibility of NAEP reporting on 12th grade academic preparedness, the Governing Board recognized that the central concern would be establishing the validity of inferences about 12th grade academic preparedness that are to be made from NAEP scores and used in NAEP reports. The Governing Board would need to ensure that the content of NAEP 12th grade reading and mathematics assessments was appropriate for measuring academic preparedness and that research was conducted to collect evidence by which the validity of
proposed inferences could be evaluated. Finally, a formal validity argument would need to be developed, specifying the proposed inference(s) for NAEP reporting, the underlying assumptions or propositions, and the evidence related to the assumptions or propositions.

Accordingly, the Governing Board

- revised the NAEP assessment frameworks for the 2009 12th grade reading and mathematics with the explicit purpose of measuring academic preparedness for college and job training,
- appointed a special panel of technical experts to recommend a program of research on 12th grade academic preparedness (National Assessment Governing Board, 2009),
- approved and conducted a comprehensive set of preparedness research studies, and
- adopted the model for a validity argument described by Michael Kane (Kane).

The first phase of the Governing Board’s program of preparedness research is completed. The studies were conducted in connection with the 2009 NAEP 12th grade assessments in reading and mathematics. More than 30 studies of five distinct types have been conducted. Study results are available and the complete studies are posted at [http://www.nagb.org/what-we-do/preparedness-research.html](http://www.nagb.org/what-we-do/preparedness-research.html). The National Center for Education Statistics (NCES) has provide additional data drawn from analyses of the 2005 and 2009 High School Transcript Studies conducted in connection with the NAEP 12th grade assessments in those years.

From this research, Governing Board staff developed a proposed interpretation of NAEP performance in reading and mathematics related to 12th grade academic preparedness for college. Following below is the validity evidence for the proposed interpretation, presented in the form of a validity argument. The validity argument provides a statement of the proposed interpretation and the main assumptions inherent in the proposed interpretation in terms of academic preparedness for college. These assumptions are then evaluated using several lines of evidence, which were found to converge for both reading and for mathematics.

### Validity Argument

#### Overview

The National Assessment of Educational Progress (NAEP) program is designed to provide information about student achievement in reading, mathematics and other content areas at the 4th, 8th, and 12th grades. The items for the assessments are developed according to content frameworks and test specifications developed by the National Assessment Governing Board. Scientific sampling procedures are used to produce estimates of score distributions representative of the national population of students at each grade level, as well as estimates representative of public school students in individual states and in 21 urban school districts. The NAEP results do not produce scores for individual students, but rather, group estimates. The NAEP results are reported, based on the estimated score distributions, by average score, percentiles, and in terms of the percentages of students at or above three performance standards used for NAEP reporting, called achievement levels, that are designated Basic, Proficient, and Advanced.
The purpose of the research reported here was to examine whether the interpretation of 12th grade NAEP results in reading and mathematics could be extended to include statements about the percentage of U.S. 12th graders who are academically prepared for college and, if such an interpretation were found to be defensible, to determine the specific statements about academic preparedness that were supportable by the research evidence. The specific statements would be based on the following general definition for academic preparedness, used in relation to the NAEP preparedness research:

the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

The NAEP assessment program is well-established and regularly evaluated, with ample technical documentation of the interpretation of the results at all three grade levels. Therefore, the technical quality, accuracy, and representativeness of the NAEP results in terms of the estimated distributions of U.S. 12th graders on the NAEP scales in reading and mathematics will be taken as a given and as a starting point for additional inferences about the academic preparedness of U.S. 12th graders for college.

In particular, the intent of this validity argument is to examine the evidence in support of statements related to academic preparedness for college for use in reporting NAEP 12th grade results that would have the following general form:

The percentage of students in the NAEP distribution at or above a particular score level in reading or mathematics on 12th grade NAEP is a plausible, or reasonable, estimate of the percentage of 12th grade students who are academically prepared for college.

This interpretation would depend on four prior claims (or assumptions):

1. The 12th grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

2. Performance on 12th grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

The first claim is supported by the combination of the content of the NAEP assessment frameworks and the NAEP test items, the NAEP sampling designs, and the statistical models used to generate estimates of score distributions at each grade level and in each content area. These claims are well-established, documented, and evaluated; therefore, the attention of the validity argument will be directed primarily to the second, third, and fourth claims.

The second claim is supported by a statistical relationship study that examined student performance on the NAEP 12th grade reading and mathematics assessments to performance on the SAT reading and mathematics tests, as well as the respective college readiness benchmarks established by the College Board for these tests, which, in turn, are related to outcomes associated with academic preparedness for college.

The third claim was evaluated with multiple sources of evidence that were highly convergent. These include the SAT/NAEP statistical relationship study, a longitudinal study of Florida 12th grade students, and analyses of the 2005 and 2009 NAEP High School Transcript Studies.

The fourth claim is supported by the fact that the Governing Board reviewed the NAEP 12th grade reading and mathematics frameworks for the purpose of making NAEP a measure of academic preparedness for college; made changes to the frameworks accordingly; and conducted a comprehensive set of content alignment studies to determine the degree of match between NAEP and tests that are used for college admission and placement.

Further, the results from the examination of the NAEP content provide a counter argument to a possible falsifying claim about the positive relationships discussed in the second and third claims. The falsifying claim would be that the positive relationships between NAEP and the other indicators were merely statistical artifacts, due to factors extraneous to academic preparedness for college, akin to finding a high correlation between height and passing rates on a state driving test. The counter argument is that the relationships are meaningful because the NAEP 12th grade reading and mathematics assessments were intentionally designed to measure academic preparedness for college and that the evidence supports the conclusion that the NAEP 12th grade assessments do measure academic preparedness for college.
Proposed Inferences

For reading:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

the percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.

For mathematics:

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

In contrast to the inference for reading, which is set at the Proficient level, the inference for mathematics is set at a score on the NAEP mathematics scale of 163. This score is strongly supported by the consistent research results across years and data sources, but is below and significantly different from the cut-score for the Proficient level for NAEP 12th grade mathematics, which is 176.

The research results for mathematics do support a related inference—that students in the distribution at or above the NAEP Proficient level in mathematics are likely to be academically prepared for college. However, the percentage of such students would be substantially less than the percentage in the distribution at or above 163, and thus, would underestimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.

For these reasons, and to have the proposed inferences for reading and mathematics as parallel as possible, the proposed inference for reading is formulated in relation to the Proficient achievement level and the proposed inference for mathematics is formulated in relation to the NAEP mathematics scale score of 163.

Limitations on Interpretation and Other Caveats

False Negatives and False Positives

Some proportion of 12th grade students scoring below Proficient on the 12th grade NAEP Reading or below a score of 163 on the Mathematics Assessment are
likely to be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and

not likely to need remedial/developmental courses in reading or mathematics in college, but with a lower probability than those at or above Proficient in reading or 163 in mathematics.

In addition, some proportion of 12th grade students scoring at or above Proficient on the 12th grade NAEP Reading or 163 on the Mathematics Assessment may not

be academically prepared for ECNRG college courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement into degree-bearing programs designed to transfer to 4-year institutions, and

need remedial/developmental courses in reading or mathematics in college.

Not a Preparedness Standard
The proposed inferences are not intended to represent or be used as standards for minimal academic preparedness for college. The proposed inferences are intended solely to add meaning to interpretations of the 12th grade NAEP reading and mathematics results in NAEP reports.

Academically Prepared for College
The proposed inferences are intended to apply to the typical degree-seeking entry-level college student at the typical college. Thus, “academically prepared for college” refers to the reading and mathematics knowledge and skills needed for placement into ECNRG courses in broad access 4-year institutions and, for 2-year institutions, the general policies for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

It is important to note the focus on “placement” rather than “admission.” This distinction is made because students who need remedial courses in reading, mathematics or writing may be admitted to college, but not placed into regular, credit-bearing courses. The criterion of importance is qualifying for regular credit-bearing courses, not admission.

The proposed inferences are not intended to reflect academic requirements for highly selective postsecondary institutions; to the additional academic requirements for specific majors or pre-professional programs, such as mathematics, engineering, or medicine; or to academic requirements applicable to entry into certificate or diploma programs for job training or professional development in postsecondary institutions.

The proposed inferences are focused on the first year of college; they do not support conclusions about college persistence beyond the first year or completion of a degree. The inferences will necessarily apply in general across a broad range of programs and majors, but should not be applied specifically to any particular program or major.
GPA of B- or Better
The selection of “first-year GPA of B- or better” as a referent was made because of its use as a research-based criterion in defining college readiness benchmarks developed by an acknowledged leader in college testing programs—the College Board. The College Board had agreed to partner with the Governing Board in a study linking performance on 12th grade NAEP with the SAT. Another leader in college testing programs, ACT, Inc. has developed similar benchmarks for its college admission assessments using a similar criterion and similar methodology. Because they are based on credible research related to college outcomes, and because performance on the respective tests could be linked to performance on NAEP, the college readiness benchmarks used by these testing programs were embraced as relevant, useful points of reference for the NAEP preparedness research.

The College Board has set a score of 500 on the SAT Mathematics and Critical Reading tests as its college readiness benchmarks in those areas. Based on its research, the College Board has determined that the score of 500 predicts, with a probability of .65, attainment of a first-year overall GPA of B- or higher. Similarly, the ACT college readiness benchmarks are based on research indicating a .50 probability of attaining first-year grades in relevant courses (e.g., college algebra and courses requiring college level reading) of B or better and .75 probability of C or better.

The proposed inferences are not intended to convey that a B- or any particular grade should be deemed a standard or goal for postsecondary student outcomes. This criterion was selected to foster comparability across the preparedness research studies, where applicable. However, it does seem self-evident that achieving a first-year GPA of B- or better, without enrollment in remedial/developmental courses, lends support to the likelihood of having possessed academic preparedness for first-year college courses upon entry to college.

Data Limitations
Although the preparedness research studies are comprehensive and the results consistent and mutually confirming, for reading they are limited to one year for data at the national level and to one state-based longitudinal study. For mathematics, there are two separate years of data at the national level and one state-based longitudinal study. Therefore, more evidence exists to support the plausibility of inferences related to mathematics than to reading.

Preparedness for Job Training
The completed research with respect to academic preparedness for job training does not support conclusions relative to the NAEP scale and will not be addressed at this time.
Discussion of the Claims and Evidence

1. The 12th-grade NAEP results in reading and mathematics provide unbiased, accurate estimates of the percentages of students at or above specified score levels on the NAEP scales in reading and mathematics for 12th-grade students in the United States.

The proposed inferences are premised in part on the capability of NAEP to report percentages of students scoring at or above a certain score on the NAEP 12th grade reading and mathematics scales. The technical qualities of the NAEP scales make them well suited to this purpose.

The NAEP sampling, scaling, IRT modeling, and statistical procedures are widely accepted, well documented (for example, see National Center for Education Statistics, pp. 70-71) and have been periodically evaluated over two decades (for example, see complete list of research conducted by the NAEP Validity Studies Panel at http://www.air.org/reports-products/index.cfm?fa=viewContent&content_id=890 and “Evaluation of the National Assessment of Educational Progress: Study Reports” at http://www2.ed.gov/rschstat/eval/other/naep/naep-complete.pdf).

Other than issues relating to the comparability among the state-level NAEP samples of inclusion rates of students with disabilities and students who are English language learners (about which the Governing Board and NAEP have taken and continue to take significant action), there is little dispute about the appropriateness of the NAEP sampling, scaling and statistical procedures for estimating the percentage of students scoring at or above a selected NAEP scale score.

This is relevant because the proposed inferences that are the subject of this validity argument are interpretations to add meaning to the reporting of NAEP 12th grade reading and mathematics results at particular score levels. The percentages of students at or above particular score levels (e.g., the NAEP achievement levels) have been estimated with accuracy and reported regularly, beginning with assessments in 1992. The proposed inference for reading would use the cut-score for 12th grade Proficient as the basis for reporting. The proposed inference for mathematics would use the score of 163 on the NAEP 12th grade scale as the basis for reporting, which is between the Basic and Proficient achievement levels. Clearly, reporting NAEP results using the proposed inferences will not impair the accuracy of the estimates of the percentages of students scoring at or above the identified points on the NAEP score scales.

2. Performance on 12th-grade NAEP assessments in mathematics and reading is positively related to other measures associated with outcomes reflecting academic preparedness for college.

In designing the NAEP preparedness research program, the Governing Board determined that it would be essential to examine how performance on NAEP relates to performance on other
measures and outcomes associated with academic preparedness for college. The research program studied the relationship between performance on NAEP and performance on the SAT and ACT college admission tests, including the respective college readiness benchmarks that had been established by these testing programs.

The data sources for the analyses that were conducted are: the NAEP/SAT linking studies (see report at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/SAT-NAEP_Linking_Study.pdf); the Florida longitudinal study (see report at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/statistical-relationships/Florida_Statistical_Study.pdf); the 2005 and 2009 NAEP High School Transcript Studies; and the Governing Board’s survey of postsecondary education institutions’ use of tests and the cut-scores on those tests for determining whether incoming students need remedial instruction in reading and mathematics (Fields and Parsad).

In addition, the research program examined directly the relationship between performance on NAEP and postsecondary outcomes analyzing data from the Florida longitudinal study.

The results of these studies will be discussed both in this section and the next section of the validity argument. In this section, background is provided on the indicators that were examined and the results of the NAEP/SAT linking study. The NAEP/SAT linking study is discussed in this section because, as the most recent large-scale national study, it serves as a focal point for discussing the results of the other studies. Thus, in section 3, the results of the other statistical linking studies are discussed in relation to the NAEP/SAT linking study.

**Indicators: College Board and ACT College Readiness Benchmarks**

The College Board and ACT, Inc. have established college readiness benchmarks for the SAT and the ACT in a number of subjects tested, including reading and mathematics. The SAT College Readiness Benchmark for critical reading and mathematics is a score of 500 on the respective tests. According to the College Board’s research, a score of 500 predicts, with a .65 probability, a first-year GPA of B- or better. The ACT College Readiness Benchmark for reading is a score of 21. According to ACT’s research, a score of 21 predicts, with a .50 probability, a grade of B or better (or .75 probability of a C or better) in first-year courses requiring college reading, such as history and the social sciences. A score of 22 on the ACT mathematics tests predicts a .50 probability of a grade of B or better in a first-year mathematics course, or a .75 probability of a grade of C or better. The College Board research and the ACT research are based on the first-year outcomes of their respective test takers.

**Indicators: First Year GPA of B- or Better and Remedial/non-Remedial Placement**

The Governing Board has a partnership with the state of Florida as a part of the Board’s program of preparedness research. Florida was one of 11 states that volunteered to provide state-
representative samples of 12th grade students for the 2009 NAEP reading and mathematics assessments. Under the partnership, the Florida 12th grade sample is being followed through the postsecondary years via the highly developed Florida longitudinal education data system. For comparability with the SAT College Readiness Benchmarks, the Governing Board analyzed the Florida data to determine the average score and interquartile range for the NAEP test takers with a first year GPA of B- or better. In addition, the Governing Board analyzed the Florida data to determine the average score and interquartile range for the NAEP test takers who were and who were not placed into remedial reading or remedial mathematics in their first year of college.

Analysis of Results for Mathematics

The statistical linking study examining performance on the NAEP 12th grade mathematics assessment and performance on the SAT mathematics test yielded a correlation of .91. This high correlation clearly supports inferences about NAEP performance in relation to SAT performance. The study also examined how performance on NAEP relates to the SAT College Readiness Benchmark for mathematics (i.e., a score on the SAT mathematics test of 500). The SAT benchmark provides “an indication of college readiness at a typical college (College Board).” This is consistent with the Governing Board’s definition of academic preparedness cited previously:

> Academic preparedness for college refers to the reading and mathematics knowledge and skills needed to qualify for placement into entry-level, credit-bearing, non-remedial courses that meet general education degree requirements in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions.

The SAT College Readiness Benchmark for mathematics is relevant to student outcomes in college, for it is “the SAT score associated with a 65 percent probability of earning a first-year GPA of B- (i.e., 2.67) or higher (College Board).” The average NAEP score of students scoring at the College Readiness Benchmark for mathematics was 163 (see Figure 1). As will be demonstrated in the discussion of the third claim, there are additional data corroborating this level of performance on the 12th grade NAEP mathematics assessment to outcomes in college.

Analysis of Results for Reading

The statistical linking study examining performance on the NAEP 12th grade reading assessment and the SAT critical reading test resulted in a correlation of .74. Although it may not be high enough to predict the performance of individual students from one test to another (which is not required to support the proposed inference for reading), it is sufficient to support the group-level inferences reported by NAEP.
Performance on NAEP was also examined in relation to the SAT College Readiness Benchmark for critical reading (i.e., a score on the SAT critical reading test of 500). The SAT benchmark provides “an indication of college readiness at a typical college (College Board).” This is consistent with the Governing Board’s definition of academic preparedness discussed in the results for mathematics above.

The SAT College Readiness Benchmark for critical reading is relevant to student outcomes in college, for it is “the SAT score associated with a 65 percent probability of earning a first-year GPA of B- (i.e., 2.67) or higher (College Board).” The average NAEP score of students scoring at the College Readiness Benchmark for reading was 301(see Figure 2). As will be demonstrated in the discussion of the third claim, there are additional data corroborating this level of performance on the 12th grade NAEP reading assessment to outcomes in college.

3. There is a point on the NAEP scale that corresponds to other measures, indicators, and outcomes associated with academic preparedness for college (i.e., possession of a specific level of academic proficiency, attainment of a first-year overall college GPA of B- or better, and placement into entry-level, credit bearing non-remedial college courses).

In addition to the NAEP/SAT Linking Studies (NSLS) described above, analyses were conducted using data from several other studies. There was a high degree of convergence found across the studies. The results are described below, first for mathematics and then for reading.

**Analysis of Results for Mathematics**

Companion statistical relationship studies to the NSLS for mathematics examined data from the 2005 and 2009 national NAEP High School Transcript Studies (HSTS) and from a longitudinal study under a partnership with the Florida Department of Education (FLS). In 2009, Florida was one of eleven states that volunteered to participate in 12th grade state NAEP in reading and mathematics. Using the highly developed Florida longitudinal data base, the students in the 12th grade NAEP samples were followed into postsecondary public institutions.

Analyzing data from the transcripts of NAEP test takers, the HSTS examined performance on 12th grade NAEP mathematics in relation to performance in mathematics on the SAT and ACT college admissions tests in 2005 and 2009. The FLS study examined performance on the NAEP 12th grade mathematics assessment in relation to the SAT and ACT college readiness benchmarks, first year overall college GPA, and whether students were placed into non-remedial college courses. The study results are displayed in Figure 1.

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) because it is the most recent of the national studies. The average NAEP score is 163 for students with an SAT score at the College Readiness Benchmark for mathematics of 500.
The other study results are consistently convergent with the NSLS results. The average NAEP mathematics scores for 12th grade students scoring at the SAT College Readiness Benchmark of 500 for mathematics are compared first for the 2005 HSTS and the 2009 NSLS. The average scores are 161 and 163 respectively.

These results are confirmed by the FLS. The average NAEP mathematics score for the 12th grade Florida NAEP test takers who scored at the SAT College Readiness Benchmark of 500 was 160, much like the 2009 NSLS results and the 2005 HSTS results.

As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for mathematics is defined somewhat differently than the SAT College Readiness Benchmark for mathematics. However, it is noteworthy that even with this different definition, the results from the 2005 HSTS, 2009 HSTS, and 2009 FLS analyses for the ACT (169, 166, and 164, respectively) are consistent and very similar to the results for the 2009 NSLS.

To answer the question, "What is the relationship between performance on NAEP and actual student outcomes?", we look to the FLS results. First we examine the average NAEP mathematics score for the 12th grade Florida NAEP test takers who attained a first-year GPA of B- or better. The average NAEP score for these students was 162. This is consistent with the SAT College Readiness Benchmark analyses and further supports the inference that students at
or above 163 on the 12th grade NAEP mathematics scale are likely to be academically prepared and attain a first-year GPA of B- or better. It follows, of course, that students who are academically prepared will not require remedial courses.

Thus, another outcome of interest is placement of entry-level students into remedial college courses versus non-remedial credit-bearing courses. Here again, we look to the FLS as a data source. The average NAEP mathematics score was 165 for the Florida NAEP test-takers not placed into remedial courses, which is consistent with the NSLS score of 163 on the NAEP 12th grade mathematics scale. Furthermore, the average NAEP score of students who were placed into remedial mathematics courses in college was 136, much lower and significantly different from the NSLS score of 163.

The FLS results, together with the SAT and ACT analyses, lend support to the conclusions that students scoring at or above 163 on the 12th grade mathematics scale are likely to be academically prepared for ECRNG college courses and not likely to need remedial courses in mathematics.

These convergent, consistent results across years and across studies support the proposed inference that the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

**Analysis of Results for Reading**

The companion statistical relationship study to the NSLS for reading examined data from a longitudinal study under a partnership with the Florida Department of Education (FLS). In 2009, Florida was one of eleven states that volunteered to participate in 12th grade state NAEP in reading and mathematics. Using the highly developed Florida longitudinal data base, the students in the 12th grade NAEP samples were followed into postsecondary public institutions.

The FLS study examined performance on the NAEP 12th grade reading assessment in relation to the SAT and ACT college readiness benchmarks for reading, first year overall college GPA, and whether students were placed into non-remedial college courses. The study results are displayed in Figure 2.

The focal point for the discussion of these results is the 2009 NAEP/SAT Linking Study (NSLS) for reading, because it is the most recent of the national studies. The average NAEP score is 301 for students with an SAT score at the College Readiness Benchmark for critical reading of 500. A NAEP score of 301 in 12th grade reading is not significantly different from the cut-score for the 12th grade Proficient achievement level (302).
The FLS results are consistently convergent with the NSLS results. The average NAEP reading score was 299 for the 12th grade Florida NAEP test takers who were not placed into remedial courses in their first year. The average score was 298 for those who had a first year overall GPA of a B- or better. These data, which show the relationship between performance on NAEP and actual student outcomes, provide strong confirmation that students scoring at or above Proficient on the NAEP 12th grade reading assessment are likely to be academically prepared for ECNRG college courses.

As discussed elsewhere in this validity argument, the ACT College Readiness Benchmark for reading is defined differently than the SAT College Readiness Benchmark for reading. However, it is noteworthy that even with this different definition, the ACT results from the 2009 FLS analysis are similar to the NSLS analysis and the FLS outcome data.

Taken together, these results support the inference that students scoring at or above Proficient on the NAEP 12th grade reading scale are likely to be academically prepared for ECNRG college courses.

In conclusion, these results suggest that the percentage of students at or above the Proficient level in reading on 12th grade NAEP would provide a plausible (or reasonable) estimate of the percentage of 12th grade students in the U.S. who are academically prepared for college.
4. The positive relationship between NAEP and the other indicators and outcomes is meaningful in terms of academic preparedness for college, not merely a statistical artifact, because the 12th grade reading and mathematics domains measured by NAEP were specifically designed to measure academic preparedness for college.

- NAEP Assessment Frameworks Were Revised to Measure Academic Preparedness

The National Assessment Governing Board intentionally revised the NAEP 12th grade reading and mathematics assessment frameworks with the purpose of measuring academic preparedness for college.

On March 5, 2004, the Governing Board accepted the report of the Commission on NAEP 12th Grade Assessment and Reporting. The Commission recommended that “NAEP should report 12th grade students’ [academic preparedness] for college-credit coursework, training for employment, and entrance into the military.”

For NAEP to report on 12th grade academic preparedness for college, it must measure relevant content at the 12th grade. The content of each assessment is determined by the NAEP assessment frameworks, which the Governing Board is responsible for developing and approving. Accordingly, the Governing Board decided that the extant NAEP frameworks intended for the 2009 for reading and mathematics at the 12th grade would be reviewed. The review would identify changes needed to measure 12th grade academic preparedness for college. Examples of the changes made are described in the next two subsections.

Assessments at the 12th grade in reading and mathematics are conducted at least once every 4 years. In 2004, when the Board decided to proceed with the 12th grade academic preparedness initiative, 2009 was the next assessment year in which the 12th grade reading and mathematics assessments could be affected by framework changes.

In September 2004, the Governing Board contracted with Achieve, Inc. (Achieve) to review the NAEP 12th grade reading and mathematics assessment frameworks and identify where changes, if any, would be needed. Achieve had established the American Diploma Project (ADP) “…to improve postsecondary preparation by aligning high school standards, graduation requirements and assessment and accountability systems with the demands of college and careers (see www.achieve.org/adp-network).” The ADP had conducted research to identify key competencies in English and mathematics needed for high school graduates who aspire to higher education. They refer to these as the “ADP benchmarks.” The type of colleges that were the target for the ADP research was similar to the “typical colleges” in the Governing Board’s research. These were the “two- and four-year colleges and universities in each of the ADP partner states...[that] enroll the vast majority of high school graduates going on to college:

---

2 The review also addressed academic preparedness for job training, but that part of the NAEP preparedness initiative is not being addressed in this validity argument.
community colleges, as well as four-year state institutions, but generally not the more highly selective “flagship” campuses.” (Achieve, 2004, p. 107)

The research and expertise of the American Diploma Project was widely accepted and was brought to bear in reviewing the NAEP frameworks for 12th grade reading and mathematics. Achieve convened a panel of nationally recognized experts in reading and a panel of nationally recognized experts in mathematics. The panels were comprised of individuals from the K-12, postsecondary, research, and policy spheres, knowledgeable about academic preparedness for college reading and college mathematics. The panels compared the 12th grade NAEP reading and mathematics frameworks and the ADP benchmarks.

**Reading**

The Achieve reading panel found considerable similarity between NAEP and the ADP benchmarks for English, although not perfect agreement. This is displayed in the side-by-side chart on pages 30-40 of the Achieve Reading Report (http://www.nagb.org/content/nagb/assets/documents/commission/researchandresources/Achieve%20Reading%20Report.pdf). The English benchmarks have eight major components and objectives under each component. Three of these major components were deemed “Not Applicable” to the reading domain: writing, research, and media.

For almost all of the applicable objectives under the five major components that were applicable to the reading domain, the Achieve reading panel found matches in the NAEP 2009 reading framework. Overall, the panel concluded that “…the 2009 NAEP Reading Framework…was aligned to the ambitious [ADP] benchmarks” (Achieve Reading Report, p. 2).

The reading panel also listed items in the NAEP framework that are not found in the ADP English benchmarks. For example, under Argumentation and Persuasive Text, figurative language and rhetorical structure, including parallel structure and repetition, was present in the NAEP reading framework at grade 12, but not in the ADP benchmarks. Under Poetry, tone, complex symbolism, and extended metaphor and analogy were present in the NAEP reading framework but not the ADP benchmarks. A complete listing of the items in the NAEP framework not present in the ADP benchmarks appears on page 41 of the Achieve Reading Report.

Although the Achieve reading panel concluded that the 12th grade NAEP reading framework for 2009 was aligned with the ADP benchmarks applicable to reading, the panel’s report does include six recommendations. The Governing Board approved these recommendations on February 14, 2005. For example, the Achieve reading panel recommended increasing the percentage of informational text passages from 60% to 70% and to feature additional items that ask students to compare texts. The changes were modest, sufficiently so to permit continuation of the 12th grade trend line from its initiation in 1992.

The NAEP reading framework used for the 2009, 2011, and 2013 assessments contains the following statement
In May 2005, the Governing Board adopted a policy statement regarding NAEP and 12th-grade preparedness. The policy states that NAEP will pursue assessment and reporting on 12th-grade student achievement as it relates to preparedness for post-secondary education and training. This policy resulted from recommendations of the Board’s National Commission on NAEP 12th Grade Assessment and Reporting in March 2004. Subsequent studies and deliberations by the Board took place during 2004 and 2005.

In reading, the Board adopted minor modifications to the 2009 NAEP Reading Framework at grade 12 based on a comprehensive analysis of the framework conducted by Achieve, Inc. The current version of the reading framework incorporates these modifications at grade 12 to enable NAEP to measure and report on preparedness for postsecondary endeavors (National Assessment Governing Board, 2008, *Reading Framework*, p. v).

**Mathematics**

The mathematics review began with the 2007 NAEP mathematics framework, which was the most current and included the changes approved for the 2005 12th grade mathematics assessment. The Achieve panel examined the NAEP mathematics framework at the 12th grade in relation to the ADP benchmarks for mathematics. The Achieve panel developed proposed revisions to the assessment objectives for grade 12. While acknowledging differences in language and purpose, the Achieve mathematics panel concluded that the “overall mathematics frameworks of ADP and [12th grade] NAEP are remarkably similar” (see http://www.nagb.org/content/nagb/assets/documents/commission/researchandresources/Achieve-Mathematics-Report.pdf, Achieve Mathematics Report, p.9).

The Governing Board convened a panel of mathematicians and mathematics educators to review and revise the objectives in relation to the objectives for grades 4 and 8. The panel conducted focus groups with various NAEP constituents, using repeated rounds of reviews. The Governing Board approved the final set of grade 12 objectives on August 5, 2006. The changes to the framework were sufficiently modest to permit the continuation of the 12th grade trend line begun with the 2005 12th grade mathematics assessment under the previous 12th grade framework. Like the reading framework, the 2009/2013 mathematics framework for grade 12 states the Board’s intention to measure 12th grade academic preparedness (National Assessment Governing Board, 2008, *Mathematics Framework*, pp. 2-3).

**Conclusion**

The Governing Board, by official action, revised the NAEP 12th grade reading and mathematics frameworks with the explicit purpose of measuring 12th grade academic preparedness for college, beginning with the 2009 assessments. Setting forth the measurement purpose and making relevant revisions to the NAEP assessment frameworks are necessary elements of the validity argument; however, they are not sufficient. Evidence must be considered with respect to the alignment of the framework and the test questions administered to the measurement purpose. This will be addressed in the next section.
Examples of Objectives added to the 2009 Grade 12 Mathematics Framework

**Number properties and operations**
b) * Analyze or interpret a proof by mathematical induction of a simple numerical relationship.

**Measurement**
d) Interpret and use the identity $\sin^2 \theta + \cos^2 \theta = 1$ for angles $\theta$ between $0^\circ$ and $90^\circ$; recognize this identity as a special representation of the Pythagorean theorem.

e) * Determine the radian measure of an angle and explain how radian measurement is related to a circle of radius 1.

f) * Use trigonometric formulas such as addition and double angle formulas.

g) * Use the law of cosines and the law of sines to find unknown sides and angles of a triangle.

**Geometry**
e) * Use vectors to represent velocity and direction; multiply a vector by a scalar and add vectors both algebraically and graphically.

g) * Graph ellipses and hyperbolas whose axes are parallel to the coordinate axes and demonstrate understanding of the relationship between their standard algebraic form and their graphical characteristics.

h) * Represent situations and solve problems involving polar coordinates.

**Data Analysis, Statistics, and Probability**
c) * Draw inferences from samples, such as estimates of proportions in a population, estimates of population means, or decisions about differences in means for two “treatments.”

e) * Recognize the differences in design and in conclusions between randomized experiments and observational studies.

k) * Use the binomial theorem to solve problems.

e) * Recognize and explain the potential errors caused by extrapolating from data.

**Algebra**
e) Identify or analyze distinguishing properties of linear, quadratic, rational, exponential, or trigonometric functions from tables, graphs, or equations.

j) * Given a function, determine its inverse if it exists and explain the contextual meaning of the inverse for a given situation.

h) * Analyze properties of exponential, logarithmic, and rational functions.

g) * Determine the sum of finite and infinite arithmetic and geometric series.
Content Alignment Studies Found Significant Overlap between NAEP and the ACT, SAT and ACCUPLACER

The Governing Board conducted studies to determine the degree of content similarity between NAEP 12th grade reading and mathematics assessments and relevant tests used for college admissions and placement.

The studies had two objectives. The first objective was to determine the degree to which the content of 12th grade NAEP in reading and mathematics covers the reading and mathematics knowledge and skills needed for first year college work. The SAT, ACT, and ACCUPLACER are well-established tests that assess individual students’ reading and mathematics proficiency in relation to college level expectations.

The ACT is developed with the purpose of “…[measuring] as directly as possible the degree to which each student has developed the academic skills and knowledge that are important for success in college…” (ACT Technical Manual, p. 62).

The SAT is developed “to ensure that the topics measured on the SAT…reflect what is being taught in the nation’s high schools and what college professors consider to be required for college success.” (Kim, Wiley, and Packman, p.1)

The ACCUPLACER has the purpose of “…[determining] which course placements are appropriate for [incoming college] students and whether or not remedial work is needed.” (ACCUPLACER, p. A-2)

The SAT, ACT and ACCUPLACER in reading and mathematics are widely used for these purposes by admissions and placement professionals in postsecondary education institutions. These testing programs regularly conduct curriculum surveys, validity studies and other research to support their claims that the content measured is directly related to the reading and mathematics knowledge and skills needed to qualify for entry-level credit-bearing courses (e.g., see the ACT curriculum studies for 2012, 2009, 2005, and 2002 at http://www.act.org/research-policy/national-curriculum-survey/, and the College Board National Curriculum Survey on English and Mathematics at http://research.collegeboard.org/publications/content/2012/05/national-curriculum-survey-english-and-mathematics).

Therefore, with the assumption that the SAT, ACT, and ACCUPLACER do measure the content needed for college level work, significant content overlap between NAEP and these other assessments would support the conclusion that what NAEP measures covers the knowledge and skills needed by college freshmen to be placed into entry-level credit bearing courses.

The second reason for conducting the content alignment studies was to provide information for interpreting the results of planned statistical linking studies between NAEP and the other tests, which measure academic preparedness for college. The linking studies were designed to examine how performance on NAEP compares with performance on the other tests, with the
purpose of supporting inferences about academic preparedness for college. For NAEP to support inferences about academic preparedness for college based on the linking studies, a sufficient content match would be needed between NAEP and the other tests, not just a statistical relationship.

The Content Alignment Studies: Overview
The Governing Board conducted content alignment studies in reading and mathematics comparing the 2009 12th grade NAEP and the ACT, SAT, and ACCUPLACER reading and mathematics tests. Overall, considerable overlap was found between the ACT and NAEP and the SAT and NAEP, with some differences. NAEP was found to measure much of what is measured on the ACCUPLACER, but the reading and mathematics domains measured by NAEP were much broader than ACCUPLACER. More details are provided in the summaries of the individual studies below.

The general design for the content alignment studies was to compare the 12th grade NAEP frameworks in reading and mathematics with the analogous document for the other test, and then to compare the test items from one test to the framework/analogous document of the other test. The reviews were performed by subject specific (i.e., mathematics, reading) panels, composed of experts in mathematics or reading and English instruction at the high school and college levels.

Alignment studies that compare an assessment to the content standards on which it is based are relatively common and have well-established methodologies. However, this is not true for the types of alignment studies the Governing Board planned to conduct: content alignment studies comparing different assessment programs. Different assessment programs have different purposes, different approaches to describing the domain being measured, and, possibly, different “grain size” in the level of detail in describing the domain.

The Governing Board contracted with Norman Webb, a noted expert in content alignment studies, to prepare a design document for conducting the assessment to assessment alignment studies. The purpose was to put in place a methodology that considered the special challenges of assessment alignment studies and to foster comparability in the conduct of the studies and the reporting metrics across studies and contractors. The link to the Webb design document is at (http://www.nagb.org/content/nagb/assets/documents/publications/design-document-final.pdf).

The Webb design was developed after the ACT alignment studies were completed. It was used in conducting the SAT and ACCUPLACER content alignment studies.

In the following sections are summaries of the content alignment study results, excerpted from the study reports. The results for the three content alignment studies in reading are presented first, followed by the three content alignment studies for mathematics, along with summary discussions for the reading and mathematics results.
The Content Alignment Studies: Reading Results

Reading: ACT
The Governing Board contracted with ACT, Inc. to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the ACT reading test. The full report can be found at http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf.

The reading panel was composed of 7 members, with expertise in reading and/or English instruction at the high school and college levels. The panel was about evenly divided in terms of prior familiarity with either the ACT or NAEP reading domains.

The panel found considerable similarity in the content of the NAEP 12th grade reading assessment and the ACT. For example, the NAEP 12th grade reading framework was compared to the ACT reading domain and the ACT College Readiness Standards for reading. The ACT College Readiness Standards (CRS) are descriptions of the content (i.e., the knowledge and skills) measured by the ACT reading test in score bands along the ACT 1-36 point scale from 13-36 (see http://www.act.org/standard/planact/reading/). The panel concluded that

“All of the skills highlighted in the ACT [reading] domain and in the [ACT] College Readiness Standards [for reading] were identified within the NAEP Reading framework. In performing the comparison in the other direction—NAEP to ACT—it was the sense of the panel that the ACT measured primarily those skills that NAEP identifies as Locate/Recall and Integrate/Interpret skills, those that pertain primarily to finding explicit information in text (what the ACT would call Referring skills) and to making inferences, drawing conclusions, and making generalizations from information within text (what the ACT would call Reasoning skills). The panel saw less evidence of the higher-level analytical and evaluative Critique/Evaluate skills in the ACT domain, and attributed that to the multiple-choice format of the ACT [whereas NAEP includes constructed response items as well as multiple choice]. Another difference is that NAEP includes items and texts measuring how well an examinee can apply reading skills across texts, whereas the paired passage format is not a feature of the ACT. So, while the NAEP Reading framework and the ACT Reading domain, test specifications, and College Readiness Standards share similarities, important differences in what and how the assessments measure suggest caution when drawing comparisons between the assessments.” (p.17)

The reading panel also conducted an item classification study, in which the NAEP 12th grade reading items were classified in relation to the ACT College Readiness Standards for Reading.

“A total of 152 Reading items (comprising 17 blocks) were classified in [the reading] study. Of these, 97 were multiple-choice (MC). Nine were dichotomously-scored (“incorrect” or “correct”) short constructed-response (DSCR) items. Thirty-three were polytomously-scored short constructed-response (PSCR) items, each scored using a three-point scoring rubric. Thirteen were extended constructed-response (ECR) items,
each scored using a four-point rubric. Each DSCR had one creditable score category, each PSCR had two, and each ECR had three. Each Reading panelist, therefore, assigned a total of 211 classifications to the NAEP Reading items [and rubric scoring categories].” (p.54)

An item or score category was deemed “classified” if there was majority agreement; that is, if at least 4 of the 7 panel members agreed about the score band to which an item (or creditable score category under an item rubric) was assigned.

Of the 211 determinations to be made, there was only one for which there was no majority agreement (the assignment of a PSCR rubric to a CRS score band). Of the remaining 210 determinations, 181 were unanimous.

The reading panel was able to classify 137 items or rubric categories (about two-thirds of the determinations to be made) to the CRS score bands. Of the 97 multiple choice items, 81 (or 84%) were classified. Of the 113 rubric score categories for items, 56 (or 50%) were classified. The reasons some multiple choice items and rubric score categories could not be classified were related to the differences in the ACT and NAEP reading domains described above. These reasons include the presence of constructed response items in NAEP but not the ACT, the presence of items involving multiple texts in NAEP but not the ACT, and the greater presence of “Critique/Evaluate” type items in NAEP than the ACT.

Of the 137 classifications, 24 were in the score bands from 13-19; 113 of the classifications were in the score bands from 20-36. This is noted because the ACT College Readiness Benchmark for reading is 21. The ACT College Readiness Benchmark signifies the score at which a student has a 50% chance of attaining a grade of B or better in a relevant subject and a 75% chance of a C or better. In addition, the Governing Board conducted a survey of postsecondary institutions’ use of tests in making entry-level decisions about placement into remedial or regular credit-bearing courses. With respect to the ACT, 18 was the mean reading score below which students were deemed to need remedial course work (Fields and Parsad, P. 19). Whereas this provides a context for the study results, it must be kept in mind that in making their judgments about item classifications, the panelists did not have data about NAEP item difficulty or data on how performance on NAEP compares with performance on the ACT.

Finally, although the study results support the conclusion that the 12th grade NAEP reading assessment measures content directly related to academic preparedness for college, it is noted that the study was conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in reading.

Reading: SAT
The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the SAT critical reading test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found
Overall, the study found similar content in the NAEP 12th grade reading assessment and the SAT critical reading test. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

What is the correspondence between the reading content domain assessed by NAEP and that assessed by SAT?

The greatest commonality between the two tests is their shared emphasis on the broad skills of integrating and interpreting both informational and literary texts. This is evident in the majority of items from both tests aligned to NAEP Standard 2, Integrate/Interpret,” including many to Goal 2.1, “Make complex inferences within and across both literary and informational texts.”

Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), there is also considerable overlap at the level of more specific skills.

To what extent is the emphasis of reading content on NAEP proportionally equal to that on SAT?

Both tests had many of their item alignments to the same NAEP “Integrate/Interpret” objectives, often with similar percentages of alignments. Although there were some differences in emphasis, both tests also had notable percentages of alignments to SAT Objectives B.1.1–B.1.3 and B.1.5. Skills with overlap include inferring/analyzing the following:

- the “main idea” and “author’s purpose” (SAT Objective B.1.1 and NAEP Objectives 2.3.a and 2.1.f);
- the “tone and attitude” of an author or character (NAEP Objectives 2.2.a and 2.2.c and SAT Objective B.1.4);
- the use of “rhetorical strategies” (NAEP Objective 2.1.d and SAT Objective B.1.2); and
- connections between ideas, perspectives, or problems (NAEP Objective 2.1.b and SAT Objectives B.1.3 and B.1.5).

Additionally, in the area of greatest content overlap—items on both tests aligned to objectives for NAEP “Integrate/Interpret” and aligned to SAT “Passage-Based Reading” Objectives B.1.1–B.1.5—both tests met the typical threshold criteria for depth of knowledge consistency…

Despite these similarities, there are some notable differences in emphasis between the two assessments. Both tests assess vocabulary skills. However, NAEP addresses vocabulary exclusively in the context of passage comprehension, while the majority of SAT vocabulary items are in a sentence-completion format, in which context plays a more limited role. This difference reflects NAEP’s emphasis on the understanding of
word meaning in context; the assessment is not intended to measure students’ prior knowledge of word definitions. The SAT sentence-completion items provide some context within the single sentence text, but in many cases, students’ success on the items almost certainly depends on their prior knowledge of word definitions.

In addition, panelists found considerably less emphasis in SAT than in NAEP on literal comprehension and critical evaluation, particularly the evaluation of the quality or effectiveness of an author’s writing, skills covered in the NAEP standards “Locate/Recall” (locating/recalling specific details and features of texts) and “Critique/Evaluate” (evaluating texts from a critical perspective), respectively. This difference suggests a greater emphasis on these skills in NAEP.

Even with the minimal coverage of NAEP “Locate/Recall” and “Critique/Evaluate” standards by SAT items, all NAEP items found a match in the SAT framework. However, the broad language of the SAT framework can encompass the range of the NAEP items. For example, SAT Goal B.2, “Literal Comprehension,” refers to items that “ask what is being said” in a “small but significant portion of a reading passage,” a description that can easily accommodate most NAEP “Locate/Recall” items and objectives. In fact, nearly all items on the NAEP short version that were coded to “Locate/Recall” objectives in the NAEP framework were matched to SAT Goal B.2 in the SAT framework.

Similarly, SAT Objective B.1.3, to which approximately one-quarter of NAEP items aligned, includes “Evaluation,” the primary focus of NAEP “Critique/Evaluate.” The description in SAT Objective B.1.3 of items that “ask the test taker to evaluate ideas or assumptions in a passage” is compatible at a very general level with NAEP “Critique/Evaluate” objectives addressing the author’s point of view, logic, or use of evidence. SAT Objective B.1.2, “Rhetorical Strategies,” is also broad enough in its language to make it a reasonable match for some NAEP “Critique/Evaluate” items focused on “author’s craft” or use of “literary devices.” In the NAEP short version, all items that aligned to “Critique/Evaluate” objectives in the NAEP framework were aligned to either SAT Objectives B.1.2 or B.1.3, or both.

Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire reading subdomains are missing or not aligned?

With regard to differences in content as described in the NAEP framework, SAT items had limited coverage of the knowledge and skills described by the NAEP standards “Locate/Recall” and “Critique/Evaluate.” This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.

27
With regard to differences in complexity, NAEP items and objectives had a range of depth of knowledge including items at DOK Levels 1, 2, and 3, while SAT items and objectives were coded primarily at Levels 2 and 3.

Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT “Passage-Based Reading” objectives and NAEP “Integrate/Interpret” objectives, as well as some important differences.

**Reading: ACCUPLACER**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade reading assessment and the ACCUPLACER reading test. The ACCUPLACER is used specifically to determine whether entry-level students have the reading skills necessary for college level work or require remedial reading courses. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf).

Overall, the study found similar content in the NAEP 12th grade reading assessment and the ACCUPLACER reading test, although the content of NAEP is much broader and complex. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

**What is the correspondence between the reading content domain assessed by NAEP and that assessed by ACCUPLACER?**

The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts.

The shared emphasis on the comprehension and interpretation of informational text can also be seen in the alignments on both tests to the ACCUPLACER framework. Although the ACCUPLACER standards do not explicitly refer to text type, they focus almost exclusively on elements typical of informational text. A majority of both NAEP and ACCUPLACER items were matched to the ACCUPLACER standard “Inferences,” and both tests had notable percentages of alignments to “Direct statements and secondary ideas” and “Applications.” A smaller percentage of items on both tests were aligned to “Identifying main ideas.”
To what extent is the emphasis of reading content on NAEP proportionally equal to that on ACCUPLACER?

As previously discussed, the alignments both within and across frameworks show that both tests emphasize the comprehension and interpretation of informational text, particularly through the use of inference. Within this broad area of convergence, however, there are differences in emphasis revealed in the alignments to specific objectives within both frameworks. In relation to the NAEP framework, the NAEP short-version items showed a far greater emphasis on the comprehension of vocabulary in context (Objective 4.a) and on the analysis of an author’s use of language (Objective 1.d). In relation to the ACCUPLACER framework, NAEP items showed more emphasis on the use of inference to interpret text (“Inferences”). The higher percentage of NAEP items aligned to “Applications” also reflects the greater emphasis in NAEP on understanding authors’ use of language.

In relation to the ACCUPLACER framework, the ACCUPLACER items showed a greater emphasis than the NAEP items on the identification of main ideas. In relation to the NAEP framework, the ACCUPLACER items showed more emphasis on the recall of specific details, facts, and information (NAEP 1.1.a).

In general, in the cross-framework alignments, the matches found in each test to the other’s framework (NAEP to ACCUPLACER and ACCUPLACER to NAEP) tended to be for the most general objectives within that framework. For example, the great majority of hits for ACCUPLACER items to NAEP objectives for “Integrate/Interpret” were to two of the most broadly stated NAEP objectives, “Draw conclusions” (2.3.b) and “Compare or connect ideas” (2.1.b). Many of the more specific NAEP objectives for “Integrate/Interpret,” such as “Find evidence in support of an argument” (2.2.c), received far fewer or no hits from ACCUPLACER items. Compared to ACCUPLACER, the NAEP items were more evenly distributed among NAEP objectives.

The majority of alignments for NAEP items to ACCUPLACER standards were also to the broadest of those standards—“Inferences” and “Applications,” both of which overlap in content with a number of NAEP objectives but at a higher level of generality. The more specific ACCUPLACER standard, “Identifying main ideas,” received far fewer alignments from NAEP items.

Are there systematic differences in content and complexity between the NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between the NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire reading subdomains are missing or not aligned?

In regard to differences in content, NAEP addresses reading skills related to both literary and informational text, while ACCUPLACER does not address reading skills specific to literary text. As expected, based on the framework-to-specifications [review]… ACCUPLACER items had minimal matches to NAEP objectives for literary text. The main area of alignment of ACCUPLACER items to the NAEP framework, NAEP
objectives in “Locate/Recall” and “Integrate/Interpret,” applied to informational text only or to both informational and literary text.

The ACCUPLACER items also had minimal to no coverage of the NAEP standard “Critique/Evaluate.” Overall, the language of the ACCUPLACER objectives (“understand,” “comprehend,” “recognize”) places more emphasis on comprehension and interpretation of text (“distinguish the main idea from supporting ideas” or “perceive connections between ideas made—implicitly—in the passage”) than on critical analysis or evaluation (“Evaluate the strength and quality of evidence used by the author to support his or her position” in NAEP Objective 3.3.b, or “Judge the author’s craft and technique” in NAEP Objective 3.1.a).

In regard to complexity, both assessments were found to meet the criteria for depth of knowledge consistency in relation to their own framework. In relation to the NAEP framework, however, only the NAEP items met the criteria for DOK consistency for all NAEP standards. The ACCUPLACER items met the criteria for depth of knowledge consistency only for NAEP “Locate/Recall.”

Although the majority of the ACCUPLACER item alignments were to objectives for NAEP “Integrate/Interpret,” over half of these items were found to have a DOK level below that of the standard. In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the more in-depth reasoning required by DOK Level 3. NAEP, by contrast, includes much longer reading passages and both multiple-choice and constructed-response items.

NAEP covers skills specific to the comprehension and analysis of literary text while ACCUPLACER does not. In addition, NAEP covers the skills of evaluating and critiquing text, skills not addressed by ACCUPLACER. Finally, NAEP has a wider range of cognitive complexity than ACCUPLACER, with a substantially higher percentage of items at DOK Level 3, requiring more in-depth analysis or evaluation. However, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

Overall, the NAEP items covered a broader range of cognitive complexity than the ACCUPLACER items. This is also apparent in the frameworks. The three NAEP standards, defined in terms of three different “cognitive targets” (“Locate/Recall,” “Integrate/Interpret,” and “Critique/Evaluate”), cover a broader range of cognitive complexity supported by the use of longer reading passages and the inclusion of both short and extended constructed-response items. The language of the ACCUPLACER standards (“understand,” “comprehend,” “recognize”) places more emphasis on comprehension and interpretation of text (e.g., “distinguish the main idea from supporting ideas” in ACCUPLACER A, “Identifying main ideas,” or “perceive connections between ideas made—implicitly—in the passage” in ACCUPLACER C, “Inferences”) than on critical analysis or evaluation (e.g., “Evaluate the strength and quality of evidence” in NAEP 3.3.b, or “Judge the author’s craft” in NAEP 3.1.a). In addition, the use of very short reading passages and exclusively multiple-choice items in ACCUPLACER may be less conducive to the cognitive complexity typical of DOK Level 3 items. Although the
NAEP items show a greater range of cognitive complexity and a greater emphasis on critical thinking, both tests show a similar emphasis on applying interpretive skills and inferential reasoning to the understanding of informational text.

The Content Alignment Studies: Summary Discussion for Reading

The NAEP 12th grade reading framework, test questions, and, for constructed response items, the score category rubrics, were compared with the analogous domain descriptions and test questions for the ACT, SAT, and ACCUPLACER reading tests. These three tests are used for college admissions and placement. They are well established and have been used for these purposes for many years by professionals in postsecondary education. The test publishers regularly survey secondary and postsecondary educators about relevant content and have conducted research that supports the validity of the test content for the intended inferences and uses. The underlying assumption is that if the content of the 12th grade NAEP reading assessment is similar to the content of these reading tests, then the NAEP content is directly related to “academic preparedness for college.”

The ACT study found that “All of the skills highlighted in the ACT [reading] domain and in the [ACT] College Readiness Standards [for reading] were identified within the NAEP Reading framework.” At the same time, there was content measured by NAEP that was not present in the ACT reading test. In assigning 211 NAEP 12th grade reading items and rubric score categories to the ACT College Readiness Standards for reading, there were 137 positive classifications, or about 65% of the possible classifications. The multiple choice items and rubric score categories that could not be classified were those that measured content not measured by the ACT reading test.

The SAT study found that “Overall, the alignment results across the two sets of items and frameworks show a strong area of overlap in their coverage of SAT “Passage-Based Reading” objectives and NAEP “Integrate/Interpret” objectives, as well as some important differences.” With respect to the differences, “…SAT items had limited coverage of the knowledge and skills described by the NAEP standards “Locate/Recall” and “Critique/Evaluate.” This difference is also reflected in test format, with the use of longer reading passages and both constructed-response and multiple-choice items in NAEP. In comparison, all SAT items are multiple-choice. With regard to differences in content as described in the SAT framework, NAEP does not include sentence-completion items.”

The ACCUPLACER study found that “The greatest commonality between the two tests is in their shared emphasis on the broad skills of comprehending and interpreting informational text, primarily through inferential reasoning. This is evident in the majority of items on both tests (two-thirds to three-fourths) matched to the NAEP standard “Integrate/Interpret: Make complex inferences within and across texts.” On both tests, the majority of alignments to “Integrate/Interpret” were to objectives that apply to informational text only or across both informational and literary texts…Overall, the NAEP [frameworks and] items covered a broader range of cognitive complexity than the ACCUPLACER items…The three NAEP standards, defined in terms of three different “cognitive targets” (“Locate/Recall,” “Integrate/Interpret,” and “Critique/Evaluate”), cover a broader range of cognitive complexity supported by the use of
longer reading passages and the inclusion of both short and extended constructed-response items.”

The results across the three studies are consistent. In general, the content of the ACT, SAT, and ACCUPLACER reading tests are present in NAEP, but NAEP is generally broader. Alignment between NAEP and the other three respective assessments is substantial, but not perfect; perfect alignment is not expected. A component of the SAT critical reading assessment not present in NAEP is sentence completion, measuring vocabulary knowledge in a different way than NAEP does.

These results support the conclusion that

- The NAEP 12th grade reading assessment measures academic knowledge and skills that are also covered by other assessments designed and used to make judgments about the academic preparedness of college freshmen for placement into entry-level, credit bearing, non-remedial college courses that meet general education degree requirements, and
- NAEP 12th grade reading test items and rubric scoring categories for items are appropriate for obtaining evidence of test takers’ possession of knowledge and skills needed for college freshmen to be placed into ECNRG courses requiring college level reading.

**The Content Alignment Studies: Mathematics Results**

**Mathematics: ACT**

The Governing Board contracted with ACT, Inc. to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACT mathematics test. The full report can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACT-NAEP_Math_and_Reading_Content_Comparison.pdf).

The mathematics panel was composed of 7 members, with expertise in mathematics instruction at the high school and college levels. The panel was about evenly divided in terms of prior familiarity with either the ACT or NAEP mathematics domains.

The panel found considerable similarity in the content of the NAEP 12th grade mathematics assessment and the ACT. For example, the NAEP 12th grade mathematics framework was compared to the ACT mathematics domain and the ACT College Readiness Standards for mathematics. The ACT College Readiness Standards (CRS) are descriptions of the content (i.e., the knowledge and skills) measured by the ACT mathematics test in score bands along the ACT 1-36 point scale from 13-36 (see [http://www.act.org/standard/planact/math/index.html](http://www.act.org/standard/planact/math/index.html)). The panel concluded that

“... the two assessments have much of their content domains in common. However, in the NAEP-to-ACT comparison, the difference in specificity with which the domains are articulated in the assessment documents left the panel uncertain as to whether a number
of NAEP content topics—those pertaining to transformations, probability, statistics, and data analysis—are assessed by the ACT. In addition, there was some uncertainty within the panel on the degree to which higher-order analytic skills were assessed, and it was the sense of the panel that the ACT Mathematics Test contained few items involving high mathematical complexity, at least as the NAEP defines it. With regard to the ACT to-NAEP comparison, the Mathematics panel found nearly all of the ACT Mathematics domain and College Readiness Standards reflected in the NAEP Mathematics domain, but determined that a number of the lower-level topics in the ACT Pre-Algebra subdomain were more consistent with Grade 8 NAEP topics. All of these points suggest that while there may be substantial overlap in what the two assessments measure and how they measure it, there are areas of difference, as well. (p. 17)

The mathematics panel also conducted an item classification study, in which the NAEP 12th grade mathematics items were classified in relation to the ACT College Readiness Standards for Mathematics.

An item or score category was deemed “classified” if there was majority agreement; that is, if at least 4 of the 7 panel members agreed about the score band to which an item (or creditable score category under an item rubric) was assigned.

Of the 229 determinations to be made, panel members believed that every item or rubric category could be classified to some CRS score range. However, there were 39 for which there was no majority agreement (17 multiple choice items and 22 rubric categories) on what the classification should be; therefore those items were not considered assigned to a CRS score band. Of the remaining 190 determinations, 24 were unanimous, 142 involved classifications to adjacent score ranges and 24 involved classifications to non-adjacent score ranges.

Of the 108 multiple choice items, 91 (or 84%) were classified. Of the 121 rubric score categories for items, 99 (or 82%) were classified.

Of the 190 classifications, 10 were in the score bands from 13-19; 180 of the classifications were in the score bands from 20-36. This is noted because the ACT College Readiness Benchmark for mathematics is 22. The ACT College Readiness Benchmark signifies the score at which a student has a 50% chance of attaining a grade of B or better in a relevant subject and a 75% change of a C or better. In addition, the Governing Board conducted a survey of postsecondary institutions’ use of tests in making entry-level decisions about placement into remedial or regular credit-bearing courses. With respect to the ACT, 19 was the mean mathematics score below which students were deemed to need remedial course work in mathematics (Fields and Parsad, p. 13). Although this provides a context for the study results, it must be kept in mind that in making their judgments about content, the panelists did not have data about NAEP item difficulty or data on how performance on NAEP compares with performance on the ACT.

Finally, although the study results support the conclusion that the 12th grade NAEP mathematics assessment measures content that is also covered by other assessments designed and used to make judgments about academic preparedness for college, it is noted that the study was
conducted by ACT, Inc., not an independent third party. Further, because a different methodology was used, the study results are not directly comparable to the results for the SAT and ACCUPLACER alignment studies in mathematics.

**Mathematics: SAT**
The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the SAT mathematics test. WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf).

Overall, the study found similar content in the NAEP 12th grade mathematics assessment and the SAT mathematics test. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

*“What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by SAT?*

At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard. Each framework contains both general and specific objectives, although the SAT objectives, which are presented as content topics without indication of the cognitive level at which that content would be assessed, may be interpreted as more general than the NAEP objectives.

Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions.

*To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on SAT?*

The greatest commonality between the two tests is their emphasis at the standard level. This is evident in the distribution of percentages of total hits from both assessments matched to each set of standards. Although there are some differences of emphasis, such as the full NAEP item pool’s greater proportion of alignment to SAT “Data analysis, statistics, and probability,” and the SAT short-version’s greater proportion of alignment to SAT “Geometry and measurement,” the proportions of alignments to “Algebra and functions” and “Number and operations” are comparable. There is also considerable overlap among some specific skills, with both assessments addressing many of the same NAEP “Number properties and operations” objectives and SAT objectives…
Despite the difference in the degree of specificity of the two frameworks (most NAEP objectives are much more finely grained than the SAT objectives), it is clear that both assessments emphasize a number of the same or closely related skills. These include properties, equivalence, and operations on rational numbers (included in NAEP Goals 1.1 and 1.3 and included in SAT Objective N.2) and properties of two-dimensional shapes (included in NAEP Goals 3.1 and 3.3 and included in SAT Objective G.6).

Are there systematic differences in content and complexity between NAEP and SAT assessments in their alignment to the NAEP framework and between NAEP and SAT assessments in their alignment to the SAT framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework. The language of the objectives in the SAT framework is sufficiently broad to encompass the range of the NAEP items. For example, SAT Objective A.10, “Basic concepts of algebraic functions,” may accommodate most of the items aligning to the seven objectives within NAEP Goal 5.1, “Patterns, relations, and functions.” Finally, some NAEP items were found to be uncodable to the SAT objectives. These items assessed skills not present in the SAT framework.

The two tests are also similar in the average DOK [Depth of Knowledge] levels of items. However, while most items in both tests were found to be at DOK Level 2, NAEP items had a wider range of DOK than did SAT items, with more NAEP items coded to Levels 1 and 3. The Level 3 NAEP items often involved application of concepts through short or extended constructed-response items. Both tests also met depth-of-knowledge consistency overall (with each not meeting this criterion for only one standard as rated by one panel).

Overall, despite differences in alignment at the detailed specific objective level, differences in emphasis at the standard level, and a small difference in ranges of depth of knowledge, there is considerable overlap of content and complexity between [the NAEP 12th grade mathematics assessment and the SAT mathematics test].”

**Mathematics: ACCUPLACER**

The Governing Board contracted with WestEd, an independent third party, to conduct the content alignment study comparing the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test. The ACCUPLACER is used specifically to determine whether entry-level students have the mathematic knowledge and skills necessary for college level work or require remedial mathematics courses.
WestEd conducted the content alignment study using the design developed for the Governing Board by Norman Webb. The full report of the content alignment study can be found at [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Math_Content_Comparison.pdf).

Overall, the study found similar content in the NAEP 12th grade mathematics assessment and the ACCUPLACER mathematics test, although the content of NAEP is much broader and complex. Following below is an excerpt from the Executive Summary of the report (pp. iv-vi).

“What is the correspondence between the mathematics content domain assessed by NAEP and that assessed by ACCUPLACER?

The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra (included in NAEP’s “Number properties and operations” and “Algebra” standards and in ACCUPLACER’s “Arithmetic,” “Elementary algebra,” and “College level math” standards), although their respective degrees of alignment and focus in these subdomains vary. Whereas the NAEP items focus primarily on number or number operations and algebra content at the grade 12 level, with an emphasis on problem solving and application of concepts at that grade level, the ACCUPLACER items span a wider developmental and grade-level range (from basic to more advanced). This difference in focus is consistent with the purposes of the two assessments and their frameworks. The NAEP objectives are written to describe assessable content for grade 12 mathematics; thus, the 130 objectives tend to address the skills and concepts specific to that grade. The purpose of ACCUPLACER is to help determine appropriate placement for an individual student, and so the 87 ACCUPLACER objectives are spread more broadly across grade levels and are intended to be more general.

To what extent is the emphasis of mathematics content on NAEP proportionally equal to that on ACCUPLACER?

Regarding alignment to the NAEP framework, within the “Number properties and operations” and “Algebra” standards, NAEP items had broader overall coverage of the NAEP objectives than did ACCUPLACER. The 42 NAEP items (the short version used for within-framework alignment) aligned to 72 NAEP objectives, whereas the 105 ACCUPLACER items (one complete form of each of the three ACCUPLACER Mathematics Core tests) aligned to only 56 NAEP objectives, with 44% of the ACCUPLACER item alignments aligning to only three NAEP objectives (all in “Number properties and operations” and “Algebra”). These differences in breadth and emphasis between the two assessments were evident across all NAEP standards. For example, in each assessment, items were aligned to four NAEP “Algebra” objectives for which the other assessment had no alignments, reflecting differences in emphasis within that standard.

Regarding alignment to the ACCUPLACER framework, ACCUPLACER items in the short version of 45 items covered all three standards—“Arithmetic,” “Elementary algebra,” and “College level math”—with a relatively even distribution, although
“College level math” had the lowest percentage of item alignments. NAEP items in the full pool of 164 items also covered “Arithmetic,” “Elementary algebra,” and “College level math,” with a fairly even distribution of approximately one-third of NAEP codable items aligned to each standard, although “Elementary algebra” received somewhat fewer item alignments. Despite these differences in emphasis, however, considering only codable items, the percentages of alignments to each ACCUPLACER standard were relatively evenly distributed in both assessments and similar in distribution across assessments. At the objective level, the distribution of item alignments to objectives was relatively even on both tests, although each assessment was aligned to some objectives to which the other was not.

In summarizing cross-framework alignment, there was somewhat less even distribution of items than observed in within-framework alignment. The majority of items on each test were found to align to objectives on the other test. However, the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from “Number properties and operations” in the NAEP framework, and four of the five goals in “Algebra.” Conversely, the NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives.

Are there systematic differences in content and complexity between NAEP and ACCUPLACER assessments in their alignment to the NAEP framework and between NAEP and ACCUPLACER assessments in their alignment to the ACCUPLACER framework? Are these differences such that entire mathematics subdomains are missing or not aligned?

Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives (20 were rated uncodable by the majority of panelists in each panel). For other NAEP items that were aligned to an ACCUPLACER objective, there were often parts of those items not addressed by the objective. These items were coded as aligned, since they do assess an ACCUPLACER objective, but parts of the items also cover other skills not included in the ACCUPLACER framework.

Regarding differences in alignment of complexity, the items from both tests that aligned to the NAEP standards met the typical depth-of-knowledge (DOK) consistency threshold; that is, the items assessed the objectives at or above the DOK level of the objective. The items from both tests that aligned to the ACCUPLACER standards had somewhat different ranges of DOK. The ACCUPLACER short-version items were divided fairly evenly between Level 1 and Level 2. The NAEP items aligned to the ACCUPLACER framework had a wider range of DOK, with items at Level 1, 2, and 3, and a greater emphasis on Level 2 than was in the ACCUPLACER items.”
The Content Alignment Studies: Summary Discussion for Mathematics

The NAEP 12th grade mathematics framework, test questions, and, for constructed response items, the score category rubrics, were compared with the analogous domain descriptions and test questions for the ACT, SAT, and ACCUPLACER mathematics tests. These three tests are used for college admissions and placement. They are well established and have been used for these purposes for many years by professionals in postsecondary education. The test publishers regularly survey secondary and postsecondary educators about relevant content and have conducted research that supports the validity of the test content for the intended inferences and uses. The underlying assumption is that if the content of the 12th grade NAEP mathematics assessment is similar to the content of these mathematics tests, then the NAEP content is directly related to “academic preparedness for college.”

The ACT study found that “With regard to the ACT to-NAEP comparison…nearly all of the ACT Mathematics domain and College Readiness Standards [are] reflected in the NAEP Mathematics domain, but…a number of the lower-level topics in the ACT Pre-Algebra subdomain were more consistent with Grade 8 NAEP topics.” In the NAEP-to ACT comparison, there was uncertainty about “…whether a number of NAEP content topics—those pertaining to transformations, probability, statistics, and data analysis—are assessed by the ACT….and the degree to which higher-order analytic skills were assessed…and it was the sense of the panel that the ACT Mathematics Test contained few items involving high mathematical complexity, at least as the NAEP defines it.”

The SAT study found similar content in the NAEP 12th grade mathematics assessment and the SAT mathematics test. “At the standard level, the wording of the standards in the two frameworks is very similar. Both the NAEP and SAT frameworks include virtually the same five broad content categories, with SAT combining geometry and measurement into one standard…Although the structures of the two frameworks differ greatly beyond the standard level (including the NAEP framework having three levels while SAT has two), the mathematics areas typically expected of grade 12 students—number and operations, geometry and measurement, data analysis and probability, and algebra—are addressed in somewhat similar proportions…While there is considerable overlap between the two assessments, primarily in the intersection of the NAEP “Algebra” and SAT “Algebra and functions” standards, there are notable differences as well. The SAT items had a somewhat limited range of coverage of the NAEP standards “Measurement,” “Geometry,” and “Data analysis, statistics, and probability,” with several goals receiving few item alignments. Even given the minimal coverage of some of the goals within each NAEP standard by SAT items, however, almost all NAEP items found a match in the SAT framework.

The ACCUPLACER study found that “The NAEP and ACCUPLACER assessments both cover certain content traditionally expected of grade 12 students, namely the two content subdomains of number or number operations and algebra…although their respective degrees of alignment and focus in these subdomains vary…the 105 ACCUPLACER items aligned primarily (90%) to a total of seven out of 24 NAEP goals: three of the six goals from “Number properties and operations” in the NAEP framework, and four of the five goals in “Algebra.” Conversely, the
NAEP items from the full pool of 164 items that aligned to the ACCUPLACER framework were distributed fairly evenly across the three ACCUPLACER standards and found to align to 75 ACCUPLACER objectives...Regarding differences in alignment of content, ACCUPLACER items had very limited coverage of measurement, geometry, and data analysis, content that is not included in the ACCUPLACER framework but that is included in the NAEP framework. Many NAEP items assessing these subdomains were found to be uncodable to the ACCUPLACER objectives…”

The results across the three studies are consistent. In general, the content of the ACT, SAT, and ACCUPLACER mathematics tests are present in NAEP, but NAEP is generally broader. Alignment between NAEP and the other three respective assessments is substantial, but not perfect; perfect alignment is not expected.

These results support the conclusion that
- The NAEP 12th grade mathematics assessment measures academic knowledge and skills that is also covered by other assessments designed and used to make judgments about the academic preparedness of college freshmen for placement into entry-level, credit bearing, non-remedial college courses that meet general education degree requirements for mathematics, and
- NAEP 12th grade mathematics test items and rubric scoring categories for items are appropriate for obtaining evidence of test takers’ possession of knowledge and skills needed for college freshmen to be placed into ECRNG college mathematics courses.

**Discussion of Test Uses and Consequences in Relation to the Proposed Inferences**

The National Assessment of Educational Progress is an independent monitor of student academic achievement in the United States. It reports on achievement at specific points in time and trends in achievement over time. NAEP reports to the public, national and state policymakers, and education leaders. It assesses student achievement at grades 4, 8, and 12 in important subjects. NAEP is used to compare performance across states and for 21 urban school districts. NAEP results are reported by gender, race/ethnicity, socioeconomic status, and for students with disabilities and students who are English language learners.

The audiences and the uses of NAEP are well established. They will not change as a result of the added meaning afforded by the inferences proposed in this validity argument. However, providing familiar external referents for performance on 12th grade NAEP will greatly enhance the understanding of NAEP results by its audiences.

Currently, there are either no or very low stakes consequences associated with the use of NAEP results. NAEP is not used as a basis for evaluating or diagnosing individual students, classroom or school performance, the effectiveness of individual teachers or administrators, or for any other accountability purpose. This will not change as a consequence of the inferences proposed in this validity argument.
Although the uses and consequences of NAEP will not change, employing the proposed inferences for NAEP reporting will bring a potential for misinterpretation. NAEP reports should include text explaining the limitations on interpretation and other caveats that were discussed in detail on pages 8-10 above.

**Summary and Conclusion**

The National Assessment Governing Board decided to determine the feasibility of transforming NAEP into a measure of academic preparedness for college. Consequently, the Governing Board made changes to the NAEP 12th grade reading and mathematics frameworks with the explicit purpose of measuring academic preparedness for college. The Governing Board conducted research that established a high degree of overlap between the content of the NAEP 12th grade reading and mathematics assessments and the content of widely used college admissions and placement tests.

Through a partnership with the College Board, performance on 12th grade NAEP was compared with performance on the SAT mathematics and critical reading assessments, with correlations of .91 and .74 respectively. Analyses of these data examined the average NAEP scores and interquartile ranges for students scoring “at” and “at or above” the College Board College Readiness Benchmarks for reading and mathematics. Similar analyses were conducted using data from the 2005 and 2009 NAEP High School Transcript Studies, using the college readiness benchmarks developed by ACT and by the College Board. A longitudinal study was conducted in partnership with the Florida Department of Education, following the 12th grade students in the state NAEP sample into Florida public postsecondary institutions, employing Florida’s longitudinal data base. The average NAEP scores and interquartile ranges were calculated for the Florida students in relation to the ACT or SAT college readiness benchmarks, whether they achieved a first-year GPA of B- or better, and whether they were placed into a remedial course in their first year of college.

The results of these analyses were consistent across studies and across years. They support the conclusions that students in the NAEP 12th grade distribution at or above the Proficient achievement level in reading and at or above 163 on the NAEP score scale for mathematics are

- likely to be academically prepared for entry-level, credit-bearing non-remedial courses in broad access 4-year institutions and, for 2-year institutions, for entry-level placement, without remediation, into degree-bearing programs designed to transfer to 4-year institutions, and
- not likely to need remedial/developmental courses in reading or mathematics in college

That the NAEP sampling, scaling and statistical procedures yield accurate estimates of the percentage of students scoring at or above a selected cut-score (i.e., NAEP achievement level) is well established as a result of numerous validity studies and evaluations.

Thus, the NAEP 12th grade preparedness research results support the inferences that...
For reading:

Given the design, content, and characteristics of the NAEP 12th grade reading assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness:

the percentage of students scoring at or above Proficient on Grade 12 NAEP in reading is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in reading that would make them academically prepared for college.

For mathematics:

Given the design, content, and characteristics of the NAEP 12th grade mathematics assessment, and the strength of relationships between NAEP scores and NAEP content to other relevant measures of college academic preparedness,

the percentage of students scoring at or above a score of 163 on the Grade 12 NAEP scale in mathematics is a plausible (or reasonable) estimate of the percentage of students who possess the knowledge, skills, and abilities in mathematics that would make them academically prepared for college.

Including these inferences in NAEP 12th grade reports will add meaning to the interpretation of the NAEP 12th grade results. However, steps must be taken to avoid potential misinterpretation. NAEP reports using these inferences must also include the limitations on interpretation and caveats described previously in this validity argument. In addition, the reports should explain the rationale for NAEP reporting on academic preparedness and describe appropriate and inappropriate uses of the results.
References


WestEd ACCUPLACER Reading Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/ACCUPLACER-NAEP_Reading_Content_Comparison.pdf).

WestEd SAT Mathematics Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Mathematics_Content_Comparison.pdf).

WestEd SAT Reading Report; [http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf](http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/content-alignment/SAT-NAEP_Reading_Content_Comparison.pdf).

**National Assessment Governing Board**

**Nominations Committee**

**August 3, 2013**

**7:30 – 8:15 am**

**AGENDA**

**Closed Session 7:30 – 8:15 am**

<table>
<thead>
<tr>
<th>Time</th>
<th>Agenda Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 – 7:35 am</td>
<td>Welcome, Introductions, and Agenda Overview</td>
</tr>
<tr>
<td></td>
<td><em>Tonya Miles, Chair</em></td>
</tr>
<tr>
<td>7:35 – 8:15 am</td>
<td>Status of Nominations for Board Terms Beginning on October 1, 2013 and Planning for 2014 Cycle</td>
</tr>
<tr>
<td></td>
<td><em>Committee Members</em></td>
</tr>
</tbody>
</table>
May 2013 Board Blue Sky Discussion – Staff Proposals for Next Steps

**Category:** Understanding Educational Assessment and Defining NAEP’s Unique Role in the Assessment Landscape

**Board Member Ideas:**
- Explain different types of tests and the purposes served by each.
- Explain the role and importance of NAEP in the “assessment landscape”—how NAEP’s unique mission complements the contributions of other kinds of assessments.
- Develop an assessment literacy initiative for a general public audience (particularly parents and grand-parents of school-age children), and students.

**Staff Proposal for Next Steps:**
1. Appoint a small working group of Board members (e.g., 3-4 individuals from various NAGB standing Committees) to oversee the assessment literacy initiative.
2. Identify goals and strategies for the assessment literacy initiative.
3. Define the intended audiences and develop initial outreach strategies, including effective ways to reach parents (e.g., through their children’s teachers)
4. Explore creative, effective ways to use technology to advance the assessment literacy goals including TED talks, interactive website materials, etc.
5. Work with the Board’s communications contractor (under the new contract, beginning in the fall of 2013) to develop an assessment literacy outreach plan.
6. Coordinate the assessment literacy initiative with planning for the Board’s January 2014 Parent Summit.
7. Obtain input and feedback on the plan from various groups including NCES, the Governing Board’s CCSSO Policy Task Force, the Board’s Business Policy Task Force, and others.
8. Determine effective ways the Board can collaborate with NCES and their contractors on this initiative.
9. Present a revised outreach plan to the full Board for discussion in December 2013. The plan should include a timeline, priorities for implementation, and ways to evaluate the effectiveness of various components before, during, and after the outreach initiative.
10. Begin to implement the plan and monitor its effectiveness using various measures.
**Board Blue Sky Discussion – Staff Proposals for Next Steps**

**Category:** Assessing Affective Skills (AKA Work Readiness Skills, 21st Century Skills, and Soft Skills)

**Board Member Ideas** (from May 2013):
- Measure important affective attributes.
- Add affective skills survey questions to measure important determinants of academic success.
- Make NAEP as relevant as the U.S. Census.

**Definition:** Although there is no universal definition of these affective skills, the National Academy of Sciences has been exploring strategies for assessing these skills. Their Committee on the Assessment of 21st Century Skills initially identified five skill areas and later collapsed them into the following three clusters.
- **Cognitive skills:** nonroutine problem solving, critical thinking, systems thinking
- **Interpersonal skills:** complex communication, social skills, teamwork, cultural sensitivity, dealing with diversity
- **Intrapersonal skills:** self-management, time management, self-development, self-regulation, adaptability, executive functioning

**Staff Proposal for Next Steps:**

Explore the current state-of-the-art in assessing or conducting surveys on affective skills before making a determination about how or if NAEP can/should be used to collect and report information for this area.

1. Have staff or a consultant prepare a “white paper” summary or issues paper about the research and publications 1 related to assessing affective skills for Board review/discussion.
2. Invite a panel of the experts (identified via the process of preparing the “white paper”) to discuss with the Board the state-of-the-art of assessments or self-report surveys on which affective skills can be reported, as well as the viability for NAEP to utilize these methods.
3. If warranted, establish an expert panel or ad hoc Board committee to examine what NAEP already collects in the area of affective skills (e.g., the cognitive skills area as defined above) and to make recommendations to the Governing Board about how NAEP could provide nationally representative information on affective skills, and whether implementation should be considered by the Board.
4. If warranted, begin planning for how NAEP could collect and report information on affective skills, possibly including the reporting of existing data, development of a framework, designing and scheduling a special survey (assessment and/or background questions) and reporting these results.

---

May 2013 Board Blue Sky Discussion – Staff Proposals for Next Steps

**Category:** Information for Educators about NAEP Data and Resources

**Board Member Ideas:**
- Focus on educator preparation about NAEP (along with parental education) for future teachers and administrators.
- Create a presentation to send to schools of education for pre-service teachers about how to use NAEP tools.
- Help teachers inform their practice, thinking of NAEP as a summative assessment in relation to teachers’ use of formative assessments.
- Add more content to the NAEP website for teachers to use.
- Do more outreach to principals and school systems.
- Develop a campaign to make NAEP more relevant to all stakeholders.

**Staff Proposal for Next Steps:**

Board members Shannon Garrison and Father Joseph O’Keefe, working with other Board members as a small planning group, will develop proposals for providing information to pre- and in-service teachers and administrators on using NAEP data and resources, including a presentation that will be piloted at Boston College.

1. Appoint planning group.
2. Identify extant pre-service and in-service professional development programs that use NAEP data and resources.
3. Identify extant pre-service and in-service professional development programs that address data-based instructional decision making.
4. Identify extant courses for teachers and leaders that include topics on classroom and/or formative and summative assessment.
5. Identify the target audience for the pilot at Boston College.
6. Design and conduct a needs analysis of the target pilot audience to determine the NAEP data and resources of greatest relevance and various print and on-line media for dissemination.
7. Develop the pilot presentation.
8. Conduct the pilot at Boston College.
9. Evaluate the results of the pilot, the applicability to other venues, and conduct one or more tryouts in other settings.
10. Develop recommendations about the potential for broader use and dissemination of NAEP data and resources to the target population for Board consideration.
Monitoring What Matters About Context and Instruction in Science Education: A NAEP Data Analysis Report

PREPARED FOR THE NATIONAL ASSESSMENT GOVERNING BOARD

By Alan Friedman and Alan Ginsburg

July 2013

Alan Friedman is a museum and science communication consultant and former director of the New York Hall of Science. He is a member of the National Assessment Governing Board. Alan Ginsburg is an education consultant and analyst. He is former director of policy and program evaluation services for the U.S. Department of Education. The data analyses and interpretations in this report are those of the authors and do not necessarily represent the views of the National Assessment Governing Board.
# Table of Contents

Executive Summary ............................................................................................................................................. 3

1. Introduction ........................................................................................................................................... 6

2. Methodology ......................................................................................................................................... 9
   Analytic Framework for Science Education Background Questions .......................................................... 9
   Data Characteristics and Issues .................................................................................................................. 12
   Organization of the Background Variable Analyses of Science Education .................................................. 15

3. Student Engagement: Interest, Attitudes, Behavior and Identity .......................................................... 15
   Research on Student Engagement ............................................................................................................... 15
   NAEP Background Questions About Students’ Engagement Toward Science ............................................ 18
   NAEP Background Variable Data on Student Engagement in Science ..................................................... 19

4. Teachers of Science .............................................................................................................................. 26
   A. Teachers with Science Degrees ............................................................................................................. 27
   B. Alternative Certification of Teachers of Science .................................................................................... 31
   C. Teachers of Science Professional Experiences ...................................................................................... 35

5. Science Physical Resources: Availability and Use ............................................................................... 38
   NAEP Science Resource Background Questions ......................................................................................... 38
   Science Labs ................................................................................................................................................ 40
   Instructional Technology for Science in Classrooms .................................................................................. 42
   Science non-technology instructional materials in classrooms ..................................................................... 43
   Use of available science resources ........................................................................................................... 44
   Equity in distribution of science resources ............................................................................................... 45

6. Curriculum and Instruction .................................................................................................................... 47
   Research and Context ................................................................................................................................. 47
   NAEP Background Variables For Science Instruction and Curriculum ...................................................... 48
   Analyses and Findings ............................................................................................................................... 49
   How much weekly time is spent on science instruction? ........................................................................... 49
   What is the basis for structuring a school’s science program? ................................................................. 50
   What science content areas are taught at grade 8? ................................................................................. 51
   What are the science practices and attitudes that receive the greatest instructional emphasis? ............... 52

7. Assessment and Ability Grouping/Differentiated Instruction ............................................................... 55
   What types of assessments are used to assess science? ............................................................................ 56
   How are the uses of assessments tied to educational purpose? ............................................................... 56
   How prevalent is ability grouping and differentiated instruction in science education? .......................... 57

8. Improving the NAEP Science Background Questions and Their Use ................................................ 58

Acknowledgements ....................................................................................................................................... 64
References ....................................................................................................................................................... 64

The authors wish to thank Lawrence Feinberg, Assistant Director for Reporting and Analysis of the National Assessment Governing Board, for his many analytic and editorial contributions to strengthen this report.
Executive Summary

This report explores background variables in the National Assessment of Educational Progress (NAEP) to examine key context and instructional factors behind science learning for eighth grade students. Science education is examined from five perspectives: student engagement in science, science teachers’ credentials and professional development, availability and use of science resources, approaches to science instruction, and methods and uses of science assessment.

KEY FINDINGS

• **Contribution of student engagement in science to student achievement.**
  
  o Students who strongly exhibit attributes representing positive engagement in science—like science, do science after school, science is a favorite subject—tend to have significantly higher average NAEP science scores. For example, the 16 percent of the students who strongly agree that science is a favorite subject have a 25-point NAEP score advantage in grade 8 science over the 21 percent of students who strongly disagree that science is a favorite subject.

  o Students who agree strongly that they take science for extrinsic reasons—science is required or science is of benefit for the future—tend to have lower NAEP science scores. The 17 percent of students who strongly agree that they take science because it is required score 26 points lower in eighth grade science than the 18 percent who answer that they strongly disagree.

• **Decline in teachers of science with science majors.** Between 1996 and 2011, the percentage of grade 8 students taught by teachers who have science majors declined by about 10 percentage points both for teachers in earth/space science and in physics/chemistry. The percentages of students taught by a teacher with a biology or life science major held constant between 1996 and 2011.

• **Alternative certification routes for science teachers.**
  
  o In 2011, the percentage of grade 8 students’ taught science by teachers who entered teaching through an alternative certification route was 5 percentage points higher than for reading.

  o Science teachers with alternate certification are about 20 percentage points more likely to have a science major compared with teachers of science who
have traditional certifications. All of the difference occurs for teachers who major in biology or life science.

- Alternative certification is a major source of teachers for Black and Hispanic students. About 30 percent of grade 8 Black students had science teachers with alternate certified and 25 percent of Hispanic students compared with only 16 percent of White and Asian students.

- **Current science education emphasizes use of conventional science education resources.** Conventional education resources, such as a digital projector, science textbook, and DVD’s, were the most widely used science resources by almost all teachers. By contrast, about 40 percent of students never used a computer to create a chart or graph for science or do a science simulation experiment. The least used resources were very specific computerized technologies, such as graphing calculators, tablet PC’s and handheld computing devices.

- **Science instruction is structured heavily by state and district standards.** The schools clearly identify state and district standards as the key drivers in structuring their science education programs:
  - 86% of students were in schools that structured their program “a lot” according to state standards.
  - 71% of students were in schools that structured their program “a lot” according to district standards.

- **The percentage of students assigned to classrooms by ability has not changed much since 1996. Ability grouping is far less common than differentiated science instruction within classrooms.**
  - In 2011, 24 percent of students were in schools where assignments to science classes were by ability, about the same as 1996 when 21 percent of the students were in schools using ability-grouping assignments.
  - The most common practice for adjusting instruction to student performance is differentiated instruction within science classrooms, used by teachers of 75 percent of eighth grade students.

**RECOMMENDATIONS**

**Expanding current questionnaire topics:**

- Recommendation 1: *NAGB should consider adding questions about the amount of time spent learning science, the nature of science activities in out‐of‐school or after‐school settings, and the coordination of these activities with the regular classroom science program.*
Recommendation 2. NAGB should consider expanding background questions about teacher professional development to obtain information on its nature, duration and quality.

Adding a new questionnaire topic:

Recommendation 3: NAGB should consider adding NAEP background questions to monitor how changes in science standards are affecting instruction, the challenges schools and teachers face in implementing new or changed standards, and whether they are receiving the needed technical assistance and professional development to bring about effective implementation of new standards.

Technical background questionnaire issues:

Recommendation 4. NAGB should explore offsetting the additional time burden from adding additional science background questions through rotating questions in and out of the science assessments and matrix sampling so that a respondent answers only a sample of the background questions.

Recommendation 5. NAGB should recommend to NCES the use of cognitive laboratories to clarify questions and responses in three areas of the science background variables: (1) understanding the generally more positive responses by schools than by teachers to questions about resource availability; (2) exploring the accuracy of school, teacher or student question responses when responses are qualitative and judgmental, such as “a little” or “a lot;” and (3) taking advantage of future science assessments that will be done on the computer to replace interval responses (e.g. 0-2 hours, 2-5 hours) with continuous sliders enabling respondents to drag an arrow to any point along a continuum.

Extending the Usefulness of the Findings

Recommendation 6. NAGB should explore ways to support greater use of the current findings by policymakers and educators and to stimulate further analyses by academic researchers. NAGB should explore with NCES coordinated support for the further use of the NAEP science background variables.
1. Introduction

This report analyzes the background variables collected during administration of the 2011 National Assessment of Education Progress (NAEP) science assessment. The NAEP background variables measure research-supported factors that contribute to student performance. Five sets of background factors are explored: student engagement, teachers of science, physical science resources, curriculum and instruction, and assessment and ability grouping. This exploratory analysis of the background variables in science is part of an effort by the National Assessment Governing Board (NAGB) to enhance the knowledge value derived from the NAEP assessments in science and other content areas.

The National Assessment for Educational Progress (NAEP) was established by Congress in order to provide information describing how well education is functioning in the United States. Starting with its first assessment in 1969 and operating since 1988 under policies set by the National Assessment Governing Board (NAGB), NAEP has proven to be a unique and valuable resource for the public and for everyone involved in the education enterprise.

Congress has given NAGB the final authority on all cognitive assessment and noncognitive background items used in NAEP. Among the noncognitive items Congress directed NAEP to include are race, ethnicity, socioeconomic status, disability, limited English proficiency, and gender. The assessment program has collected such data for more than four decades plus additional noncognitive information about teachers, students, schools, and classroom practices and resources when such information can shed light on academic performance.

While the personnel involved in NAEP are prohibited by statute from using the assessment “to establish, require, or influence the standards, assessments, curriculum, including lesson plans, textbooks, or classroom materials, or instructional practices of States or local educational agencies,” NAEP reports and data can and are used extensively by academics and policy makers to inform their own decisions about important factors in educational achievement. For this reason, the decisions NAGB makes on what background variables to include in the assessment are critical in keeping NAEP useful to the education community.

What background and context items are included in NAEP and the science assessment reviewed in this report must change from time to time, as actual classroom practices and resources change. When NAEP began, there was no such thing as the Internet, there were no computers in K-12 classrooms, few afterschool classes in science, and under two dozen science-technology museums in the entire country. Today on-line science courses are the subject of intense interest; computers have become commonplace in schools and many science laboratories rely on “probeware” and other computer-assisted tools; there are thousands of afterschool science programs across the nation; and over 400 science-technology
centers and museums are visited by tens of millions of students, teachers, and families every year. To provide useful information on background variables and contexts like these, which can influence student achievement, NAGB has to pay attention to current practices and concerns in science, technology, and engineering education.

In the preparation of this paper, we asked several leaders in STEM education what kinds of information they would find valuable from NAEP’s background questions, and we examined various discussion forums for a sample of “hot” topics among STEM educators. Among the common “hot” issues we found under active discussion were:

- How much time is spent in teaching science using the various strategies such as lecture, laboratory, and discussion?
- What factors motivate students to learn science?
- How do teachers deal with student misconceptions and wrong answers about science?
- Do out-of-school time activities, such as science clubs and museum visits, impact classroom performance?
- Are school laboratories used for exploring or for confirming?
- To what extent are student laboratories using virtual science experiments, or using data and instruments available on-line?
- What kinds of teacher professional development are commonly available, such as on-line, blended, active, or passive, and does the kind or duration chosen correlate with student performance?
- Which topics in STEM are of the greatest interest to students, and how does their performance relate to their interests?
- How extensive is participation in “citizen science” projects, and do they contribute to science learning?
- Are school systems prepared to respond to reforms underway in science education to focus teaching and learning on the big cross-cutting topics and issues in science and teach these with greater depth and understanding?

Another list of requested K-12 STEM education indicators, summarized in Exhibit 1-1, was published in the National Research Council (NRC) report Monitoring Progress Toward Successful K-12 STEM Education (National Research Council, 2013). There is substantial overlap with the informal survey we took before we had read the NRC’s report. One key difference is that while affective domain traits such as motivation and interest appear in our informal survey, they do not appear in the NRC’s summary table of desired indicators, below. The NRC’s report discusses the importance of interest, attitudes, and other affective domain indicators, but this realm did not make it into the list of 14 key indicators in the table reproduced above. The NRC says that affective domain indicators and other topics are important but require separate studies (NRC, 2013, p. 33). The NAEP background questions do provide some evidence for these parameters, as we will discuss.
<table>
<thead>
<tr>
<th>Recommendations from Successful K-12 STEM Education (NRC, 2011)</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Districts Should Consider All Three Models of STEM-Focused Schools</strong></td>
<td>1. Number of, and enrollment in, different types of STEM schools and programs in each district.</td>
</tr>
<tr>
<td><strong>Districts Should Devote Adequate Instructional Time and Resources to Science in Grades K-5</strong></td>
<td>2. Time allocated to teach science in grades K-5.</td>
</tr>
<tr>
<td><strong>Districts Should Ensure That Their STEM Curricula Are Focused on the Most Important Topics in Each Discipline, Are Rigorous, and Are Articulated as a Sequence of Topics and Performances</strong></td>
<td>4. Adoption of instructional materials in grades K-12 that embody the Common Core State Standards for Mathematics and A Framework for K-12 Science Education.*</td>
</tr>
<tr>
<td></td>
<td>5. Classroom coverage of content and practices in the Common Core State Standards and A Framework for K-12 Science Education.</td>
</tr>
<tr>
<td><strong>Districts Need to Enhance the Capacity of K-12 Teachers</strong></td>
<td>6. Teachers’ science and mathematics content knowledge for teaching.</td>
</tr>
<tr>
<td></td>
<td>7. Teachers’ participation in STEM-specific professional development activities.</td>
</tr>
<tr>
<td><strong>Districts Should Provide Instructional Leaders with Professional Development That Helps Them to Create the School Conditions That Appear to Support Student Achievement</strong></td>
<td>8. Instructional leaders’ participation in professional development on creating conditions that support STEM learning.</td>
</tr>
<tr>
<td><strong>Policy Makers at the National, State, and Local Levels Should Elevate Science to the Same Level of Importance as Reading and Mathematics</strong></td>
<td>9. Inclusion of science in federal and state accountability systems.</td>
</tr>
<tr>
<td></td>
<td>10. Inclusion of science in major federal K-12 education initiatives.</td>
</tr>
<tr>
<td></td>
<td>11. State and district staff dedicated to supporting science instruction.</td>
</tr>
<tr>
<td><strong>States and National Organizations Should Develop Effective Systems of Assessment That Are Aligned With A Framework For K-12 Science Education and That Emphasize Science Practices Rather Than Mere Factual Recall</strong></td>
<td>12. States’ use of assessments that measure the core concepts and practices of science and mathematics disciplines.</td>
</tr>
<tr>
<td><strong>National and State Policy Makers Should Invest in a Coherent, Focused, and Sustained Set of Supports for STEM Teachers</strong></td>
<td>13. State and federal expenditures dedicated to improving the K-12 STEM teaching workforce.</td>
</tr>
<tr>
<td><strong>Federal Agencies Should Support Research That Disentangles the Effects of School Practice from Student Selection, Recognizes the Importance of Contextual Variables, and Allows for Longitudinal Assessments of Student Outcomes</strong></td>
<td>14. Federal funding for the research identified in Successful K-12 STEM Education.</td>
</tr>
</tbody>
</table>
While a number of the issues raised by our informal survey and by the NRC report could be informed using current NAEP context and instruction item responses, as we will demonstrate below, the current NAEP background questionnaires have two important limitations. One, the current questions fail to address some important science education topics. These include providing almost no information about the amount and kind of exposure to out-of-school science learning, or the quality and relevance of professional development activities. Items are missing which could shed light on some of the greatest challenges schools and teachers are facing in improving science education, including concerns about having to teach too many science topics annually, lack of knowhow in integrating virtual experiments into instruction, and meeting the needs of students who differ greatly in science performance.

Two, many of these questions cannot be answered by NAEP alone, especially since NAEP data provide correlation indicators but not causality research. Nevertheless, NAEP can currently provide invaluable guidance for researchers and policy-makers in pointing to significant correlations, and can provide information about which educational variables are sufficiently common or sufficiently correlated with performance to encourage further research.

This report explores the current science background questions, the science issues and topics they do address, and offers recommendations for improving the background questions in science in the following sections of the report: section 2 describes the methodology of the analysis; sections 3-7 provide findings from the currently available NAEP background data on student engagement, science teachers, science resources, science instruction, and science assessment; and section 8 consists of recommendations to NAGB to improve the collection of the NAEP background variables in science education.

2. Methodology

Analytic Framework for Science Education Background Questions

NAGB questionnaire development process. NAGB follows a systematic process for developing its Science Background Questionnaires. General guidance for this process comes from a framework document published in 2002 (NAGB, 2002).

A 2009 Science Background Question Issues Paper (NAGB, n.d.) identified potential issues for the Science Background Questionnaires to address. The five issue areas identified are: 1. Availability and Use of Instructional Resources; 2. Organization of Science Instruction; 3. Teacher Preparation; 4. Role of Technology in Instruction; and 5. Student Engagement with Science. This listing and the specific issues in each
area represented a broad identification of potential questions with a supporting research rationale. However, for such reasons as limited questionnaire time, some identified issues, such as learning activities outside of school, were omitted in the final questionnaire.

A systematic process was then followed for generating science-specific background variables (WestEd & CCSSO, 2006):

- Informed by the Background Questions Issues Paper, a Planning Committee brainstormed a list of possible variables.
- Next, a sub-group of Steering Committee members reviewed and culled the list and suggested additional items.
- NCES staff members then refined items from this list and sent a draft of this paper out for review to members of the Planning Committee and Steering Committee, who had expressed interest in being reviewers.
- After receiving feedback from these individuals, staff made appropriate changes and sent the document to NAGB for a preliminary review.
- After making edits suggested by NAGB, the document was sent to the entire Planning Committee and Steering Committee for review.
- After incorporating members’ feedback, the document was then sent back to NAGB for presentation to the NAGB Assessment Development Committee and the full board.

The final questionnaires (http://nces.ed.gov/nationsreportcard/bgquest.aspx) for the 2009 and 2011 assessment are organized by respondent categories of students, teachers, and schools:

- “Student questionnaires include items on: ethnicity/race; resources (magazines, books, computers) available in the home; the student’s attendance; language spoken in the home; education level of the parents (8th and 12th grade questionnaires only); and the student’s program of study (12th grade questionnaire only).”
- “Teacher questionnaires include items on: race/ethnicity; years teaching; certification; number of courses in a subject area; professional development activities; leadership activities; and school improvement activities. Teacher questionnaires are completed by teachers in grades 4 and 8. NAEP typically does not collect grade 12 teacher information because of the difficulty of drawing a valid sample including the inability to relate student and teacher responses.”
- “School questionnaires included items on: general characteristics of the school such as grades taught; number of hours of instruction; enrollment; percentage of limited English proficient students; participation in the
While an organization of questions by respondent is useful for assessment administration, this report is focused on exploring the NAEP background variables organized around critical factors in the delivery of science education. These factors may cut across questionnaires. To guide the exploration of the science background variables, an analytic framework describing the key factors in the delivery of science education was developed (Exhibit 1-1). This framework identifies the potential factors influencing science education and their relationships.

- On the left side of the framework are **inputs** supporting instruction including teachers, parents, and education policies. As noted in the introduction, with respect to education policy, national education organizations in concert with many states are reviewing their K-12 science education standards and this has enormous potential for changing science instruction.

**Exhibit 2-1: Science education analytic framework (K-12)**

- Inputs feed into **formal student instruction at school**. This includes four factors of instructional time, curriculum, pedagogy and school science
laboratories and classroom science resources

- Out-of-school learning can reinforce school science instruction and student learning and can produce both cognitive and affective domain learning in parallel or independently of what happens in school (*Learning Science in Informal Environments*, NRC, 2009). Out-of-school learning can be delivered through a formal afterschool program or informal exposure to science at a museum, the web or other non-formal learning settings. Schools and providers of informal science learning may work in partnership for their activities to reinforce instruction.

Outcomes may be both non-cognitive and cognitive outcomes.

- Non-cognitive outcomes are associated with affecting student engagement. They include student attitudes and behaviors such as liking science or wanting to take science to improve future opportunities.

- Cognitive science outcomes are measured by the knowledge and skills assessed by NAEP. Cognitive outcomes would be expected to be influenced by non-cognitive outcomes as students who are motivated to learn science may put more effort and concentration in science. Also, cognitive outcomes may also influence non-cognitive outcomes, as students may like a subject in which they perform well.

This science education analytic framework is used to organize the exploration of the data produced by the NAEP science background questionnaires.

**Data Characteristics and Issues**

Data considerations for this report include the years for which science background variables will be examined, the selection of individual background variables, measurement issues, analytic capability of the NAEP data explorer, and a cautionary note in making causal interpretations based solely on the NAEP data.

**Years covered.** The NAEP science assessment began in 1996 and was administered in 2000, 2005, 2009 and 2011. It is significant for the present analysis that the NAEP science assessment frameworks were redone in a major way beginning with 2009, in response to major changes in science education curriculum and practices. Because the science assessment outcomes are not comparable for prior to 2009, this analysis will focus primarily on the most recent 2011 assessment.

For a few variables where trend data are of interest, the science background data are shown for both the earliest 1996 and the 2011 periods. An example is the trend in the proportion of teachers of science by whether they have a degree in science.
**Background variable omissions.** The background variables cover many areas of the framework, but there are several notable omissions.

- No information is available for some items. In particular, the out-of-school learning variables are almost non-existent for science education. Thus, no questions are asked about time spent on science homework nor for time spent on science formal or informal learning outside the classroom. Nor is there any data on visits to science museums or on school partnerships with science museums. The only information on out-of-school science learning is a single question on doing science activities out-of-school, without any explicit direct information about frequency or amount, as follows:

  “Please indicate how much you disagree or agree with the following statements about science: I do science-related activities that are not for schoolwork.” Responses: Strongly disagree, Disagree, Agree, Strongly agree.

- Information about professional development is available about topics taken. However, it would be informative to know the amount of time spent in professional development, what teachers thought of its quality, whether they changed their practice in response to the professional development and whether it addressed their science challenges. Given the changes in science standards expected to occur in many states, and the central role of professional development to support curriculum changes, NAEP could potentially provide important information about professional development time and quality.

**Measurement error.** Information from survey questionnaires is subject to measurement error of different types. Three examples of potential measurement errors are:

- Different respondent groups offer inconsistent answers to the same or similar factual background questions. As a case in point, the data analyses will show that schools are much more likely than teachers to indicate that they have adequate supplies of science resources.

- Judgmental measures. Many of the NAEP background variables are questions involving personal judgments. An example, from the student questionnaire:

  “Please indicate how much you disagree or agree with the following statements about science: I like science.” Responses: Strongly disagree, Disagree, Agree, Strongly agree.

  As the analyses below indicate, students’ responses to this question are associated with students’ science scores. It would be helpful in understanding the importance of motivation and liking science to understand
more about what students meant when they interpreted the question and the response of Strongly disagree, Disagree, Agree, Strongly agree.

- Some questions may not be clear. Student grade questionnaire: “I take science only because I have to.”

  It is not clear who doesn’t have to take science at grade 8, so the students’ interpretation of the question may not be clear.

It is recommended that questions where there are uncertain interpretations of response be submitted for review by the NCES cognitive laboratories. This is process where a sample of students is probed in-depth for their question interpretation and the process could help strengthen question accuracy.

**Data Explorer Strengths and Limitations.** The NAEP data set with its matrix sampling is a very large and extremely complex database to analyze statistically. The NCES online NAEP Data Explorer enormously simplifies complex computational analyses of responses. The Data Explorer is in many respects an exceptional analytical tool providing access to all NAEP background variables since the mid-nineties by subject. Nevertheless, while the Data Explorer also permits some cross-variable categorization, the Data Explorer does have its limitations in that at the time of this study it was not capable of performing multivariate statistical analyses that would permit stronger statements about the contribution of an individual variable to an education outcome. For example, multivariate analyses would permit stronger statements about how much student engagement in science differs by students’ race/ethnicity after controlling for differences in students family income and gender, as compared with analyses of a simple correlation alone. According to NCES Commissioner Jack Buckley in remarks before NAGB (May 2013), NCES is committed to adding a multivariate capability in the near future to enhance the capabilities of the NAEP Data Explorer.

**Causal interpretations.** The NAEP samples are repeated regularly at the same grade over different students. This repeated cross-section sampling does not, however, produce a measure of the change in outcomes for the same students over time. Without change data over the same respondents, causal interpretations with the NAEP data must be cautioned.

The approach of this report to making causal statements is to draw on the findings about the causality of a variable in relation to science outcomes from well-designed experimental or quasi-experimental studies. Once causality is shown with some degree of validity from non-NAEP independent sources, then these results can be used to justify examining the variable-outcome relationship from the NAEP assessment data. The NAEP data adds information by shedding light on the strength of the association under different conditions.
Organization of the Background Variable Analyses of Science Education

Consistent with the above framework (Exhibit 2-1), the background analyses explore for grade 8 the following five science education related areas:

- Student Engagement Toward Science
- Teachers of Science
- School Science Resources
- Science Instruction (Curriculum And Pedagogy)
- Science Assessment and Performance-Based Student Groupings

Note that the framework factor related to the science education a student receives out-of-school is not directly broken out because of the limited information on out-of-school learning from the NAEP student questionnaire. The only question available is one that asks students: “I do science-related activities that are not for schoolwork.” There are no measures of the actual amount of time spent in out-of-school learning nor of the nature of the out-of-school learning.

Each analysis of a science area follows a three-part discussion:

- Brief highlights of the research and policy to identify variables that are important to measure within a science area.
- Descriptions of the NAEP background questions available to describe a science area.
- Analyses of the available background question data for that science area.

3. Student Engagement: Interest, Attitudes, Behavior and Identity

Research on Student Engagement

Student engagement embraces a broad category of student characteristics including interest and attitudes toward science, behaviors such as reading about or doing science voluntarily, and identifying oneself personally with science (McCallie, et.al., 2009 offers a variety of definitions and mechanisms for “engagement”). The underlying theory of action behind the student engagement factor is straightforward. If you like a subject and have confidence in your ability in a subject you will concentrate more, work harder and do better in learning that subject.

The direction of causation is also two ways. Do you come to like a subject because
you do well and have higher achievement, or do you do well in a subject because you already liked it? These two possibilities are likely cyclic and mutually reinforcing. This section examines the research evidence about whether this theorized relationship between students’ science attitudes and behaviors and their science test scores exist.

In fact, findings from international mathematics and science assessments have been cited to challenge the common sense idea that students who have positive engagement especially attitudes toward a subject are more likely to be motivated to learn and consequently will do better in a subject. Specifically, researchers point out that some high performing countries, such as Japan and Korea, have students who on average like mathematics and science less than do students in many lower performing countries. Based on this cross-country evidence, Loveless (2006) in a Brookings report concludes: “National indices of student happiness are inversely related to achievement in mathematics (p. 2-1)..... The evidence does suggest that the American infatuation with the happiness factor may be misplaced” (p.2-6)

Often lost in the discussion is that the very same international evidence also shows that within each of these countries, students with more positive attitudes towards mathematics and science consistently outperform students with less positive attitudes (Buckley, 2009). A case in point is the conclusion of the authors of the most recent 2011 results from the TIMSS grades 4 and 8 science assessment:

“Each successive TIMSS assessment has shown a strong positive relationship within countries between student attitudes toward science and their science achievement. The relationship is bidirectional, with attitudes and achievement mutually influencing each other.”(TIMSS, 2012, p.17)

The few rigorous longitudinal studies relating student attitudes to achievement in the STEM area also support the conclusion that student attitudes affect learning in important ways. Two examples are:

- Longitudinal “research examined how motivation (perceived control, intrinsic motivation, and extrinsic motivation), cognitive learning strategies ... and intelligence jointly predict long-term growth in students’ mathematics achievement for 3500 German students over 5 years from grades 5 -10. Using longitudinal data from six annual waves, latent growth curve modeling was employed to analyze growth in achievement. Results showed that the initial level of achievement was strongly related to intelligence, with motivation and cognitive strategies explaining additional variance. In contrast, intelligence had no relation with the growth of achievement over years, whereas motivation and learning strategies were predictors of growth.” (Murayama, et.al., 2012)
- Student engagement can be defined as the level of participation and intrinsic interest that a student shows in school...Extensive evidence exists that engagement and motivation are critical elements in student success and learning. Researchers agree that engaged students learn more, retain more, and enjoy learning activities more than students who are not engaged. Studies have shown a direct link between levels of engagement and achievement in reading and mathematics. (Akey, 2006).

Two studies commissioned by the Wellcome Trust examined the role of informal STEM education in UK and beyond. These studies included an extensive literature survey and review of 553 published works. The Trust concluded that “Researchers have conclusively shown the positive impact on student attainment of learning experiences during the summer gap and the PISA 2006 study shows that school extracurricular activities relate to better performance, enjoyment and more positive attitudes to science” (Wellcome Trust, 2012).

All of these studies are consistent with the National Research Council’s report on learning science in informal environments (2009), which added two strands of affective domain parameters to the four strands of cognitive domain indicators the NRC had identified earlier as essential components of science learning:

<table>
<thead>
<tr>
<th>Six strands for impact for Informal Science Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Develop interest in STEM</td>
</tr>
<tr>
<td>2. Understand STEM knowledge</td>
</tr>
<tr>
<td>3. Engage in STEM reasoning</td>
</tr>
<tr>
<td>4. Reflect on STEM</td>
</tr>
<tr>
<td>5. Engage in the practice of STEM</td>
</tr>
<tr>
<td>6. Identify with the STEM enterprise</td>
</tr>
</tbody>
</table>

The first and sixth strands are affective domain characteristics which are accepted major goals for informal science education, but are normally not explicit goals or parameters assessed in formal education. This makes it especially valuable if NAEP background questions can shed light on how these affective domain qualities, such as interest, attitude, engagement, and identity, are reflected in the student population, and what school and non-school factors are associated with improvements in these qualities.

The NAEP background questions about student attitudes and behavior provide empirical evidence about the baseline and changing affective domain characteristics of U.S. students, and the strength of the association of these characteristics with science achievement within the U.S.
NAEP Background Questions About Students’ Engagement Toward Science

In describing students’ non-cognitive attributes, NAEP identifies eight questions reflecting students’ engagement focused on attitudes and behaviors toward science (Exhibit 3-1).

**Exhibit 3-1. Non-cognitive questions about students’ science engagement: attitudes and behaviors**

![Eight non-cognitive questions](details)

The student response categories to these questions are: “Strongly disagree, Disagree, Agree, Strongly agree”

Three of the questions are about the difficulty, effort and importance a student attaches to the NAEP science test. These questions are useful for understanding the design and interpretation of the cognitive science assessment, but are not germane for this report on how the student background variables influence science education and its effectiveness. Therefore this analysis focuses on the remaining five questions.

Two of the five student attitude and behavior questions explore how much a student enjoys science as a subject:

- How much student likes science.
- Science is a favorite subject.

A related student question is about discretionary science behavior

- How often student does science activities that are not for schoolwork.
This last question is also the only measure of out-of-school science learning. Unfortunately, the response categories do not directly ask about the frequency in time of doing out-of-school science activities.

The responses to these questions shown at the bottom of Exhibit 3-1 indicate how strongly students agree with the above three questions. Agreement is a sign that students' like and enjoy science and hence are apt to be engaged positively with science.

The two additional questions at the bottom of the list (Exhibit 3-1) deal with two primary reasons students take science.

- Take science because it will help in the future.
- Take science because required.

Yes answers to these questions suggest that the motivation to take science is not necessarily because of an intrinsic like for science, but because of some external requirement or reward. The following data analyses will compare the responses to the first three questions that reflect liking science with the last two questions that reflect taking science for reasons other than personal satisfaction with the subject.

**NAEP Background Variable Data on Student Engagement in Science**

The analyses first look at the pattern of association nationally of student responses to the five attitude and behavioral questions with student science assessment scores. These data are correlation data that show the strength of the association in the U.S. across students in grade 8 science, but by themselves do not prove causation. The research evidence supporting causation was discussed above. Following the analyses for all students, the analyses then breakout the distribution of student attitudes and behaviors toward science by gender, low-income and racial-ethnic student sub-groups.

**Student engagement toward science and science assessment scores**

Questions 1-3 in Exhibit 3-2 reflect students' positive attitudes/engagement toward science. Strongly disagree means the students do not hold positive attitudes/behaviors toward science. Strongly agree is a sign of positive attitudes/behaviors. The student responses show:

- For each of questions 1-3, there is a strong positive improvement in students’ NAEP science assessment scores, as student responses move from strongly disagree to strongly agree.
Exhibit 3-2. Average NAEP science assessment scores by student responses to questions about engagement in science for grade 8, 2011

**NAEP science scores by grade 8 students’ attitudes and engagement in science: 2011**

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
<th>Strongly agree minus strongly disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The student likes science</td>
<td>136</td>
<td>144</td>
<td>155</td>
<td>166</td>
<td>30</td>
</tr>
<tr>
<td>2. Student does science activities that are not for schoolwork</td>
<td>146</td>
<td>153</td>
<td>157</td>
<td>162</td>
<td>16</td>
</tr>
<tr>
<td>3. Science is a favorite subject</td>
<td>141</td>
<td>149</td>
<td>150</td>
<td>166</td>
<td>25</td>
</tr>
<tr>
<td>4. Take science because it will help in future</td>
<td>154</td>
<td>161</td>
<td>150</td>
<td>143</td>
<td>-11</td>
</tr>
<tr>
<td>5. Take science because required</td>
<td>164</td>
<td>160</td>
<td>146</td>
<td>138</td>
<td>-26</td>
</tr>
</tbody>
</table>

NCES, NAEP Data Explorer

- The differences in the science achievement scores range from 30 points for students liking science activities to 16 points for students doing science activities. In judging differences in student assessment scores, the 30-point differential is about the same as the difference between the cut-scores for the Basic and Proficient achievement levels. The 16-point differential is about half the difference between the two performance standards.

Questions 4 & 5 explore the reasons for taking science. They ask students about whether they are taking science because it will help them in the future or because it is required. Implicitly, a strong positive answer suggests that students who hold these reasons for taking science take science for external reasons and do not take science primarily because they are engaged in and like science. The pattern of science scores for questions 4 and 5 is different than for questions 1-3.

- For both questions, a strong positive answer that students took science for reasons other than they liked science produced an average lower grade 8 science score.

- The range in score decline between strongly disagree and strongly agree was 11 points (about 30 percent the Basic and Proficient cut-score difference) for students who took science because it would help in the future. It was 26 points (over 85 percent of the Basic and Proficient cut-score difference) for students who took science primarily because it was required.
Exhibit 3-3 shows the same pattern of correlations controlling for family-income status, as measured by the stand-in variable of eligibility for free or low-priced lunch, except for one of the engagement measures.

There are no differences in the direction of the correlations (positive for questions 1, 2, and 3, negative for questions 4 and 5), but the correlation is much weaker among low-income students on question 2. We speculate that this may be because some highly engaging activities (such as amateur astronomy, video-making, or bird watching) require equipment and/or travel, which may be less available to low-income students. Further investigation is required to determine why this particular engagement variable differs from the others so much with respect to income level.
Given the associations between these student attitude and behavioral questions about engagement in science, Exhibit 3-4 shows that the percentage distribution of student responses nationally to these questions. There is considerable variability in the percentage of students who agree or strongly agree when asked about positive science attitudes and behaviors, with a range for agree or strongly agree of:

- 69 percent agree or strongly agree that they like science.
- 29 percent agree or strongly agree that they do science activities that are not for school work.

The following sections examine how student engagement, as measured by responses to these five questions, differs by the gender, low-income or race-ethnicity of a student.

**Gender.** By gender, grade 8 students’ scores on the NAEP science assessment exhibit a modest score advantage of 5 points favoring males. Five points represent slight more than 15 percent of the much larger Basic-Proficient science cut-score difference.
Exhibit 3-5. Percentage of grade 8 students by gender and students’ engagement for science: 2011

Exhibit 3-5 explores students’ engagement in science by gender for the five questions about student engagement. The pattern of responses shows that:

- A greater percentage (between 6 and 11 percentage points) of males than females agree or strongly agree on the first three questions in which positive responses are associated with higher NAEP student achievement scores.
- A greater percentage (between 7 and 8 percentage points) of females agree or strongly agree on the bottom two questions in which positive responses are associated with lower achievement.

Thus, for both groups of questions, males hold attitudes and behaviors more favorably associated with higher achievement than do females.

**Low-income Students.** Students from low-income families identified as those who receive free or reduced price lunch, do not do as well on the grade 8 NAEP science assessments as students who are from non-low-income families, as identified by not receiving free or reduced-price lunch. On average, low-income students’ grade 8 2011 science scores are 27 points below those of non low-income students.
Exhibit 3-6. Percentage of grade 8 students by school-lunch eligibility and students’ engagement in science

<table>
<thead>
<tr>
<th>Question</th>
<th>School lunch eligibility</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
<th>Agree or strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The student likes science</td>
<td>Eligible</td>
<td>14</td>
<td>20</td>
<td>49</td>
<td>17</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>10</td>
<td>18</td>
<td>51</td>
<td>20</td>
<td>71</td>
</tr>
<tr>
<td>2. Student does science activities that are not for schoolwork</td>
<td>Eligible</td>
<td>27</td>
<td>44</td>
<td>25</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>24</td>
<td>47</td>
<td>25</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>3. Science is a favorite subject</td>
<td>Eligible</td>
<td>23</td>
<td>33</td>
<td>29</td>
<td>15</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>19</td>
<td>31</td>
<td>33</td>
<td>17</td>
<td>50</td>
</tr>
<tr>
<td>4. Take science because it will help in future</td>
<td>Eligible</td>
<td>12</td>
<td>27</td>
<td>40</td>
<td>21</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>12</td>
<td>35</td>
<td>38</td>
<td>15</td>
<td>53</td>
</tr>
<tr>
<td>5. Take science because required</td>
<td>Eligible</td>
<td>16</td>
<td>31</td>
<td>32</td>
<td>21</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>20</td>
<td>37</td>
<td>28</td>
<td>15</td>
<td>43</td>
</tr>
</tbody>
</table>

Source: NCES NAEP Data Explore, February 2013

Consistent with their lower science assessment scores, students from low-income families (school-lunch eligible) are less likely to hold attitudes and behaviors that are correlated with higher science achievement (Exhibit 3-6).

- Low-income students are 5-6 percentage points less likely to agree or strongly agree that they like science or that science is a favorite subject. There are no differences in doing science activities that are not for schoolwork.

- Low-income students are more likely to agree or strongly agree (between 8 to 10 percentage points) that they take science because it will help in the future or because it is required. Again, a positive response is associated with lower assessment scores, perhaps because implicit to these answers is that they do not take science primarily because they are engaged in the subject.

**Race/Ethnicity of Students.** On the NAEP grade 8 2011 science assessment, White and Asian science scores are substantially higher than the scores for Black, Hispanic or American Indian students. To illustrate, White student scores are 34 points higher than Black students, 26 points higher than Hispanic students, and 22 points higher than science scores for American Indian students. As with Whites, Asian student science scores are similarly higher than for other minorities.
Exhibit 3-7. Percentage of grade 8 students by race/ethnicity and students’ engagement for science: 2011

<table>
<thead>
<tr>
<th>Question</th>
<th>Race/Ethnicity</th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
<th>Agree or strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Percentage</td>
<td>Percentage</td>
<td></td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
<tr>
<td>1. The student likes science</td>
<td>White</td>
<td>10</td>
<td>19</td>
<td>50</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>16</td>
<td>20</td>
<td>46</td>
<td>19</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>13</td>
<td>21</td>
<td>50</td>
<td>16</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Asian/Pacific Islander</td>
<td>8</td>
<td>17</td>
<td>56</td>
<td>19</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>14</td>
<td>19</td>
<td>52</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>2. Student does science activities that are not for schoolwork</td>
<td>White</td>
<td>25</td>
<td>46</td>
<td>25</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>29</td>
<td>42</td>
<td>25</td>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>26</td>
<td>47</td>
<td>24</td>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Asian/Pacific Islander</td>
<td>21</td>
<td>46</td>
<td>28</td>
<td>5</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>26</td>
<td>46</td>
<td>23</td>
<td>5</td>
<td>28</td>
</tr>
<tr>
<td>3. Science is a favorite subject</td>
<td>White</td>
<td>19</td>
<td>31</td>
<td>33</td>
<td>17</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>26</td>
<td>30</td>
<td>26</td>
<td>17</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>21</td>
<td>35</td>
<td>30</td>
<td>14</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Asian/Pacific Islander</td>
<td>14</td>
<td>34</td>
<td>35</td>
<td>17</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>22</td>
<td>31</td>
<td>34</td>
<td>13</td>
<td>47</td>
</tr>
<tr>
<td>4. Take science because it will help in future</td>
<td>White</td>
<td>13</td>
<td>36</td>
<td>37</td>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>12</td>
<td>22</td>
<td>37</td>
<td>29</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>11</td>
<td>27</td>
<td>42</td>
<td>20</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Asian/Pacific Islander</td>
<td>9</td>
<td>26</td>
<td>42</td>
<td>22</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>10</td>
<td>27</td>
<td>44</td>
<td>18</td>
<td>62</td>
</tr>
<tr>
<td>5. Take science because required</td>
<td>White</td>
<td>20</td>
<td>37</td>
<td>28</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>17</td>
<td>28</td>
<td>30</td>
<td>25</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>15</td>
<td>33</td>
<td>35</td>
<td>18</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>Asian/Pacific Islander</td>
<td>19</td>
<td>37</td>
<td>30</td>
<td>14</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>American</td>
<td>14</td>
<td>29</td>
<td>37</td>
<td>19</td>
<td>56</td>
</tr>
</tbody>
</table>

Source: NCES NAEP Data Explorer, February 2013

As with gender and low-income, with respect to student attitudes and behaviors toward science by race/ethnicity (Exhibit 3-7), the overall pattern of student engagement is again generally consistent with the pattern of relative science scores across racial/ethnic groups. The largest differences in student attitudes and behaviors related to engagement are with respect to Asian students.
Compared with Asian American students, Black, Hispanic and American Indian students are about 8-10 percentage points less likely to agree or strongly agree that they like science; and 5-9 percentage points less likely to indicate that science is a favorite subject. The differential in doing science activities that are not for schoolwork are only 4-5 percentage points higher for Asian students. The patterns for White students are similar but the differences are not as large as with Asian students for questions about liking science or science is a favorite subject. Here is essentially no difference for the question about the student does science activities not for schoolwork.

Black, Hispanic and American Indian students are about 10 percentage points more likely to agree or strongly agree that they take science because it is required than Asian students. However, Asian students, but not White students, are about as likely as Black, Hispanic or American Indian students to take science because it will help them in the future. In the case of Asian Americans, they are more likely to agree or strongly agree with taking science for future benefit.

In summary, overall there are definite patterns of differences in student attitudes and behaviors among students taking grade 8 science nationally and by student subgroups of gender, low-income and race-ethnicity. In most cases, the pattern of responses to these questions about science attitudes and behaviors are correlated in the expected direction with student achievement.

4. Teachers of Science

The National Research Council’s review of teacher preparation concludes “that both strong content knowledge and familiarity with how students learn a particular subject are important for reading, math, and science teachers” (National Research Council, 2010). This section examines NAEP background variables describing teachers of grade 8 science with respect to three characteristics related to teacher preparation and training:

- **What is the proportion of teachers of science who have an undergraduate or graduate major by science area?** Having a degree in science provides evidence of a teacher’s content knowledge in science, although the relevance of a teacher’s knowledge may depend on the recency of a teacher’s degree and the specific area of science the teacher is now assigned to teach.

- **What is the proportion of teachers of science who have taken alternative routes to certification?** Alternative certification is a route more commonly relied upon in K-12 fields of teaching where there are potential shortages of qualified teachers to fill vacancies. The NAEP data provide estimates of alternative certification by science major. They also indicate whether
teachers from alternative certification routes are less experienced in teaching science, addressing a concern that students who enter teaching through alternative certification paths may be less likely to stay in teaching.

- What is the amount and type of teacher professional development in science? Professional development is a major source of teacher improvement, a means to stay current in science topics, and to develop skills in the emerging use of instructional technology in science education.

The analyses primarily focus on the grade 8 questionnaires for 2011. However, we do compare science degrees of grade 8 teachers of science between 1996 and 2011 to determine if there has been progress in exposing students to teachers with content mastery as measured by a science degree.

A. Teachers with Science Degrees

Research. The National Research Council’s Taking Science to School (2007) notes the limited number of sound research studies investigating the relationship between teachers’ postsecondary study of science with their students’ achievement scores. However, the NRC finds that the available evidence does support the importance of science teachers’ content knowledge as positively affecting the STEM achievement of their students. Two of the citations with strongest evidence are:

- “a 1983 meta-analysis (Druva and Anderson, 1983) found a positive relationship between student achievement and the number of science courses their teachers had taken. “

- “Monk (1994) presents data from a longitudinal survey that addressed teacher degrees in relation to both science and mathematics instruction and also identified positive effects.”

In the related STEM area of mathematics, the International Education study (TEDS-M) of elementary and middle-school U.S. teachers of mathematics found that those teachers with a major in mathematics had substantially greater content knowledge that was more comparable to the knowledge of teachers of mathematics in countries with high-performing students (Schmidt, et.al., 2007).

Professional consensus is also clear about teachers of science needing to have strong content knowledge base in science. The National Science Teachers Association (2011) recommends both strong knowledge of science content and knowledge of how students learn. Also, the Council of Chief State School Officers (2012, p15) in developing their new vision for teacher education stresses, "As part of the evidence of a candidate's ability to perform, states will need evidence of a candidate's content knowledge."
Exhibit 4-1. NAEP background questions on science degrees held by teachers of science, 2011

**NAEP Questions about Science Teachers’ Science Degrees.** Exhibit 4-1 displays the NAEP Background questions, which ask teachers of science about their undergraduate and graduate degrees. The focus of this background variable report is on the six questions that ask about having an undergraduate or graduate major or minor related to science. These cover the three science fields of biology or other life science, earth or space science, and physics or chemistry.

One important limitation of the analyses of teachers’ science degrees generated by the NAEP Data Explorer is that it treats each undergraduate or graduate degree in a field as if it came from a separate individual. Undoubtedly, some teachers who received an undergraduate degree in a science field also went on to obtain a graduate degree, thus creating the potential for double counting. Consequently, the numbers in the tables represent an upper limit on the number of grade 8 teachers of science with an undergraduate or graduate degree in one of the three science fields.

**Analyses of Teachers With Science Degrees.** Exhibit 4-2 compares the percentages between 1996 and 2011 of grade 8 students taught by teachers of science who have an undergraduate or graduate major in one of three science areas: biology or other life science, earth or space science, or physics or chemistry.
Exhibit 4-2

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Percent of grade 8 students taught by science teachers who have an undergraduate or graduate science major, by science field, 1996 and 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Biology or Other Life Science</td>
</tr>
<tr>
<td></td>
<td>Undergraduate</td>
</tr>
<tr>
<td>2011</td>
<td>36</td>
</tr>
<tr>
<td>1996</td>
<td>38</td>
</tr>
</tbody>
</table>

Source: NCES, NAEP Data Explorer

All students. Between 1996 and 2011, the percentage of grade 8 students taught by teachers of science who have a science major declined by about 50 percent for teachers in earth/space science or physics/chemistry. The percentages of students taught by a teacher with a biology or life science major held constant between 1996 and 2011. In both time periods, the highest percentages of teachers with a science major were in biology or other life sciences (Exhibit 4-2).

Race/ethnicity (Exhibit 4-3). In 2011, students within different subgroups of race/ethnicity generally had roughly similar percentages of teachers with science majors. However, in 1996 a significantly lower percentage of Black students compared with Whites had access to teachers with a science major, while Asian and Hispanic students had greater access compared with White students. However, the greater equality in 2011 compared with 1996 was achieved by reductions for groups with greater access to science teachers in 1996 rather than by increases in the percentages of the groups with lower access.

- In 2011, all racial/ethnic groups were within 10 percentage points of White students in access to teachers with a science major. Black and American Indian students had somewhat lower access and Asians somewhat greater access with Hispanics similar to White students.

- This contrasts with 1996, in which Black students had less access to teachers with a science major than whites by 21 percentage points, Asians 20 percentage points greater access, American Indians 14 percentage point greater access, and Hispanics 9 percentage points greater access. However the improved Black equality in 2011 occurred through reducing the 2011 percentages of teachers with a science major for those groups with higher percentages of such teachers in 1996 rather than by increasing 2011 Black access to teachers with a science major.
Exhibit 4-3

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Biology or Other Life Science</th>
<th>Earth or Space Science</th>
<th>Physics or Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Undergraduate</td>
<td>Graduate</td>
<td>Undergraduate</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>36</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Black</td>
<td>39</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Hispanic</td>
<td>38</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>41</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>32</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>1996</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>38</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>Black</td>
<td>33</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Hispanic</td>
<td>44</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>54</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>56</td>
<td>27</td>
<td>11</td>
</tr>
</tbody>
</table>

Source: NCES, NAEP Data Explorer

Gender (Exhibit 4-4). There were no systematic differences between the percentages of grade 8 male and female students who were taught science by a teacher who majored in science, by field in either 1996 nor 2011.

Exhibit 4-4

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Biology or Other Life Science</th>
<th>Earth or Space Science</th>
<th>Physics or Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Undergraduate</td>
<td>Graduate</td>
<td>Undergraduate</td>
</tr>
<tr>
<td>2011 Male</td>
<td>36</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Female</td>
<td>37</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>1996 Male</td>
<td>39</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Female</td>
<td>36</td>
<td>10</td>
<td>18</td>
</tr>
</tbody>
</table>

Source: NCES, NAEP Data Explorer
Low-income (Exhibit 4-5). There were no major differences between the percentages of grade 8 low-income and non low-income students who were taught science by a teacher who majored in science, by field, in either 1996 nor 2011.

Exhibit 4-5

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Biology or Other Life Science</th>
<th>Earth or Space Science</th>
<th>Physics or Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Undergraduate</td>
<td>Graduate</td>
<td>Undergraduate</td>
</tr>
<tr>
<td>2011 Eligible</td>
<td>36</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>2011 Ineligible</td>
<td>37</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>1996 Eligible</td>
<td>36</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>1996 Ineligible</td>
<td>39</td>
<td>11</td>
<td>18</td>
</tr>
</tbody>
</table>

Source: NCES, NAEP Data Explorer

B. Alternative Certification of Teachers of Science

Research and Background. There are several reasons why schools may hire teachers who have taken alternative routes to become a teacher. One is to fill positions for which available candidates are in short supply. Another reason is to increase the quality of teachers by drawing upon a broader group of graduates interested in teaching than only those who have matriculated through education school. Both trends are a widespread concern in science.

One well-known example of an alternative certification program is Teach for America, a program that is designed to draw into teaching graduating students from the upper academic tier of colleges and universities. These are students who are less likely to enter teaching through traditional education school preparation.

The evaluation evidence on the test scores of students of teachers entering through alternative certification routes is that their students’ scores are about the same as regular teachers. (Constantine, et.al., 2004). There is concern, however, that these teachers are less likely to stay in education (Fowler, 2003). Higher turnover would raise the likelihood of students having a less experienced teacher, a concern as research consistently shows that students of beginning teachers (first three years) have lower performance.

Questions and Analytic Approach. NAEP in 2011 asks a single question of grade 8 teachers of science,
“Did you enter teaching through an alternative certification program”
Teacher-reported responses are: Yes, No.

The analyses address the following questions:

- How does the percentage of students taught by grade 8 teachers of science with alternative certification routes compare with the percentage taught by grade 8 teachers in non-STEM subjects? This analysis compares teachers of reading with teachers of science.

- Given the potential greater difficulty in finding teachers of science for schools serving low-income or minority students, are schools with higher proportions of low-income or minority students more likely to hire teachers who have gone through alternative routes to teaching?

- Is there evidence that science teachers of students from low-income or minority families have less experience and if so could this be related to higher rates of science teachers with an alternative certificate?

**Exhibit 4-6**

<table>
<thead>
<tr>
<th>Percent of Grade 8 students taught science and reading by teachers entering teaching through alternative certification: all students, race/ethnicity &amp; low-income, 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Group</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>National</td>
</tr>
<tr>
<td>Race/ethnicity</td>
</tr>
<tr>
<td>• White</td>
</tr>
<tr>
<td>• Black</td>
</tr>
<tr>
<td>• Hispanic</td>
</tr>
<tr>
<td>• Asian/Pacific Islander</td>
</tr>
<tr>
<td>• American Indian/Alaska Native</td>
</tr>
<tr>
<td>Low-income (school lunch eligible)</td>
</tr>
<tr>
<td>• Eligible</td>
</tr>
<tr>
<td>• Ineligible</td>
</tr>
</tbody>
</table>

Source: NAEP Data Explorer, 2013
Data Analyses and Findings. Exhibit 4-6 shows the percentage of grade 8 students taught by teachers who entered teaching through an alternative certification route in science compared with reading. Overall, the percentage of grade 8 students' taught science by teachers who entered teaching through an alternative certification route was higher than for reading. Black students and to some extent Hispanic students were more likely to be taught by teachers who entered through alternative certification than White or Asian students. Similarly students from low-income families are more likely to be taught science by a teacher entering through alternative certification than for non-low-income students.

- Nationally, 20 percent of grade 8 students had teachers of science who entered through alternative certification, 5 percentage points higher than for students taught by teachers of reading.

- By race/ethnicity, 30 percent of grade 8 Black students and 25 percent of Hispanic students had science teachers who were alternatively certified compared with only 16 percent for White or Asian students. Interestingly, teachers of American Indian students are no more likely to go through alternative certification routes. Use of alternative certification in reading is also greater for teachers of Black or Hispanic students, but the rates are not nearly as high as for teachers of science.

- Grade 8 students from low-income families were 6 percentage points more likely to have science teachers who entered through alternative certification than for non-low-income students.

Exhibit 4-7
Percentage of alternative certification and non alternative certification teachers who majored in science as under graduates or graduates by field, 2011

<table>
<thead>
<tr>
<th></th>
<th>Biology &amp; Other life Science</th>
<th>Earth or Space Science</th>
<th>Physics or Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduate</td>
<td>Undergraduate</td>
<td>Undergraduate</td>
<td>Undergraduate</td>
</tr>
<tr>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
<td>Percent</td>
</tr>
<tr>
<td>Alternative Certification</td>
<td>51</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>No Alternative Certification</td>
<td>33</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Source: NAEP Data Explorer

The NAEP data also yield information about whether schools are more likely to hire teachers who have a major or minor in science by hiring alternative certification teachers for grade 8 science. Exhibit 4-7 shows that alternative certification teachers are more likely to have a majored in science, but only with respect to biology and life sciences.
• Considering grade 8 entrants through alternative certification routes, 51 percent of the teachers had an undergraduate major in biology or other life science, which is 18 percentage points higher than the 33 percent who entered teaching through other than alternative certification.

• The incidence of science majors for earth/space science or physics/chemistry is similar between alternative certification and non-alternative certification teachers.

In terms of years of experience teaching science, 26 percent of the grade 8 students had teachers who have taught science four or fewer years (Exhibit 4-8).

**Exhibit 4-8**

<table>
<thead>
<tr>
<th>Teachers of grade 8 science by years of experience for all students, students taught by alternative certification teachers, students by race/ethnicity and low-income: 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>All students</td>
</tr>
<tr>
<td>----------------------------------</td>
</tr>
<tr>
<td>Percentage</td>
</tr>
<tr>
<td>All students</td>
</tr>
<tr>
<td>Alternative Certification</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Hispanic</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
</tr>
<tr>
<td>Low-income (School-lunch Eligible)</td>
</tr>
<tr>
<td>Not eligible</td>
</tr>
<tr>
<td>Source: NAEP Data Explorer</td>
</tr>
</tbody>
</table>

• Teachers with the least science teaching experience are more likely to have entered teaching through alternative certification paths.

• Among students taught science by an alternatively certified teacher, 39 percent of the students were taught science by an alternatively certified teacher with 4 or fewer years of experience. This compared with only 22 percent of students taught science by a teacher with 4 or fewer years of experience when that teacher entered teaching through a traditional route. These numbers suggest merit in a frequent criticism of alternative
certification programs, such as Teach for America, in that teachers are less likely to stay in the system. However, because alternatively certified teachers may also tend to be disproportionately new hires, these numbers need to continue to be tracked over time.

- The least experienced teachers of science are also more likely to teach Black, Hispanic or low-income students. Blacks, Hispanic and American Indian students have 31 or 32 percent of their teachers falling in the 0 to 4 years range, about 10 percentage points more than for White or Asian students.
- Low-income students are about 7 percentage points more likely to have teachers in the 0-4 years experience range than non low-income students.

As noted above, students who are Black or Hispanic or low-income are more likely to be taught by an alternatively certified teacher, which may explain why these students are being taught, on average, by a less experienced teacher.

C. Teachers of Science Professional Experiences

Research and Background. Research on professional development consistently finds that elementary and secondary teachers of all subjects are exposed to less than optimal professional development practices. Professional development is typically delivered through one-time external workshops that often lack integration with a teacher’s regular classroom preparation and teaching. Consequently, research finds professional development as typically delivered often had little impact on changing teacher practice. (Sawchuck, Nov 10, 2010)

The Federal No Child Left Behind Act of 2001 recognized that the problems with professional development were primarily funding pullout and short term teacher workshops. The act responded to this concern by defining professional development eligible for funding to include: activities that "are not one-day or short-term workshops or conferences." However, data from federal evaluations of federally supported activities still continue to show that short-term workshops are the dominant form of teacher professional undertakings” (PPSS, 2009)

The National Academy of Sciences National Science Education Standards (http://www.nas.edu/rise/backg4.htm) are helpful in describing appropriate professional development that is ongoing and integrated:

“Professional development for teachers should be analogous to professional development for other professionals. Becoming an effective science teacher is a continuous process that stretches from pre-service experiences in undergraduate years to the end of a professional career. Science has a rapidly changing knowledge base and expanding relevance to societal issues, and teachers will need ongoing opportunities to build their understanding and ability. Teachers also must have opportunities to develop understanding of
how students with diverse interest, abilities, and experiences make sense of scientific ideas and what a teacher does to support and guide all students. And teachers require the opportunity to study and engage in research on science teaching and learning, and to share with colleagues what they have learned. “(NAS p. 55)

The next section examines how the NAEP questionnaires inform about professional development in light of research about the type of professional opportunities teachers of science need.

**NAEP Background Questions and Analyses About Teacher Professional Development.** Unfortunately, the NAEP questions of teachers of science are limited, asking only about the emphasis on professional development topics. *Missing entirely are questions about the amount of time for professional learning, the time available to develop lessons and improve, the integration of staff development into school activities, teachers perception of the value of their professional development and whether they changed their practice in response to their professional development activities.*

The questions of professional learning follow a similar format in their focus on professional development content and ask:

“Consider all of the professional development activities you participated in during the last two years. To what extent did you learn about:

- Instructional methods for teaching science
- Instructional methods for teaching technological design
- Learned about content standards in science
- Learned about curricular materials in science
- Learned about effective use of ICT in science
- Learned about effective use of lab activities in science
- learned about how students learn science
- Learned about methods for assessing in science
- Learned about preparing students for district or state assessments
- Learned about scientific inquiry and technological design
- learned about teaching science to students from diverse backgrounds”

The NAEP background questions typically cover answers of “not at all,” “small extent,” “moderate extent,” and “large extent.” No information is available on the actual amount of time teachers spent.

**Analyses and Findings about science professional development.** The analyses show that grade 8 professional development for teachers of science emphasizes the core instructional topics (Exhibit 4-9). The topics with the highest percentage of teachers
indicating they learn about to a moderate or a great extent are:

- Content standards in science: 72% to a moderate or a great extent.
- Instructional methods for teaching science: 63% to a moderate or a great extent.
- Scientific inquiry and technological design: 62% to a moderate or a great extent.

The topics receiving least emphasis in terms of professional learning were in the technology area:

- Instructional methods for teaching technological design: 29% to a moderate or a great extent.
- Effective use of ICT in science: 30% to a moderate or a great extent.

The relatively low-emphasis given to developing professional skills in using technology may be a concern for limiting improvement because technology, especially in science instruction, offers the potential to radically change the delivery of science education through such means as computerized science simulations.

**Example 4-9**

<table>
<thead>
<tr>
<th>Grade 8 teacher responses to questions asked about their professional development activities over the last two years, 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
</tr>
<tr>
<td>----------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Learned about instructional methods for teaching science</td>
</tr>
<tr>
<td>Learned about instructional methods for teaching technological design</td>
</tr>
<tr>
<td>Learned about content standards in science</td>
</tr>
<tr>
<td>Learned about curricular materials in science</td>
</tr>
<tr>
<td>Learned about effective use of ICT in science</td>
</tr>
<tr>
<td>Learned about effective use of lab activities in science</td>
</tr>
<tr>
<td>learned about how students learn science</td>
</tr>
<tr>
<td>Learned about methods for assessing in science</td>
</tr>
<tr>
<td>Learned about preparing students for district or state assessments</td>
</tr>
<tr>
<td>Learned about scientific inquiry and technological design</td>
</tr>
<tr>
<td>Learned about teaching science to students from diverse backgrounds</td>
</tr>
<tr>
<td>Participated in activities associated with school improvement efforts directed at issues such as adequate yearly progress and state accountability standards?</td>
</tr>
</tbody>
</table>

Source: NAEP Data Explorer
5. Science Physical Resources: Availability and Use

Quality physical resources for science and the use of these resources to support students’ science learning are an important part of implementing effective science instruction. The middle school science years represent an important period where students are transitioning to more rigorous science content including seeing and doing in-depth science experiments. Students’ middle school science activities need to be supported with adequate books, instructional materials, audio-visual equipment, access to computers and laboratory opportunities.

The National Science Teachers Association (http://www.nsta.org/about/positions/highschool.aspx) has established criteria covering an adequate science program including laboratory work for high schools, that is also relevant for middle schools. These criteria cover:

“Science rooms/laboratories should be used only for science classes and science activities and should be equipped with:

- Adequate laboratory space per student and sufficient gas, electrical, and water outlets for student laboratory activities
- Safety equipment, such as fire extinguisher, fume hoods, emergency showers, and eyewash stations
- Audiovisual equipment such as an overhead projector; videocassette recorder and monitor; slide projector; and one or more computers with Internet access, plus needed software and maintenance service
- Sufficient storage for equipment and supplies and preparation space close to the classroom
- Support equipment such as photocopying machines, typewriters, word processors, and telephone in a nearby and accessible area
- Textbooks for each student, laboratory guides, and references as appropriate and needed.”

The next section examines the questions NAEP asks about the physical resources to teach science.

NAEP Science Resource Background Questions

The NAEP grade 8 science questionnaires cover the essential resource categories of safety, facilities, science-specific equipment and access to general resources. The questions are of two types. One set of questions asks about the availability of these
different science resources and a second set asks about use for instruction of many of these same resources.

The background question responses are from three sources.

- Schools responded to questions about science resource availability of different types on a school-wide basis.
- Teachers responded to questions about both resources and use in their science instruction classes.
- Student responses provide responses from their personal respective about use of science resources.

The responses differ in their quantitative/qualitative nature. An example of those involving a more quantitative response is shown in Exhibit 5-1. The questions ask specifically about the percent of classrooms with access to specific science resources; in this example, handheld devices and tablet PCs.

### Exhibit 5-1. Quantitative questions about science resources

<table>
<thead>
<tr>
<th>Type of Science Resource</th>
<th>Percent of Classrooms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>percent classrooms with handheld device for 8th grade science</td>
<td>11</td>
</tr>
<tr>
<td>percent classrooms with tablet PC for 8th grade science</td>
<td>70</td>
</tr>
</tbody>
</table>

Another type of question is qualitative and asks about relative resource availability in terms of small, moderate or large extent along with “not at all.” An example of these questions is shown in Exhibit 5-2. The advantage of this format is that it is easy for the respondent to answer. However, the disadvantage is that respondents might differ in interpreting terms such as small, moderate and large. For example, respondents in schools that are generally resource rich may have a high standard for moderately available resources than respondents who are in generally resource poor schools.

### Exhibit 5-2. Qualitative questions about science resources

<table>
<thead>
<tr>
<th>Type of Science Resource (School or Teacher Reported)</th>
<th>Not at all</th>
<th>Small extent</th>
<th>Moderate extent</th>
<th>Large extent</th>
<th>Percentage</th>
</tr>
</thead>
</table>
| computerized science labs are provided for science (teacher reported) | 60 | 19 | 14 | 7 | 67%
| have computerized science labs (school reported) | 42 | 25 | 15 | 18 | 65% |
Concerns about measurement error also occur from different patterns of responses to similar questions by schools, teachers and students. In general, the pattern of responses indicates that schools are likely to say they provide more science resources than teachers are. As illustrations (Exhibit 5-3):

- With respect to the availability of science kits, 26 percent of teachers report “none at all” available compared with 7 percent of the school responses.

- With respect to science magazines and books, 22 percent of the teachers report “none at all” compared with 2 percent of the school responses.

This analysis will report both teacher and school responses, but will generally use the teacher responses in the analyses for several reasons. First, schools have a psychological incentive to boost responses to resource availability, which are a school responsibility, while teachers do not. Second, teacher responses are more fine-grained so that a school might respond they have science kits even if only one teacher or classroom does.

### Science Labs

*The overall findings with respect to the characteristics of science labs at grade 8 are that they typically have many of the laboratory essentials, but that many labs are lacking more sophisticated equipment (gas for burners) or newer technology (computerized science labs).*

Exhibit 5-4 presents the findings for the NAEP background questions on science lab resource availability. The responses within each type of question group are rank ordered from least to most science resource availability. The ranking is based on the response to the first column of “not at all” in the table.
Most but not all science labs serving grade 8 have the essentials. Looking at the percentage responding "not at all" or "a small extent," the responses were only:
- 5 percent for safety equipment
- 5 percent for running water
- 8 percent for supply storage
- 8 percent for internet connections
- 12 percent for demonstration stations

The science lab resource weaknesses at grade 8 are indicated by a relatively high percentage of "not at all" or "a small extent" responses in the area of computers and certain equipment:
- 79 percent for computerized science labs
- 54 percent for air hoses
- 39 percent for gas burners

<table>
<thead>
<tr>
<th>Exhibit 5-4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Science lab resource availability, school and teacher reported, grade 8, 201</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>computerized science labs are provided for science (teacher reported)</td>
</tr>
<tr>
<td>have computerized science labs (school reported)</td>
</tr>
<tr>
<td>science labs have hood or air hoses</td>
</tr>
<tr>
<td>science labs have gas for burners</td>
</tr>
<tr>
<td>science labs have computers (school reported)</td>
</tr>
<tr>
<td>science labs have student lab stations</td>
</tr>
<tr>
<td>space to conduct science labs is provided</td>
</tr>
<tr>
<td>lab supplies are provided for science</td>
</tr>
<tr>
<td>science labs have demonstration stations</td>
</tr>
<tr>
<td>science labs have Internet connection</td>
</tr>
<tr>
<td>have supplies for science labs</td>
</tr>
<tr>
<td>science labs have supply storage</td>
</tr>
<tr>
<td>science labs have running water</td>
</tr>
<tr>
<td>science labs have safety equipment</td>
</tr>
<tr>
<td>science labs have electricity</td>
</tr>
</tbody>
</table>

Source: NAEP Data Explorer
Instructional Technology for Science in Classrooms.

Exhibit 5-5 shows school-reported availability of instructional technology for science by percent of classrooms having a particular instructional technology resource.

- Some instructional technology items are quite common for grade 8 science. At least 75 percent of the schools have half or more of their grade 8 science classrooms equipped with digital projectors, CD/ROM, online software and a desktop computer.

- The least common items for which 75 percent or more of the schools have less than half their grade 8 science classrooms equipped with instructional technology include handheld devices, tablet PC’s, digital music,.

- Examples of mid-range items in terms of limited availability are graphing calculator, which less than half the science classrooms have available in 54 percent of the schools and cable or satellite TV with less than half the science classrooms having available 41 percent of the schools.

**Exhibit 5-5**

| Instructional technology for science in classrooms (school reported), grade 8, 2011 | Percent of Classrooms |
|---|---|---|---|---|---|---|
| | 0% | 1-25% | 26-50% | 51-75% | 76-99% | 100% |
| Type of Science Resource | Percentage | Percentage | Percentage | Percentage | Percentage | Percentage |
| percent classrooms with handheld device for 8th grade science | 70 | 18 | 4 | 2 | 1 | 4 |
| percent classrooms with tablet PC for 8th grade science | 70 | 15 | 4 | 2 | 2 | 7 |
| percent classrooms with digital music device for 8th grade science | 56 | 19 | 6 | 4 | 4 | 12 |
| percent classrooms with data collection sensor for 8th grade science | 46 | 23 | 9 | 6 | 4 | 13 |
| percent classrooms with cable or satellite TV for 8th grade science | 33 | 6 | 2 | 2 | 5 | 52 |
| percent classrooms with online course management for 8th grade science | 27 | 9 | 4 | 3 | 6 | 52 |
| percent classrooms with digital whiteboard for 8th grade science | 26 | 17 | 10 | 7 | 7 | 34 |
| percent classrooms with digital camera for 8th grade science | 21 | 31 | 12 | 7 | 4 | 26 |
| percent classrooms with graphing calculator for 8th grade science | 21 | 20 | 13 | 10 | 7 | 29 |
| percent classrooms with laptop computer for 8th grade science | 21 | 19 | 8 | 7 | 7 | 38 |
| percent classrooms with desktop computer for 8th grade science | 19 | 9 | 2 | 2 | 6 | 72 |
| percent classrooms with online software for 8th grade science | 8 | 7 | 4 | 5 | 10 | 64 |
| percent classrooms with CD/ROM for 8th grade science | 5 | 4 | 3 | 3 | 9 | 75 |
| percent classrooms with digital projector for 8th grade science | 3 | 6 | 4 | 6 | 12 | 70 |

Source: NAEP Data Explorer, Science 2011
Science non-technology instructional materials in classrooms

These include measuring instruments, books and audiovisual materials (Exhibit 5-6).

- The most widely available instructional technology items with less than 10 percent of the respondents reporting “not at all” or a “small extent” for availability include science textbooks (3 percent), scientific measurement instruments (8 percent) and supplies for science demonstrations (6 percent).

- The least available science non-technology resources are science kits and science magazines and books, with more than half the teachers reporting availability “not at all” or “small extent.” The scarcity of these resources may indicate some lack of depth in conducting hands-on science activities without access to specialized kits or scientific reading material.

<table>
<thead>
<tr>
<th>Exhibit 5–6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classroom Access to Non-instructional Technology Resources in Classrooms when Teaching Grade 8 Science</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Science Resource (School or Teacher Reported)</th>
<th>Not at all</th>
<th>Small extent</th>
<th>Moderate extent</th>
<th>Large extent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
<tr>
<td>science kits are provided for science (Teacher reported)</td>
<td>24</td>
<td>30</td>
<td>29</td>
<td>16</td>
</tr>
<tr>
<td>science magazines and books are provided for science (Teacher reported)</td>
<td>22</td>
<td>35</td>
<td>33</td>
<td>11</td>
</tr>
<tr>
<td>have science kits (School reported)</td>
<td>7</td>
<td>24</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>audiovisual materials are provided for science (Teacher reported)</td>
<td>6</td>
<td>21</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td>supplies for demonstrations are provided for science (Teacher reported)</td>
<td>3</td>
<td>18</td>
<td>43</td>
<td>36</td>
</tr>
<tr>
<td>have science magazines and books (school reported)</td>
<td>2</td>
<td>19</td>
<td>35</td>
<td>44</td>
</tr>
<tr>
<td>textbooks are provided for science (Teacher reported)</td>
<td>2</td>
<td>6</td>
<td>21</td>
<td>72</td>
</tr>
<tr>
<td>scientific measurement instruments are provided (Teacher reported)</td>
<td>2</td>
<td>16</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td>have science textbooks (Teacher reported)</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>90</td>
</tr>
<tr>
<td>have scientific measurement instruments (school reported)</td>
<td>0</td>
<td>8</td>
<td>31</td>
<td>61</td>
</tr>
<tr>
<td>have supplies for science demonstrations (Teacher reported)</td>
<td>0</td>
<td>6</td>
<td>26</td>
<td>65</td>
</tr>
</tbody>
</table>

*Source: NAEP Data Explorer, Science 2011*
Use of available science resources

Exhibit 5-7. Science resource use: teacher and student reported, grade 8, 2011

<table>
<thead>
<tr>
<th>Type of Science Resource</th>
<th>Not at all</th>
<th>Small extent</th>
<th>Moderate extent</th>
<th>Large extent</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Reported</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>use handheld device for science</td>
<td>86</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>use tablet PC for science</td>
<td>83</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>use graphing calculator for science</td>
<td>77</td>
<td>17</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>use digital music device for science</td>
<td>75</td>
<td>17</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>use cable or satellite TV for 8th grade science</td>
<td>69</td>
<td>21</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>use data collection sensor for science</td>
<td>65</td>
<td>25</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>use digital whiteboard for science</td>
<td>51</td>
<td>9</td>
<td>10</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>use digital camera for science</td>
<td>48</td>
<td>37</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>use online course management for science</td>
<td>34</td>
<td>15</td>
<td>16</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>use laptop computer for science</td>
<td>21</td>
<td>16</td>
<td>17</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>use CD/ROM for 8th grade science</td>
<td>25</td>
<td>37</td>
<td>25</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>use online software for science</td>
<td>21</td>
<td>34</td>
<td>29</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>use desktop computer for science</td>
<td>20</td>
<td>18</td>
<td>20</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>use DVDs and player for science</td>
<td>13</td>
<td>44</td>
<td>32</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>use digital projector for science</td>
<td>7</td>
<td>7</td>
<td>18</td>
<td>69</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher Reported</th>
<th>Never or hardly ever</th>
<th>Once or twice a month</th>
<th>Once or twice a week</th>
<th>Every day or almost every</th>
</tr>
</thead>
<tbody>
<tr>
<td>students science book or magazine</td>
<td>43</td>
<td>44</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>students use computer to chart or graph science projects</td>
<td>39</td>
<td>43</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>students use computer to simulate physical or biological process</td>
<td>38</td>
<td>43</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>students use computer to search for science information</td>
<td>30</td>
<td>53</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>students read science textbook</td>
<td>11</td>
<td>21</td>
<td>45</td>
<td>24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Reported</th>
<th>Never or hardly ever</th>
<th>Once every few weeks</th>
<th>About once a week</th>
<th>2-3 times a week</th>
<th>Every day or almost every</th>
</tr>
</thead>
<tbody>
<tr>
<td>use library resources for science</td>
<td>67</td>
<td>21</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>use computers for science</td>
<td>40</td>
<td>36</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>watch teacher do science experiment</td>
<td>24</td>
<td>35</td>
<td>24</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>watch science movie, video or DVD</td>
<td>20</td>
<td>42</td>
<td>22</td>
<td>12</td>
<td>5</td>
</tr>
</tbody>
</table>

Source: NAEP Data Explorer, Science 2011

Having science resources available does not necessarily translate into science resources being used for instruction. The NAEP background variables also ask explicitly about frequency of use of the science resources by teachers and students. Exhibit 5-7 organizes the responses into three tables by whether the information about use is teacher or student reported and for teacher reported into the two
different response formats. Based on the responses, resource use falls into three categories: commonly used, uneven use and uncommon use.

- **Science resources commonly used for science instruction are the familiar and traditional activities.** The four with the lowest percentages (20 percent or less) of “not at all” or “never or hardly ever” used are **the teachers use DVD and player for science, teachers use a digital projector for science, teachers use a desk top computer for science, and students read science textbook.**

- **Science resources in the mid-range (21-50 percent) of use** based on percentages of “not at all” or “never or hardly ever” include students use a computer to: search for science information (30 percent); simulate a physical or biological process (38 percent); or use a computer to chart a graph (39 percent). Perhaps surprisingly, also not having regular use is students reading a science book or magazine. About 43 percent of the teachers indicate that students “never or hardly ever” read a science book or magazine.

- **In terms of least use of science resources,** over 50 percent of the teachers report not using more recent technology including handheld devices, PC tablets, graphing calculators, digital music devices, cable or satellite TV, data collection sensors or whiteboards. As noted, access to these newer technology resources is limited, so these use findings are expected. These findings should not be interpreted as necessarily meaning that technological resources would not be used if there were greater availability.

And as noted in Introduction section of this report on the list of current topics under discussion among science educators, how such science resources are used is at least as important as the quantity and type of resources available. Computers and lab equipment used only to confirm accepted knowledge, or for drill and practice, may be associated with very different outcomes than using those resources in inquiry pedagogies.

**Equity in distribution of science resources**

Another resource issue is whether science education resources are equitably distributed across populations. Exhibit 5-8 compares resource availability for students by their school lunch eligibility. The measure of resource availability is the percent of classrooms in grade 8 science with the designated resource. The percent measure is preferred to a subjective relative measure, such as small, moderate or large extent, as schools with fewer overall economic resources may have a lower standard of meaning for terms such as small, moderate or large extent.

Overall, school lunch eligible students have slightly less access to science resources. For example, comparing the average percentage of schools with a hundred percent of their classrooms having access to particular resources, there is a 4 percentage
point advantage for non-school lunch eligible students (43 percent) compared with school lunch eligible students (39 percent) across all the resource categories listed.

### Exhibit 5-8. Availability of specific science resources by students' school lunch eligibility and percent of classrooms with resources, grade 8, 2011

<table>
<thead>
<tr>
<th>Type of Resources</th>
<th>0%</th>
<th>1-25%</th>
<th>26-50%</th>
<th>51-75%</th>
<th>76-99%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of all resources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eligible</td>
<td>28.6</td>
<td>14.3</td>
<td>6.1</td>
<td>4.9</td>
<td>7.1</td>
<td>39.2</td>
</tr>
<tr>
<td>Not eligible</td>
<td>26.7</td>
<td>13.7</td>
<td>6.1</td>
<td>4.6</td>
<td>5.7</td>
<td>42.2</td>
</tr>
<tr>
<td>percent classrooms with cable or satellite TV for 8th grade science</td>
<td>Eligible</td>
<td>35</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>29</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>percent classrooms with CD-ROM for 8th grade science</td>
<td>Eligible</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>percent classrooms with data collection sensor for 8th grade science</td>
<td>Eligible</td>
<td>48</td>
<td>23</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>43</td>
<td>23</td>
<td>10</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>percent classrooms with desktop computer for 8th grade science</td>
<td>Eligible</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>9</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>percent classrooms with digital camera for 8th grade science</td>
<td>Eligible</td>
<td>22</td>
<td>33</td>
<td>12</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>19</td>
<td>31</td>
<td>12</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>percent classrooms with digital music device for 8th grade science</td>
<td>Eligible</td>
<td>57</td>
<td>19</td>
<td>5</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>54</td>
<td>20</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>percent classrooms with digital projector for 8th grade science</td>
<td>Eligible</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>percent classrooms with digital whiteboard for 8th grade science</td>
<td>Eligible</td>
<td>25</td>
<td>17</td>
<td>10</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>25</td>
<td>17</td>
<td>10</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>percent classrooms with DVDs and player for 8th grade science</td>
<td>Eligible</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>percent classrooms with graphing calculator for 8th grade science</td>
<td>Eligible</td>
<td>22</td>
<td>21</td>
<td>13</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>20</td>
<td>19</td>
<td>14</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>percent classrooms with handheld device for 8th grade science</td>
<td>Eligible</td>
<td>71</td>
<td>18</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>70</td>
<td>18</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>percent classrooms with laptop computer for 8th grade science</td>
<td>Eligible</td>
<td>19</td>
<td>18</td>
<td>9</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>22</td>
<td>19</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>percent classrooms with online course management for 8th grade science</td>
<td>Eligible</td>
<td>29</td>
<td>19</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>24</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>percent classrooms with online software for 8th grade science</td>
<td>Eligible</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>percent classrooms with tablet PC for 8th grade science</td>
<td>Eligible</td>
<td>23</td>
<td>15</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Not eligible</td>
<td>63</td>
<td>15</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Source: NAEP Data Explorer
6. Curriculum and Instruction

Research and Context

Science curriculum and instruction, like mathematics and reading, is undergoing extensive review and will likely see major future changes as a result of nationwide efforts to improve and strengthen science education standards in the U.S. Although the U.S. has no official science standards, the standards prepared by the National Research Council (1996) have become the basis for many of the state frameworks.

A national coalition of organizations involved in science education (The National Research Council, the National Science Teachers Association, the American Association for the Advancement of Science, and Achieve) have come together to develop the “Next Generation Science Standards.” The Next Generation Science Standards are based on the Framework for K–12 Science Education (National Research Council, 2012) developed by the NRC. The major weaknesses in current standards and instruction that are cited to be addressed by the new standards include (http://www.nextgenscience.org/development-overview):

- Teaching too many science content topics at each grade.
- The science content that is taught tends to be shallow and lacks depth at providing students a real understanding of how things work. Instead, much of science teaching is about memorization. At least in part, this is the result of having to cover so many science content topics.
- The standards and curriculum lack coherence over the grades. That is, they do not build on prior knowledge but instead present fragmented learning from year to year.
- A lack of meaningful teaching of scientific and engineering practices about how scientific knowledge is acquired and applied. Scientific inquiry is often taught as a mechanical, isolated activity and disconnected with major cross-cutting concepts and disciplinary ideas.

The new “Next Generational Science Standards” (Achieve, 2013) offer a potential opportunity to provide a framework to build greater coherence, focus, and depth of understanding in science instruction and curriculum. Combined with the potential advances in teaching and learning from instructional technology, they also offer an opportunity to get across the “integration of scientific explanations and practices needed to engage in scientific inquiry and engineering design.”

The NAEP science background variables around curriculum and instruction offer an opportunity to establish a baseline and chart progress and challenges as new
standards and practices are being implemented. The NAEP analyses cover all states and it will be helpful to monitor the changes in standards and instruction both in states participating in the new standards and those states that do not choose to participate. In this context, the grade 8 questions that these background questions will explore below include:

- How important are standards and other factors perceived at the school level in driving the school curriculum and instruction?
- How much time weekly is spent on grade 8 science?
- What scientific disciplines (life science, earth or space science, physics or chemistry) are focused on at grade 8?
- What aspects of science receive the greatest and least emphasis in curriculum and instruction?
- How frequently are different modes of instruction (e.g., hands-on activities, working with others) employed in teaching?

Because the NGSS also have far greater stress on engineering and engineering design than previous standards and most current educational practice, the new Technology and Engineering Literacy assessment [National Assessment Governing Board, 2011] being developed for NAEP also has great potential to monitor the changes which result in these areas of instruction. Indeed, because so many students have more access to technology outside of the classroom than inside, through smart phones and other devices, background variables related to use of information and communication technologies outside of school could provide exceptional valuable information for developers and users of the NGSS.

**NAEP Background Variables For Science Instruction and Curriculum**

NAEP background variables are available to address each of these questions. The NAEP background questions are arrayed by curriculum, instructional time and modes of instruction or classroom activities.

Several challenges in the specification of the NAEP background variables for science instruction and curriculum arise.

- The science instructional time variable interval range is: Less than 1 hour, 1-2.9 hours, 3-4.9 hours, 5-6.9 hours, 7 hours or more. Two-thirds of the responses are in the under 5 hours a week but almost all between 3-4.9. Eliminating the less than an hour and breaking-up the 3-4.9 into 3-3.9 and 4-4.9 would add a lot of information about instructional time amounts under 5 hours.
- Again, a lot of the question formats involve qualitative interpretations of amounts (little, some, a lot) rather than quantitative responses and the interpretation of qualitative amounts may depend upon respondents’ context. For example, the question about class time spent on earth or space
science asks for responses of “none, little, some or a lot.” But a more precise response is obtained for the question about how frequently students design a science experiment, which asks for a quantitative response by selecting among “never or hardly ever, once every few weeks, once a week,” etc.)

**Analyses and Findings**

### Exhibit 6-1

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>less than 5 hours</th>
<th>5-6.9 hours</th>
<th>7 or more hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>66</td>
<td>26</td>
<td>8</td>
</tr>
<tr>
<td>Mathematics</td>
<td>63</td>
<td>28</td>
<td>9</td>
</tr>
<tr>
<td>Reading</td>
<td>46</td>
<td>32</td>
<td>22</td>
</tr>
</tbody>
</table>

NAEP, Data Explorer

**How much weekly time is spent on science instruction?**

Exhibit 6-1 displays the amount of weekly time spent on instruction in science, mathematics and reading for 2011. At grade 8, weekly instructional time spent on science is comparable to that for mathematics, but less than the weekly instructional time allocated for reading. Approximately, two-thirds of the students spend less than 5 hours a week in science, equivalent to an hour a day. At the upper end of the time range, slightly less than 10 percent of all grade 8 students spend 7 or more hours a week on science. Research exploring how science instruction varies as student time increases and the relationship of instructional time to student learning could be useful in understanding how schools could better use greater instructional time more effectively.

Exhibit 6-2 explores how the distribution of science instructional time per week differs by whether students are low-income, as defined by school-lunch eligibility, or by their race/ethnicity. *To the extent that differences in weekly instructional time in science are observed, the differences in time of exposure to science favor at-risk students groups, suggesting the extra instructional time may be compensatory.*

- Non low-income students who are ineligible for school-lunch at grade 8 are 6 percentage points more likely to receive less than 5 hours of science instruction a week than are school-lunch eligible students.
White and Asian grade 8 students are 11 percentage points more likely to receive less than 5 hours of weekly science instruction than Black students.

### Exhibit 6-2

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Weekly hours of science instruction: Grade 8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>less than 5 hours</td>
</tr>
<tr>
<td></td>
<td>Percentage</td>
</tr>
<tr>
<td><strong>Low-income</strong></td>
<td></td>
</tr>
<tr>
<td>School-lunch eligible</td>
<td>62</td>
</tr>
<tr>
<td>School-lunch ineligible</td>
<td>68</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>69</td>
</tr>
<tr>
<td>Black</td>
<td>58</td>
</tr>
<tr>
<td>Hispanic</td>
<td>64</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>69</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>67</td>
</tr>
</tbody>
</table>

NAEP, Data Explorer

### What is the basis for structuring a school’s science program?

Understanding the basis for structuring science education programs is important to understanding how to leverage changes intended to improve science education. In particular, the potential impact of the proposed changes from the Next Generation Science Standards will depend upon how closely school curriculum and instruction align with adopted standards.

Exhibit 6-3 reports on the schools’ responses when asked about the basis for how they structure their school’s science education program. The responses clearly support the central importance of state and district standards and assessment results in driving the structure of schools’ science education programs:

- 86% of the students were in schools that structured their program “a lot” according to state standards.
- 71% of the students were in schools that structured their program “a lot” according to district standards.
54% of the students were in schools that structured their program “a lot” according to state or district assessment results.

Potential leverage points over instruction that are relatively weaker, based on school responses, include:

- 6% of the students were in schools that structured their program to a large extent according to commercial programs.
- 13% of the students were in schools that structured their program to a large extent according to the discretion of teachers.

**What science content areas are taught at grade 8?**

A key element in leveraging and strengthening science education is knowing what science content domains are taught at different grades. Is instructional time about evenly divided among the major content domains: earth and space science, engineering and technology, life science, and physical science? Alternatively, are there concentrations in the teaching of particular science domains at certain grades? If so, these grade concentrations represent leverage points to focus on to improve content and instruction. Moreover, knowing present domain emphasis of instruction at a grade establishes a baseline against which to measure changes.

Exhibit 6-4 shows that instructional emphasis on different science content domains varies considerably at grade 8. Given the preponderance of grade 8 teachers with a major in the life sciences (Exhibit 6-4), this domain may be expected to receive the greatest attention, but life science was in fact among the least commonly taught in terms of spending a lot of class time.
The top two content areas in terms of “a lot” of class time were the physical science, taught “a lot” by 61 percent of the school respondents, and earth and space science, at 39 percent.

By contrast, only about 20 percent of the teachers indicated that life science received a lot of instructional emphasis at grade 8.

Despite efforts over the past few years to integrate engineering and technology into science instruction, at least at grade 8 that is not happening, with only 5 percent of the teachers giving this a lot of emphasis.

What are the science practices and attitudes that receive the greatest instructional emphasis?

### Exhibit 6-4

<table>
<thead>
<tr>
<th>Science Content Domains By Grade 8 Class Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science Domain</td>
</tr>
<tr>
<td>class time spent on earth and space science</td>
</tr>
<tr>
<td>class time spent on engineering and technology</td>
</tr>
<tr>
<td>class time spent on life science</td>
</tr>
<tr>
<td>class time spent on physical science</td>
</tr>
</tbody>
</table>

NCES, NAEP Data Explorer

### Exhibit 6-5

<table>
<thead>
<tr>
<th>Teacher reported emphasis on science practices and attitudes, 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area of emphasis</td>
</tr>
<tr>
<td>emphasis on applying science to environment</td>
</tr>
<tr>
<td>emphasis on further study</td>
</tr>
<tr>
<td>emphasis on importance of science in daily life</td>
</tr>
<tr>
<td>emphasis on inquiry skills</td>
</tr>
<tr>
<td>emphasis on interest in science</td>
</tr>
<tr>
<td>emphasis on problem-solving skills</td>
</tr>
<tr>
<td>emphasis on scientific facts and principles</td>
</tr>
<tr>
<td>emphasis on scientific methods</td>
</tr>
<tr>
<td>emphasis on scientific writing skills</td>
</tr>
<tr>
<td>emphasis on skills in lab techniques</td>
</tr>
<tr>
<td>emphasis on systematic observation skills</td>
</tr>
</tbody>
</table>

// is approximately zero

Source: NCES, NAEP Data Explorer
Exhibit 6-5 presents teacher responses to questions about degree of emphasis given in instructional to different practices and attitudes. Practices represent the types of behaviors scientists engage in, such as inquiry skills, lab techniques or scientific writing. Attitudes are about interest in science and are associated with NAEP achievement scores (Exhibit 3-2).

Among the practices and attitudes taught, four received a lot of emphasis by half or more of the teachers (Exhibit 6-5).

- The science practices emphasized by half or more of the teachers of grade 8 science are scientific facts and principles (61 percent) and scientific methods (50).
- The student attitudes toward science given “a lot” of emphasis by more than half the teachers are “interest in science” (58 percent) and the importance of science in daily life (54 percent).

Teacher instructional emphasis is consistent with the research on the importance of positive attitudes towards science. Note that the final version of the Next Generation Science Standards (NGSS) mentions the importance of the soft variables (or “21st-century skills”), including positive student attitudes toward science, in their Executive Summary: “The affective domain, the domain of learning that involves interests, experience and enthusiasm, is a critical component to science education” (Achieve 2013). But the NGSS authors decided not to include any performance expectations for interest or motivation. NGSS restrict themselves to “endpoints of learning,” not incorporating standards for any factors which are known or suspected of influencing student learning. If NGSS-related assessments follow the same endpoint focus, they may fail to provide information on those noncognitive variables. This makes NAEP background variables even more crucial to help understand how NGSS or other new standards are implemented and their impacts on factors like interest and motivation. Out of school factors may provide a significant source of soft skills, which could also be informed by appropriate new background variables.

**How frequently and in what content areas do teachers do hands-on science activities?**

A priority in science education is for students to understand and be able to apply at their own level the practices that scientists engage in as they explore and understand the world around them. Rather than having students being passively taught about scientific practices, a more authentic way for students to learn about the practices of science is for students to engage in their own hands-on, minds-on scientific inquiry activities. These activities range from design activities, through investigation, to
analysis, writing and presenting results. Because the terms “hands-on” and “inquiry” are used with a variety of meanings in education, it may be desirable in the future for these items to provide a definition in the assessment, or to break these practices down so that individual elements of inquiry pedagogy are reported.

### Exhibit 6-6

**Frequency of students carrying out hands-on science projects including designing, measuring and discussing results of science projects, 2011 (Student Reported)**

<table>
<thead>
<tr>
<th>Type of Hands-on Science Project/Activity</th>
<th>Never or hardly ever</th>
<th>Once or twice a month</th>
<th>Once or twice a week</th>
<th>Every day or almost every day</th>
</tr>
</thead>
<tbody>
<tr>
<td>how often students do hands-on science activities</td>
<td>2</td>
<td>25</td>
<td>56</td>
<td>16</td>
</tr>
<tr>
<td>do hands-on projects with chemicals</td>
<td>30%</td>
<td>70%</td>
<td>36%</td>
<td>6%</td>
</tr>
<tr>
<td>do hands-on projects with electricity</td>
<td>57%</td>
<td>41%</td>
<td>36%</td>
<td>16%</td>
</tr>
<tr>
<td>do hands-on projects with living things</td>
<td>68%</td>
<td>32%</td>
<td>43%</td>
<td>33%</td>
</tr>
<tr>
<td>do hands-on projects with rocks or minerals</td>
<td>57%</td>
<td>43%</td>
<td>56%</td>
<td>33%</td>
</tr>
<tr>
<td>do hands-on projects with simple machines</td>
<td>42%</td>
<td>58%</td>
<td>67%</td>
<td>33%</td>
</tr>
<tr>
<td>do hands-on projects with thermometer or barometer</td>
<td>42%</td>
<td>56%</td>
<td>67%</td>
<td>33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design a science experiment</th>
<th>Never or hardly ever</th>
<th>About once a few weeks</th>
<th>About once a week</th>
<th>About 2-3 times a week</th>
<th>Every day or almost every day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td></td>
</tr>
<tr>
<td>39%</td>
<td>36%</td>
<td>16%</td>
<td>6%</td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>Discuss measurements for science project</td>
<td>35%</td>
<td>33%</td>
<td>19%</td>
<td>9%</td>
<td>4%</td>
</tr>
<tr>
<td>Discuss results of science project</td>
<td>24%</td>
<td>35%</td>
<td>23%</td>
<td>12%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Exhibit 6-6 presents teachers’ responses about the range of coverage of hands-on science activities. *Teachers nearly universally conduct hands-on activities with their grade 8 students, but from 24 percent to 39 percent of the students do not design a science experiment (39 percent), discuss measurement (35 percent), or discuss the results of their science project (24 percent). A surprising result is that although the most commonly taught science domain at grade 8 was physics, electricity and simple machines were among the least frequent hands-on activities, which may raise questions about how this physics is taught.* Specific findings about hands-on activities include:

- Only 2 percent of the students respond they never carry out a grade 8 hands-on science activity.
- Students indicate students most frequent grade 8 hands-on project is with chemicals (70 percent) and the least frequent is with living things (32 percent).
percent), electricity (36 percent) and simple machines (33 percent).

- Although doing grade 8 hands-on science activities is nearly universal, carrying out the steps of an investigative process is not. Twenty-four percent of the grade 8 students never discuss their results, thirty-five percent never discuss measurement for their science project and thirty-nine percent of the grade 8 students don't design an experiment.

Many science educators have become skeptical of using “hands-on” as a proxy term for inquiry or experimental work, since merely touching an apparatus does not guarantee that the apparatus is actually being used for investigation, rather than simply to confirm or demonstrate what the teacher or textbook is saying. “Inquiry” too is a suspect term, because there are so many interpretations of inquiry as a learning pedagogy. NAEP background items could help clarify what activities are actually being conducted, by asking teachers the extent to which their use of instruments in labs is for students to explore and create hypotheses, or to test and confirm previously discussed hypotheses.

### 7. Assessment and Ability Grouping/Differentiated Instruction

Assessment and ability grouping/differentiated instruction are often treated as independent education activities, but in practice they should be linked in organizing classroom instruction. That is, while science assessment serves summatively to produce student grades or for teacher accountability, science assessments can also be powerful drivers of achievement, when assessments are used formatively for feedback to guide instructional improvement (Black & Wiliam, 1998; Ruiz-Primo & Furtak, 2006). Ability grouping and differentiated science instruction are practices designed to respond to student performance by providing different students with different pathways for accessing instruction.

The NAEP background questionnaires ask both about assessment and ability grouping/differentiated instruction. With respect to assessment, the NAEP questions ask teachers of grade 8 science about:

- How they assess students – with multiple choice, short-written responses and long written responses?

- How are assessments used with specific students – adjust teaching strategies, discuss current performance, discuss progress toward previous goals, or set specific progress goals?
• How prevalent is the grouping of students for grade 8 science across classes and the use of differentiated instruction for different students within classrooms.

What types of assessments are used to assess science?

In contrast with the state standardized assessments in which multiple choice items are used because they are easier to grade, responses of grade 8 teachers of science indicate that multiple-choice assessments are less common in their classrooms than short answer responses (Exhibit 7-1). Teachers probably see no advantage in having to develop four answers for each question given that the assessments are hand-graded in any case. On the other hand, long-written responses take the most time to grade and are time consuming in terms of test time, so they remain the least frequently employed assessment by grade 8 science teachers.

Exhibit 7-1

<table>
<thead>
<tr>
<th>Types of assessments used for grade 8 science, teacher reported, 2011</th>
<th>Never or hardly ever</th>
<th>Once or twice a month</th>
<th>Once or twice a week</th>
<th>Almost every day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Assessment</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
<tr>
<td>assess science with multiple-choice tests</td>
<td>7</td>
<td>70</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td>assess science with short written responses</td>
<td>3</td>
<td>32</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>assess science with long written responses</td>
<td>24</td>
<td>53</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>NCES, NAEP Data Explorer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Assessments with short-answer questions were used at least weekly by science teachers of 65 percent of the students.
• Multiple choice and long written responses were used at least weekly by science teachers of only 23 percent and 24 percent of the students. In fact, science teachers of 24 percent of the of students never or hardly ever used long written responses.

How are the uses of assessments tied to educational purpose?

How frequently assessments are used depends upon purpose (Exhibit 7-2)

• The most frequent use of assessments was to guide teachers in adjusting their teaching strategies, as teachers make these adjustments routinely in relation to class performance.
• Virtually all teachers use assessments to discuss current student’s performance level at least a few times a year.
While a substantial majority of teachers use assessments to set goals or assess progress toward goals, close to 20 percent do not use assessments to set student goals and measure progress.

Exhibit 7-2

<table>
<thead>
<tr>
<th>Frequency of uses of assessments (or other student evaluations) for grade 8 science, teacher reported,</th>
<th>Never or hardly ever</th>
<th>A few times a year</th>
<th>Once or twice a month</th>
<th>Once or twice a week</th>
<th>Every day or almost every day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
<tr>
<td>adjusting teaching strategies to meet needs</td>
<td>8</td>
<td>24</td>
<td>26</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>discussing current performance level</td>
<td>5</td>
<td>30</td>
<td>39</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>discussing progress toward goals</td>
<td>16</td>
<td>37</td>
<td>31</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>assess science students by setting goals for specific progress</td>
<td>17</td>
<td>40</td>
<td>29</td>
<td>12</td>
<td>3</td>
</tr>
</tbody>
</table>

How prevalent is ability grouping and differentiated instruction in science education?

Ability grouping is the assignment of students to classes based on their perceived ability as indicated by student performance. Differentiated instruction is the adjustment of instruction within classrooms to students with different needs or abilities. The NAEP background variables provide information from school and teacher respondents about how common it is for students to be assigned to science classes by a student’s ability and for instruction to be differentiated within science classrooms.

Exhibit 7-3 indicates that 24 percent of the students are in schools in which students are assigned to classrooms based on ability. The NAEP trend data also suggest that this percentage has changed very little since 1996. The NAEP data (Exhibit 7-4) indicate that it is far more common for teachers to differentiate instruction by student ability or other attributes within science.

Exhibit 7-3

<table>
<thead>
<tr>
<th>Percent of grade 8 students assigned to science class based on ability, 1996–2011</th>
<th>Year</th>
<th>% Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2011</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>1996</td>
<td>21</td>
</tr>
</tbody>
</table>

NCES, NAEP Data Explorer
classrooms than to differentiate students by ability across classrooms.

- Differentiating instructional methods and materials when teaching science is used to a moderate or large extent by about 75 percent of the teachers.
- Creating ability groups with classrooms is used by 44 percent of the teachers. This compares with 24 percent of the schools responding that they assign students to classrooms based on ability.
- It is also quite common for teachers to set different standards for some students when teaching (48 percent) and to engage students in different science activities (44 percent).

### Exhibit 7-4

**Percent of grade 8 students by strategies used within a science classroom to differentiate science instruction for some students, 2011**

<table>
<thead>
<tr>
<th>Type of Strategy</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>create groups within this class for science instruction on the basis of ability (% Yes)</td>
<td>44</td>
</tr>
<tr>
<td>engage some students in different activities when teaching science (% moderate/large extent)</td>
<td>44</td>
</tr>
<tr>
<td>set different standards for some students when teaching science (% moderate/large extent)</td>
<td>48</td>
</tr>
<tr>
<td>use different methods for some students when teaching science (% moderate/large extent)</td>
<td>75</td>
</tr>
<tr>
<td>use other materials for some students when teaching science (% moderate/large extent)</td>
<td>73</td>
</tr>
</tbody>
</table>

NCES, NAEP Data Explorer

### 8. Improving the NAEP Science Background Questions and Their Use

The current NAEP background questions have provided useful information about a number of aspects of the current status of science education. At the same time, it is helpful to guide development of future NAEP questionnaires to return to the list of expert-identified “hot issues,” and the requested indicators from the NRC (2013), as described in the introduction to this report. Exhibit 8-1 shows that while none of the top “hot issues” were fully addressed by the 2011 science questionnaire, six of ten were partially addressed, while four were not explored at all.
Exhibit 8-1. How well NAEP addresses “hot” science issues identified by leading STEM educators

<table>
<thead>
<tr>
<th>Expert Identified “hot” topics</th>
<th>Issue Explored</th>
<th>Provides Some information</th>
<th>Issue Not Explored</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Are school systems prepared to respond to reforms underway in science education to focus on big cross-cutting topics and issues in science and teach these with greater depth and understanding?</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>• What factors motivate students to learn science?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>• Which topics in STEM are of the greatest interest to students? How does their performance relate to their interests?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>• Do out-of-school time activities, such as science clubs and museum visits, impact classroom performance?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>• How extensive is participation in “citizen science” projects, such as bird counts and other research by non-specialists, and do they contribute to science learning?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>• What kinds of teacher professional development are commonly available, such as online, blended, active, or passive, and does the kind or duration chosen correlate with student performance?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>• How do teachers deal with student misconceptions and wrong answers about science?</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>• To what extent are student laboratories using virtual science experiments, or data and instruments available online?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>• Are school laboratories used for exploring or for confirming?</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>• How much time is spent in teaching science using various strategies such as lecture, laboratory, and discussion?</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Exhibit 8-2. How well NAEP addresses indicators requested by
the National Research Council’s 2013 report (NRC, 2013)

<table>
<thead>
<tr>
<th>Issue</th>
<th>Explored</th>
<th>Provides Some Information</th>
<th>Not Explored</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Number of, and enrollment in, different types of STEM schools and programs in each district.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Time allocated to teach science in grades K-5.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Science-related learning opportunities in elementary schools.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Adoption of instructional materials in grades K-12 that embody the Common Core State Standards for Mathematics and A Framework for K-12 Science Education.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Classroom coverage of content and practices in the Common Core State Standards and A Framework for K-12 Science Education.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Teachers’ science and mathematics content knowledge for teaching.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Teachers’ participation in STEM-specific professional development activities.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Instructional leaders’ participation in professional development on creating conditions that support STEM learning.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Inclusion of science in federal and state accountability systems.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Inclusion of science in major federal K-12 education initiatives.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. State and district staff dedicated to supporting science instruction.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. States’ use of assessments that measure the core concepts and practices of science and mathematics disciplines.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. State and federal expenditures dedicated to improving the K-12 STEM teaching workforce.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Federal funding for the research identified in Successful K-12 STEM Education.</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The boldface items in Exhibit 8-2 are the NRC’s most important indicators selections. Our judgment on the extent to which the NAEP questionnaires currently cover these points includes the following observations:

- Main NAEP has data for 4, 8, and 12th grade, while the NRC calls for more grades, as in indicator 2, which asks for K-5.
- NAEP asks about teacher backgrounds, but doesn’t measure teacher knowledge, as in indicator 6.
- NAEP doesn’t collect any data funding, as requested for indicator 13 and 14.
- Data on state and federal policy and funding requested by the NRC would be much more efficiently collected by separate means from NAEP, which is organized around assessments administered to tens or hundreds of thousands of individual students, teachers, and administrators, but not state or federal officials.
Recognizing the severe limitations of respondent time and the overall structure of NAEP, we propose three additional question areas to strengthen the NAEP questionnaire’s ability to monitor key changes underway in science education in several of the areas mentioned above.

Expanding current questionnaire topics:

- **Recommendation 1**: *NAGB consider adding questions about the amount of time spent learning science and the nature of these activities in out-of-school or after-school settings, and the coordination of these activities with schools.*

  Increasingly, instructional technology is making the learning of science through computers and the Internet feasible in a variety of settings such as after-school programs or in the home. Moreover, science museums are increasingly partnering with schools to extend and enrich school science programs, yet currently NAEP has little information about these activities.

  NAEP background questions could ask students about unique out-of-school and informal science learning activities like building apps, using high-tech toys such as remote controlled airplanes, participating in citizen science projects, taking courses in science museums or nature centers, watching science shows on television or following science expeditions live on the Internet. All these could inform policy makers about new opportunities for synergies between in-school and out-of-school learning.

- **Recommendation 2**: *NAGB consider expanding background questions about teacher professional development to obtain information on the nature, duration and quality of that professional development.*

  The range of quality and quantity of professional development is huge, and practice is inconsistent from year to year and from topic to topic. How much professional development time do teachers receive and what is the fraction devoted to science? To what extent does the professional development for science involve inquiry or hands-on activity by the teachers? To what extent have teachers changed their science teaching based on professional development? What are areas in which teachers would like to have increased professional development and are there areas where current professional development has not been useful?

Adding a new questionnaire topic:

- **Recommendation 3**: *NAGB consider adding NAEP background questions to monitor how changes in science standards are affecting instruction, the challenges schools and teachers face in implementing new or changed standards, and whether they are receiving the needed technical assistance and
professional development to bring about effective implementation of the new standards.

NAEP teacher surveys can go beyond documenting the formal adopted changes in standards and can ask on a regular basis schools and teachers directly about the how the standards are affecting their instruction and the challenges faced in implementing changes in standards.

**Technical background questionnaire issues:**

Several technical survey recommendations are also proposed for offsetting the data burden of additional questions and to better understand survey responses.

- **Recommendation 4.** NAGB explore offsetting the additional time burden from adding additional science background questions through rotating questions in and out of the science assessments and matrix sampling so that a respondent answers only a sample of the background questions.

  Holding down background questionnaire time is a NAEP priority especially for the student questionnaire, which is administered to the student along with taking the NAEP science assessment. Rotating items recognizes that for many questions it is sufficient to have broad trend data spaced perhaps every half decade, so that questions can be interlaced among science surveys without adding to overall survey burden. Matrix sampling of background questions would parallel the matrix sampling of science assessment items so that no respondent answers all the background questions. However, statistical analyses should be carried out to assess the impacts of the reduced sample for any single question on the ability to disaggregate findings by student subgroups.

  A fifth recommendation to NAGB is that certain science items be designated as priority for NCES to conduct cognitive science labs to clarify and better understand survey responses to background variable questions.

- **Recommendation 5.** NAGB should recommend to NCES the use of their cognitive science laboratory to clarify questions and responses in three areas of the science background variables: (1) understanding the schools generally more positive responses than teachers to questions about resource availability; (2) exploring the accuracy of school, teacher or student question responses when responses are qualitative and judgmental, such as “a little” or “a lot;” and (3) taking advantage of future science assessments that will be done on the computer to replace interval responses (e.g. 0-2 hours, 2-5 hours) with continuous sliders enabling respondents to drag an arrow to any point along a continuum.
Extending the Usefulness of the Findings

The ultimate value of the NAEP science background findings is when they are used to inform decisions, which policymakers, educators and researchers make regularly about science education and how to improve it.

- Recommendation 6. NAGB should explore ways to support the use of the current findings by policymakers and educators and to stimulate further analyses by other researchers. NAGB should explore with NCES coordinated support for the further use of the NAEP science background variables.

NAGB in cooperation with NCES could provide technical online guidance, workshops or financial support to facilitate: (1) Policymakers developing current and leading indicators for the state of science education; or tracking and making adjustments when the data shows trends in background variables that they are trying to support (e.g., new standards or new strategies for professional development); 2) Educators comparing their own system to national (or state) averages on science education characteristics (e.g., % of science teachers with a major in science); and (3) Researchers conducting follow-up research including multivariate analyses based on interesting/provocative/counter-intuitive correlations (e.g., Buckley (2009), cited earlier, employed student response-style adjustments by country on international PISA survey results to find a positive but non-linear attitude-achievement relationship within countries.
Acknowledgements

The authors would like to thank Arthur Eisenkraft, Dennis Schatz, Martin Storksdieck, and Gerald Wheeler, who consented to discuss with us for this article the kinds of issues that science educators engaged currently, and which NAEP background variables might illuminate. We also thank the Board, Executive Director, and senior advisor of the Noyce Foundation, which examined a draft of Exhibit 2-1 and gave us major suggestions for improvements. We appreciate all this generous input, but of course responsibility for any errors and the conclusions made here remains with the authors.

References


National Assessment Governing Board (n.d.). Science issue paper: Potential Background issues for the NAEP science assessments based on the 2009 science framework


National Research Council (2009). *Learning science in informal environments: people, places, and pursuits* / Philip Bell ... [et al.], editors ; Committee on Learning Science in Informal Environments, Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Available April 2013 online: http://www.nap.edu/catalog.php?record_id=12190#toc.


National Science Teachers Association (2002). *NSTA position statement: Learning conditions for high school science.* Available July 2013 online:

National Science Teachers Association (2012). *NSTA Standards for Science Teacher Preparation.* Available April 2013 online:


EXECUTIVE SUMMARY

By Alan Ginsburg and Marshall S. Smith

Introduction

This is the first of two reports exploring the use of the background data collected by the Nation Assessment of Educational Progress (NAEP) to develop key education indicators at national, state, and urban district system levels. Key indicators are statistics that regularly measure an important condition of education. For example, NAEP can tie to its student achievement results the reporting of background conditions on: students’ attitudes, motivation and excessive absenteeism; measures of teacher quality; and indicators of the nature of reading and math instruction (e.g., amount of instructional time).

The GAO identifies three broad purposes of indicators:

- Increase transparency and public awareness.
- Foster civic engagement and collaboration.
- Monitor progress, establish accountability for results, and aid decision-making.

In a NAEP context, indicators also serve to:

- Identify for each cognitive assessment (e.g., reading) a set of key indicators, which are derived from the background variables and are monitored regularly.

Specifically, this first report is intended to develop a general indicators framework specifying an organizing structure, potential indicators, measurement criteria and reporting design. This report is based on a review of several major international or domestic data collections and reports produced by organizations other than NAEP:

**International**

- OECD’s *Education At a Glance*
- International Education Association’s 2011 TIMSS *Mathematics Assessment* covering grade 4 and 8.
- OECD’s 2009 *PISA Report*

**Domestic**

- National Center for Education Statistics *Condition of Education*
- Education Week’s *Quality Counts*
- U.S. Department of Education’s *Annual Priority Performance Goals*
- National Academy of Sciences’ *Key National Education Indicators*
## Potential Indicators by Organizing Structure

### Exhibit EX-1

<table>
<thead>
<tr>
<th>Locus of Education Activity</th>
<th>Key Drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student</strong></td>
<td>- Command of core content, using NAEP scores</td>
</tr>
<tr>
<td></td>
<td>- College readiness levels by age and grade</td>
</tr>
<tr>
<td></td>
<td>- Career readiness (21st century skills)</td>
</tr>
<tr>
<td></td>
<td>- Attended preprimary education</td>
</tr>
<tr>
<td></td>
<td>- Chronic absenteeism</td>
</tr>
<tr>
<td></td>
<td>- Student motivation and belief that hard work is more important than luck</td>
</tr>
<tr>
<td></td>
<td>- Student positive attitudes toward subject</td>
</tr>
<tr>
<td></td>
<td>- Student uses research-based approaches to learning subject</td>
</tr>
<tr>
<td></td>
<td>- Student respect for teacher and visa versa</td>
</tr>
<tr>
<td></td>
<td>- Participation in extra-curricular activities including community service</td>
</tr>
<tr>
<td></td>
<td>- Home learning environment</td>
</tr>
<tr>
<td></td>
<td>- Formal and informal learning outside school – nature of the their neighborhood</td>
</tr>
<tr>
<td><strong>Teacher</strong></td>
<td>- Teachers with less than 3-years experience</td>
</tr>
<tr>
<td></td>
<td>- Teachers with mastery-level and current knowledge of content they are teaching.</td>
</tr>
<tr>
<td></td>
<td>- Teachers with mastery-level and contemporary knowledge of child and adolescent development</td>
</tr>
<tr>
<td></td>
<td>- Teacher-student interactions that demonstrate high levels and qualities of involvement, language, stimulation, and expansion of thinking and cognitions, and sensitivity to students’ perspectives, individual experiences, and backgrounds</td>
</tr>
<tr>
<td></td>
<td>- Teacher-student interactions that indicate that teachers respect students.</td>
</tr>
<tr>
<td></td>
<td>- Teacher working conditions</td>
</tr>
<tr>
<td></td>
<td>- Average district teacher salary</td>
</tr>
<tr>
<td></td>
<td>- Time teachers spend teaching</td>
</tr>
<tr>
<td></td>
<td>- Teacher has high quality professional development and comprehensive induction programs Quality of the principal</td>
</tr>
<tr>
<td></td>
<td>- Teachers belong to professional learning communities</td>
</tr>
<tr>
<td><strong>School/Classroom</strong></td>
<td>- Content of instruction aligned with standards</td>
</tr>
<tr>
<td></td>
<td>- Effective use of technology to support instruction</td>
</tr>
<tr>
<td></td>
<td>- School Climate – whether the school is a learning organization – do teachers work together?</td>
</tr>
<tr>
<td></td>
<td>- Instructional time per subject</td>
</tr>
<tr>
<td></td>
<td>- Engaged instruction in subject</td>
</tr>
<tr>
<td></td>
<td>- Emphasis on continuous improvement on outcomes through both formative and summative assessments aligned with standards</td>
</tr>
<tr>
<td></td>
<td>- Emphasis on continuous improvement of practices of teaching</td>
</tr>
<tr>
<td></td>
<td>- School SES Composition</td>
</tr>
<tr>
<td></td>
<td>- Safe &amp; orderly school climate</td>
</tr>
<tr>
<td></td>
<td>- Teacher-student ratio</td>
</tr>
<tr>
<td></td>
<td>- School resource shortages</td>
</tr>
<tr>
<td></td>
<td>- School lacks key characteristics, coaches for teachers, support systems for students, technology, books</td>
</tr>
<tr>
<td><strong>System (district, state or nation)</strong></td>
<td>- Support for implementation of new content standards</td>
</tr>
<tr>
<td></td>
<td>- Alignment of assessment with content standards</td>
</tr>
<tr>
<td></td>
<td>- Accountability with emphasis on continuous improvement</td>
</tr>
<tr>
<td></td>
<td>- K-12 education spending as a share of gross domestic product</td>
</tr>
<tr>
<td></td>
<td>- K-12 spending per student</td>
</tr>
<tr>
<td></td>
<td>- Disparity in resources across districts within states</td>
</tr>
</tbody>
</table>

The indicator structure in Exhibit EX-1 is focused primarily around variables at the student, teacher, school/classroom and system levels that support learning outcomes.
across three aspects of education conditions:

- **Results indicators** include student assessment outcomes (such as from NAEP), but also teacher evaluations that reflect student outcomes, and other outcomes such as secondary school completion and parent satisfaction with the school.

- **Enabler indicators** reflect formal learning at different levels of education. These include student exposure to preschool, teachers’ knowledge and skills and their ability to apply them to create a challenging and supportive classroom learning environment; and school instructional time and student engagement in the content areas. Enablers also include system policies and regulations at district, state and national levels regarding teacher certification, standards, assessment, and accountability.

- **Context/constraint indicators** reflect factors not readily manipulable by the education system, although conditions may be changeable with proper interventions, such as schools intervening in the home learning environment. These factors include: learning at home and outside the school in formal and informal settings; factors influencing teacher quality including salaries and working conditions; and factors affecting the school learning environment including school safety, climate and class size.

**Indicator Measurement**

A sound measure for an indicator should meet criteria of validity, reliability, and consistency overtime.

**Validity of Indicators.** A valid measure is one that adequately captures the underlying education condition of interest. Combining responses from a number of questions around a topic into a larger comprehensive indicator scale can produce richer indicator measures than reporting on a single question, but this approach currently is not used in NAEP background factor analyses. Exhibit EX-2 illustrates a scale developed for TIMSS at grade 4 measuring students’ early numeracy activities before beginning primary school.
Exhibit EX-2 Development of Indicator Scales from Multiple Questions

Reliability of Indicators. A reliable indicator measure is one that produces consistent results when repeatedly measuring the same underlying condition. Qualitative responses may be unreliable when sensitive to the position of the respondent. In the recently completed NAEP background paper on science, Exhibit EX-3 was presented showing that teachers were more likely than school principals to indicate that resources were “not at all available” within a school. This is not surprising as it is principals who are responsible for school resource availability.
**Consistency over time.** A consistent measure requires using the same measure for an indicator over time. To the extent that measures are changed from time period to time period then it is unclear whether a change comes about because of a real change in the underlying condition or changes in the measure. The report by the Expert Panel on Strengthening NAEP Background Questions (2012) addressed this issue in its recommendation 1d:

"Use consistency over time as a criterion to consider for question selection and wording. NAEP’s inconsistent inclusion of background questions weakens its potential to track trends and improvements within a subject area and topic.

For example, the Expert Panel found that only one-third of the 2011 questions asking about course offerings provided at least a 6-year trend. None of the questions about curriculum or school resources used in 2011 were found on the 2005 or earlier questionnaires.

**Sources of Indicator Data.** The reports that were studied use two ways to obtain indicator data, which differentiate them from NAEP.

First, TIMSS and PISA both conduct a household survey to obtain information directly from parents or guardians about socio-economic status and the home learning environment. TIMSS innovatively combined with PIRLS to develop a joint sample household survey for grade 4 students. The household survey included questions about:

- Early numeracy activities in the home before beginning primary school (See Exhibit EX-2)
- Early literacy activities in the home before beginning primary school
- Amount of exposure to preschool
- Family perception about child’s literacy and numeracy skills before entering primary school
- Family interaction with the child about school work
- Family perceptions about school
- Family literacy environment
- Family SES

A second source of data that is different from NAEP is the pooling of information across different surveys. Both the *Condition of Education* and *Education at a Glance* are drawn almost entirely from data series generated by other surveys. *Quality Counts* is a state-level amalgam of direct analyses of state policies by Education Week combined with data from other surveys, which prominently features NAEP assessment results.

A form of pooling could be the aligning of NAEP survey questions with international assessment items, as illustrated in Exhibit EX-4. The exhibit suggests that at least for U.S. middle schools, only about 12 percent of U.S. principals are having at least some difficulty filling vacancies for mathematics teachers. This compares with other Western
English-speaking countries of 41 percent of the principals having difficulty hiring math teachers in Australia, 37 percent in England, and 44 percent in New Zealand. Adding the same question about vacancies to the NAEP principal survey for mathematics would yield U.S. state-by-state comparisons.

### Exhibit EX-4

#### Schools Having Difficulties Filling Vacancies With Mathematics Teachers, Grade 8

<table>
<thead>
<tr>
<th>Country</th>
<th>No Vacancies Percent of Students</th>
<th>Vacancies Are Easy To Fill Percent of Students</th>
<th>Vacancies Are Somewhat Difficult To Fill Percent of Students</th>
<th>Vacancies Are Very Difficult to Fill Percent of Students</th>
<th>Total of Vacancies Somewhat or Very Difficult To Fill Percent of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>25</td>
<td>34</td>
<td>31</td>
<td>10</td>
<td>41</td>
</tr>
<tr>
<td>Chinese Taipei</td>
<td>46</td>
<td>44</td>
<td>10</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>England</td>
<td>28</td>
<td>35</td>
<td>27</td>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>Finland</td>
<td>42</td>
<td>46</td>
<td>10</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Hong Kong SAR</td>
<td>48</td>
<td>44</td>
<td>8</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Japan</td>
<td>82</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Korea, Rep. of</td>
<td>67</td>
<td>16</td>
<td>15</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>New Zealand</td>
<td>30</td>
<td>27</td>
<td>38</td>
<td>6</td>
<td>44</td>
</tr>
<tr>
<td>Norway</td>
<td>38</td>
<td>40</td>
<td>20</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>81</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

Next Steps: Using the International and Domestic Indicator Framework as a Guide, Develop a NAEP Education Indicators Framework and Provide Examples with Current Data

A second report will be prepared for NAGB by December 2013 with a recommended set of *Key Indicators* and recommended improvements in NAEP data to strengthen indicator measurement or fill indicator gaps. The December 2013 report will:

- Specify a NAEP Education Indicators Framework for the background variables applicable across cognitive assessments.
- Propose indicators that are research-based and estimable by:
  - offering examples using current NAEP data.
  - suggesting changes to the current NAEP questionnaires.
  - introducing a fundamentally new NAEP questionnaire or drawing data from education surveys other than NAEP.
- Explore opportunities for combining NAEP with International or other NCES indicator-supporting data.
- Explore how NAEP reports could best display a pyramid information approach along the lines of an indicator dashboard.
DEVELOPING A NAEP INDICATORS FRAMEWORK: LESSONS FROM MAJOR INTERNATIONAL AND DOMESTIC EDUCATION INDICATOR REPORTS

An Exploratory Analysis
Prepared for the National Assessment Governing Board

By Alan Ginsburg and Marshall S. Smith

August 2013

Alan Ginsburg is an education consultant and analyst. He is former director of policy and program evaluation services for the U.S. Department of Education. Marshall S Smith is visiting Scholar, Carnegie Foundation for the Advancement of Teaching and former U.S.Under Secretary of Education. The data analyses and interpretations in this report are those of the authors and do not necessarily represent the views of the National Assessment Governing Board.
Table of Contents

DEVELOPING A NAEP INDICATORS FRAMEWORK: LESSONS FROM MAJOR INTERNATIONAL AND DOMESTIC EDUCATION INDICATOR REPORTS 3

1. INTRODUCTION 3

2. INDICATOR BACKGROUND AND METHODOLOGICAL CONSIDERATIONS 4

3. INTERNATIONAL AND NATIONAL INDICATOR REPORTS 7
   INTERNATIONAL: OECD’S EDUCATION AT A GLANCE 7
   INTERNATIONAL: IEA’S TIMSS (TRENDS IN INTERNATIONAL MATHEMATICS AND SCIENCE STUDY) 11
   INTERNATIONAL: OECD’S PISA (PROGRAM FOR INTERNATIONAL STUDENT ASSESSMENT) 14
   DOMESTIC: U.S. DEPARTMENT OF EDUCATION’S PRIORITY PERFORMANCE GOALS 18
   DOMESTIC: EDUCATION WEEK’S QUALITY COUNTS 19
   DOMESTIC: NCES’S CONDITION OF EDUCATION 21
   DOMESTIC: NATIONAL ACADEMY OF SCIENCES’ KEY NATIONAL EDUCATION INDICATORS 22
   IMPLICATIONS 26

4. PROPOSED DESIGN OF KEY INDICATORS FRAMEWORK 27
   INDICATOR ORGANIZING STRUCTURE 27
   INDICATOR SELECTION 29
   INDICATOR MEASUREMENT 32
   SOURCES OF INDICATOR DATA 36
   REPORTING ON INDICATORS 38

5. NEXT STEPS: USING THE INTERNATIONAL AND DOMESTIC INDICATOR FRAMEWORK TO GUIDE DEVELOPMENT OF A NAEP INDICATOR FRAMEWORK AND PROVIDE EXAMPLES WITH CURRENT DATA 39

REFERENCES 41

The authors wish to thank Lawrence Feinberg, Assistant Director for Reporting and Analysis of the National Assessment Governing Board, for his many analytic and editorial contributions to strengthen this report.
DEVELOPING A NAEP INDICATORS FRAMEWORK: LESSONS FROM MAJOR INTERNATIONAL AND DOMESTIC EDUCATION INDICATOR REPORTS

1. Introduction

This is the first of two reports exploring the use of the background data collected by the National Assessment of Educational Progress (NAEP) to develop key education indicators at national, state, and urban district system levels. The report has been commissioned by the National Assessment Governing Board (NAGB) to follow-up on recommendations by its Expert Panel on Strengthening the NAEP Background Questions to “develop key background indicators.” (Smith, et.al., 2012).

Exploring NAEP’s potential as part of a national and sub-national system of key education indicators is a priority for enhancing the value of NAEP. Indicators are statistics that regularly measure an important condition of education (NRC, 2012). The regular (every two years) collection of background variables for NAEP cognitive assessments, has the unique potential to strengthen our understanding of the assessment results and expand the richness of statistics on the quality and equality of our educational systems at the national, state, and urban district levels. For example, NAEP can tie to its student achievement results the reporting of background conditions on students’ attitudes, motivation and excessive absenteeism, measures of teacher quality, and the nature of reading and math instruction, including the amount of instructional time at multiple levels of the education system.

Specifically, this first report explores a number of the major international and domestic data collections and reports produced by organizations other than NAEP. These data collection efforts, including those associated with international assessments of school achievement, specifically generate data organized around well-defined indicators or they produce and report data describing key components of the education system much like an indicators report.

The aim of this report is to distill from these other education data and reports an indicator framework that might help structure and guide similar work on NAEP. A NAEP indicators framework would specify key features of a useful and valid indicator system including its design. Within the NAEP context, an indicators framework could identify for each cognitive assessment (e.g. reading) a set of key indicators, which are derived from the background variables and regularly monitored.
A second report, due in December 2013, will apply this report’s indicator framework and use current NAEP data to develop some of the indicators proposed. It will also identify priority areas where NAEP could collect additional background data to address indicator gaps, or draw from other NCES statistics collections to fill these gaps.

The following sources of indicator or large-scale data reports similar to an indicator report are explored in this first report:

International

- OECD’s *Education At a Glance*
- International Education Association’s TIMSS *Math Assessment* covering grade 4 and 8.
- OECD’s 2009 *PISA Report*

Domestic

- National Center for Education Statistics *Condition of Education*
- Education Week’s *Quality Counts*
- U.S. Department of Education’s *Annual Priority Performance Goals*
- National Academy of Sciences’ *Key National Education Indicators*

The examination of these sources of indicator frameworks proceeds as follows:

- Section 2: Indicator Background and Methodological Considerations
- Section 3: International and National Indicator Reports
- Section 4: Proposed Design of Key Indicators Framework
- Section 5: Next Steps: Using the International and Domestic Indicator Framework to Guide Development of a NAEP Education Indicator Framework

2. Indicator Background and Methodological Considerations

The design of an education indicators framework begins with identifying the different purposes of education indicator systems. The focus must be on key national education indicators that are the ones most important to track over time. A key indicator system is thus distinguished from a statistical effort, such as the NCES *Digest of Education Statistics (http://nces.ed.gov/Programs/digest/)*, which attempts to produce an inclusive volume describing many aspects of the education system in some detail.

Within the scope of a key national indicator system, the Government Accountability Office (GAO) has identified three main purposes:
• *Increase transparency and public awareness* by giving decision-makers and the public easy, single-source access to credible and relevant information.

• *Foster civic engagement and collaboration* by bringing “increased pressure to bear on diverse parties in the public and the private sectors...to break down traditional boundaries between various actors and organizations” (GAO, 2011).

• *Monitor progress, establish accountability for results, and aid decision-making.* This focus includes assessing performance, discussing options, and making choices about major education concerns of fundamental and long-term significance.

This third purpose of monitoring progress and establishing accountability for results has been a particularly prominent aim of federal government-wide performance indicator policy including performance indicators established by the U.S. Department of Education. When an education indicator is associated with a defined policy or education objective, it becomes an *education performance indicator* that over time measures progress toward that objective. At the federal level, the Government Performance and Results Act (GPRA) requires every federal department, including the U.S. Department of Education, to establish performance indicators for every program. These indicators have been required since 1997.

Currently, as part of its GPRA reporting, the U.S. Department of Education has established for FY13 “priority performance goals” that illustrate the strategic importance of indicator measurement of educational conditions that might influence achievement. For pre-K through secondary school, the Education Department’s priority performance goals include:

• *“Improve outcomes for all children from birth through third grade.* By September 30, 2013, at least nine states will implement a high-quality plan to collect and report disaggregated data on the status of children at kindergarten entry.

• *Improve learning by ensuring that more students have an effective teacher.* By September 30th, 2013, at least 500 school districts will have comprehensive teacher evaluation and support systems and a majority of States will have statewide requirements for comprehensive teacher and principal evaluation and support systems.

• *Demonstrate progress in turning around the nation’s lowest-performing schools.* By September 30th 2013, 500 of the nation’s persistently lowest-achieving schools will have demonstrated significant improvement and serve as potential models for future turnaround efforts.
• Make informed decisions and improve instruction through the use of data. By September 30th, 2013 all states will implement comprehensive statewide longitudinal data systems.
• Prepare all students for college and career. By September 30th 2013, all states will adopt internationally-benchmarked college- and career-ready standards.”

NAEP background variable data could potentially provide independently generated indicators in most of these priority areas that would add information value in several ways.

One important way NAEP would add value is through its capacity to *disaggregate national data* to provide comparable indicators across all states and more than 20 urban districts. In this context, NAEP started out reporting only nationally representative assessment data. Following the report, *A Nation At Risk* (1983), then Secretary of Education Ted Bell wanted a way to hold States accountable for improving measurable performance of student outcomes. At his direction, the Department of Education staff developed and published a Wall Chart, a one-page summary set of less-than-perfect state-by-state indicators of student outcomes, education services and context (student characteristics). Between 1984 and 1989 the announcement of the Wall Chart annually produced the Department’s largest press conferences.

But methodological limitations of the available State-by-State data led Congress, heavily supported by the Council of Chief State School Officers, to fund in 1988 a larger NAEP sample representative state-by-state (Ginsburg, Noell, and Plisko, 1988). NAEP is unique among NCES databases in yielding *comparable student assessment data for math, reading or science* subjects coupled with student, teacher and school background data for every state and currently for 21 individual urban districts.

A second way NAEP could potentially add value is *consistency of the measurement of the indicator series over time*. The real power of indicators comes from establishing baselines and then measuring change regularly over time. NAEP offers the potential for consistent measures from repeated administrations. Great care is taken to ensure comparability of NAEP assessments over time.

Unfortunately, as the Expert Panel report (2012) concluded, “NAEP’s inconsistent inclusion of background questions weakens its potential to track trends and improvements within a subject area and topic.” Nonetheless, more consistent and useful data series could be created from existing or new NAEP measures.

Along with different characteristics of individual indicators, *different perspectives of the education system can generate different frameworks* for identifying key indicators. One system perspective is by the *stages of student learning*. NAEP currently does not address preschool, but begins with a grade 4 assessment. An end-
of-preschool assessment has been discussed. Also, other surveys ask retrospective questions of parents or guardians as shown below.

A second system perspective is a *production function model*. This applies an economic model to education typically covering school contexts, school inputs, school services and student outcomes. Each of these components has multiple factors and research on the significance of each component for outcomes that can help distill the key indicators.

A third perspective is reflected in a National Academy of Sciences (1991) report *Education Counts*, which guides indicator development based on identifying six *critical issue areas* that an indicator system should address:

1. learner outcomes;
2. quality of educational institutions;
3. readiness for school;
4. societal support for learning;
5. education and economic productivity; and
6. equity (measures of resources, demographics, and students at risk).

Development of an indicators framework faces the challenge of defining a limited set of key education indicators or performance indicators within an overall system of indicators at different levels of disaggregation and over time. Specifying the features of the indicators framework is a critical initial step toward this prioritization. To move the indicator framework selection process forward, the next section explores key features related to indicators identified in the seven major international and national indicator or data reports listed above.

### 3. International and National Indicator Reports

This section examines seven international and domestic indicator-producing data systems and reports. Each system is described with respect to purpose, organizing framework and an overview of indicators focused around NAEP-relevant ages of early childhood and K-12 education.

**International: OECD’s Education at a Glance**

This annual report draws on various OECD surveys to measure the current condition of education internationally. The indicators “provide information on the human and financial resources invested in education, how education and learning systems operate and evolve, and the returns to educational investments.” (OECD, 2013, p.17).
Organizing framework. Exhibit 3-1 displays the three factors that form the basis for the organizing framework for Education at a Glance 2013 (OECD, 2013).

- The left-hand column organizes indicators by the level of the “actors” in the education system. It assesses the functioning and impact of education systems at four levels contributing to overall learning outcomes. These include: I. Individual participants in education and learning; II. The instructional setting and the learning environment involving teachers and classrooms within institutions; III. The educational institutions (e.g., schools, informal education providers) that are providers of educational services; and IV. The national education system as a whole. In U.S. NAEP terms, the different levels approximately correspond to students, teachers, schools and classrooms, and the national, state and district systems.

- The columns cluster the indicators by whether the focus is on: 1. The education outcomes/outputs of the education system; 2. Policy levers and contexts which shape the outputs/outcomes; and 3. The antecedents that

<table>
<thead>
<tr>
<th>Indicator Level of Actors of Education System</th>
<th>Indicator Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Individual participants in education and learning</td>
<td>1. Education and learning outcomes; 2. Policy levers and contexts shaping educational outcomes; 3. Antecedents or constraints that contextualize policy</td>
</tr>
<tr>
<td>II. Instructional settings</td>
<td>1. The quality of instructional delivery; 2. Pedagogy, learning practices and classroom climate; 3. Student learning conditions and teacher working conditions</td>
</tr>
<tr>
<td>III. Providers of educational services</td>
<td>1. The output of educational institutions and institutional performance; 2. School environment and organisation; 3. Characteristics of the service providers and their communities</td>
</tr>
<tr>
<td>IV. The education system as a whole</td>
<td>1. The overall performance of the education system; 2. System-wide institutional settings, resource allocations, and policies; 3. The national educational, social, economic, and demographic contexts</td>
</tr>
</tbody>
</table>

Cross-cutting policy issues addressed:
- Quality of educational outcomes and educational provision;
- Equality of educational outcomes and equity in educational opportunities; and
- Adequacy, effectiveness and efficiency of resource management.

Source: Adapted from Education At a Glance (2013).
define or constrain policy (e.g., teacher working conditions).

- At the bottom of the table, the cross-cutting policy issues indicate that each of the cells in the framework can be examined from three policy perspectives of quality, equality, and efficiency characteristics of resource management.

**Indicators.** The indicators reported in *Education at a Glance* are categorized in Exhibit 3-2 and are intended to emphasize one or more cells of this framework. The indicator structure resembles a production model with outputs, resources, access to education services and the learning process. With respect to a focus on early childhood or K-12 education, *Education at a Glance* covers:

- Chapter A, *Output of Educational Institutions and the Impact of Learning*, focuses only on students expected to complete upper secondary education.
- Chapter B, *Financial and Human Resources Invested in Education*, measures the amount of resources and types of expenditures on education at different levels with limited data on preprimary education.
- Chapter C, *Access to Education, Participation and Progression*, includes an important focus on access to early childhood education and transitions from secondary to tertiary education or from school to work.
- Chapter D, *The Learning Environment and Organization of Schools*, focuses on indicators of instructional time, teacher qualifications, teachers’ working time and teacher salaries.

Looking across *Education at a Glance*, the Chapter A indicators relate largely to the first column. However, as *Education at a Glance* notes the indicators for the other chapters are a mixture of policy levers and contexts.
Exhibit 3-2. Education At a Glance Indicators

Chapter A. The Output of Educational Institutions and the Impact of Learning

| Indicator A1 | To what level have adults studied? |
| Indicator A2 | How many students are expected to complete upper secondary education? |
| Indicator A3 | How many students are expected to complete tertiary education? |
| Indicator A4 | How many students complete tertiary education? |
| Indicator A5 | How does educational attainment affect participation in the labour market? |
| Indicator A6 | What are the earnings premiums from education? |
| Indicator A7 | What are the incentives to invest in education? |
| Indicator A8 | What are the social outcomes of education? |

Chapter B. Financial and Human Resources Invested In Education

| Indicator B1 | How much is spent per student? |
| Indicator B2 | What proportion of national wealth is spent on education? |
| Indicator B3 | How much public and private investment in education is there? |
| Indicator B4 | What is the total public spending on education? |
| Indicator B5 | How much do tertiary students pay and what public support do they receive? |
| Indicator B6 | On what resources and services is education funding spent? |
| Indicator B7 | Which factors influence the level of expenditure on education? |

Chapter C. Access to Education, Participation and Progression

| Indicator C1 | Who participates in education? |
| Indicator C2 | How do early childhood education systems differ around the world? |
| Indicator C3 | How many students are expected to enter tertiary education? |
| Indicator C4 | Who studies abroad and where? |
| Indicator C5 | Transition from school to work: where are the 15-29 year-olds? |

Chapter D. The Learning Environment and Organisation of Schools

| Indicator D1 | How much time do students spend in the classroom? |
| Indicator D2 | What is the student-teacher ratio and how big are classes? |
| Indicator D3 | How much are teachers paid? |
| Indicator D4 | How much time do teachers spend teaching? |
| Indicator D5 | Who are the teachers? |

Source: OECD, 2013
International: IEA’s TIMSS (Trends in International Mathematics and Science Study)

TIMSS 2011, like NAEP, administers mathematics and science assessments to grade 4 and 8 students. TIMSS 2011 gathered nationally representative samples of students in 63 countries and 14 benchmarking entities (including U.S. states). Fifty-two countries and seven benchmarking entities participated in the fourth grade assessment, and 45 countries and 14 benchmarking entities participated in the eighth grade assessment. While many European OECD countries did not participate in TIMSS when the more progressive PISA assessment was launched in 2000, the 2011 TIMSS survey involved many OECD countries including Finland, which is one of the highest scorers on PISA. Note that a companion survey to TIMSS, Progress in International Reading Literacy Study (PIRLS), assesses reading at grade 4. TIMSS and PIRLS collaborated on a common grade 4 household survey.

<table>
<thead>
<tr>
<th>Exhibit 3-3. TIMSS 2013 Questionnaires</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student Questionnaires (grads 4 &amp; 8):</strong> basic demographic information, their home environment, school climate for learning, and self-perception and attitudes toward learning mathematics and science.</td>
</tr>
<tr>
<td><strong>Home Questionnaire</strong> (grade 4 to homes of students participating in joint TIMSS &amp; PIRLS survey) preparations for primary schooling, including attendance in preschool and literacy- and numeracy-centered activities in the home before the child began school, such as reading books, singing songs, or playing with number toys. Parents answered questions about home resources in addition to information about their highest level of education and employment situations.</td>
</tr>
<tr>
<td><strong>Teacher Questionnaires:</strong> Asked students’ teachers about their education, professional development, and experience in teaching, coverage of the mathematics and science curriculum and about the instructional activities and materials used in the class.</td>
</tr>
<tr>
<td><strong>School Questionnaires:</strong> Principals answered questions about student demographic characteristics, the availability of resources, types of programs, and environments for learning in their schools.</td>
</tr>
<tr>
<td><strong>Curriculum Questionnaires:</strong> Answered by TIMSS 2011 National Research Coordinator questions centered on the organization and content of the curriculum in mathematics and science.</td>
</tr>
</tbody>
</table>

Source: TIMSS, 2011

TIMSS, like NAEP, is a general-purpose data collection from which indicators about key conditions of math and science education are derived. For purposes of indicator development, TIMSS includes contextual surveys that explore the “interplay of societal, school, and home environmental factors” on the achievement results.

The TIMSS contextual questionnaires (Exhibit 3-3) measure factors that affect student learning by administering questionnaires to students, parents or guardians (grade 4), teachers, schools, and the TIMSS national research coordinator (about the organization and content of the curriculum questionnaire). These questionnaires report on a variety major influences on student learning, including student engagement in classroom learning, home support for student learning, the scope and coverage of the mathematics or science curriculum, teacher preparation for mathematics or science instruction, and school resources and learning climate.
An extensive *home questionnaire* for the parents of students at grade 4 is of particular note in the 2011 TIMSS; there is no such survey in NAEP. The home questionnaire is part of a joint administration with the 2012 Progress in International Reading Literacy Study (PIRLS) for grade 4. It covers family well-being and learning opportunities and expectations for both math and reading. This contrasts with NAEP’s reliance on a brief and somewhat unreliable student questionnaire for grade 4 that covers home environment and family socio-economic status.

An example of the comprehensive home context information TIMSS gathers is the questions that ask parents about education activities prior to the child entering primary school (Exhibit 3-4). The left-hand column asks a parent/guardian to...
report on the frequency of various learning activities they did with their grade 4 child before the child began school. The right hand column asks about their child's reading ability before entering school. There are similar questions for mathematics. These data are subject to the normal error of retrospective responses, but the TIMSS analyses indicate they have predictive validity with current TIMSS scores.

Additional home questions inquire about current activities of the family in helping their child with math, reading and homework. There are also questions to the responder at the child’s home about their perceptions of a child’s school, personal reading habits, books in the home and how the child and parent talk together. Socio-economic status measures include questions about education, occupation, and income that are not accurately obtained from a fourth grader. TIMSS grade 8 responses about the home are from a student questionnaire.

A second feature of TIMSS, which differentiates it from NAEP’s treatment of background variables, is essential for indicator development: TIMSS creates numeric scales by combining answers to questions focused around a similar background condition. Because a single question rarely captures the richness and multi-dimensionality characterizing an education condition, TIMSS has created numeric scales that weight answers from multiple questions about a topic. The following are the Contextual Questionnaire Scales in the 2011 TIMSS for grade 4 mathematics:

- Home Resources for Learning
- Early numeracy activities before beginning primary school
- Could do early numeracy tasks when began primary school
- Instruction affected by mathematics resource shortages
- Teacher working conditions
- School emphasis on academic success –principal reports
- Safe and orderly school
- School discipline and safety
- Students bullied at school
- Confidence in teaching mathematics
- Teacher career satisfaction scale
- Students like learning mathematics
- Students confident in mathematics
- Collaborate to improve teaching
- Instruction to engage students in learning
- Students engaged in mathematics lessons
To illustrate these scales, Exhibit 3-5 displays the items for the condition of early numeracy activities before beginning primary school scale in the fourth grade. The responses to these six items are pooled through an IRT statistical procedure to yield a scale with a mean across all countries of 10 and a standard deviation of 2. Cut points were then established on the scale to create three categories of doing early numeracy activities often, sometimes, and never or almost never (Exhibit 3-5 bottom).

**International: OECD’s PISA (Program for International Student Assessment)**

The Organization for Economic Co-operation and Development (OECD) launched the OECD Program for International Student Assessment (PISA) in 1997 to monitor the outcomes of education through measuring student achievement on a regular basis. PISA chose age 15 because “at this age students are approaching the end of compulsory education in most OECD countries.” PISA is known to focus on the use of knowledge in everyday tasks and challenges in reading, mathematics and science. PISA assessments are administered every three years, with each assessment emphasizing one of the three subjects, while assessing the other two subjects less extensively. The latest released assessment 2009 focused on reading. The 2009
assessment covered 67 countries and jurisdictions and, included for the first time the high performing jurisdictions of Singapore and Shanghai, China.

Along with the assessment of reading, mathematics and science literacy, PISA, like NAEP and TIMSS, includes a set of contextual questionnaires that ask students and the principals of their schools to respond to background questionnaires of around 30 minutes in length. Note that unlike TIMSS, PISA does not include a teacher questionnaire. For 2009 PISA, the questionnaires yielded information on:

- *Students and their family backgrounds*, including their economic, social and cultural capital.
- *Aspects of students’ lives*, such as their attitudes towards learning, their habits and life inside school, and their family environment.
- *Aspects of schools*, such as the quality of the schools’ human and material resources, public and private control and funding, decision-making processes, staffing practices and the school’s curricular emphasis and extracurricular activities offered.
- *Context of instruction*, including institutional structures and types, class size, classroom and school climate and reading activities in class.
- *Aspects of learning and instruction* in reading, including students’ interest, motivation and engagement.

The 2009 PISA, unlike NAEP, offered three optional questionnaires.

- A *computer familiarity questionnaire* focusing on the availability and use of information and communications technology (ICT), including where ICT is mostly used, as well as on the students’ ability to carry out computer tasks and their attitudes towards computer use.
- “An *educational student career questionnaire* collecting additional information on interruptions of schooling and changes of schools, expected educational attainment and lessons or tutoring outside of school.”
- “A *parent questionnaire* focusing on a number of topics including the student’s past reading engagement, the parents’ own reading engagement, home reading resources and support, and the parents’ perceptions of and involvement in their child’s school.”

In the context of developing indicators PISA, like TIMSS but unlike NAEP, creates *scales* from multiple questions around a construct. But PISA goes beyond the TIMSS approach in studying the determinants of learning, by *estimating the quantitative relationship between the scales with learning*. This estimation process is illustrated for PISA scales for students’ *reading habits*, as an indicator of engagement in reading activities and *approaches to learning*, as an indicator of learning strategies (Exhibit 3-6). Each of these scales is made up of five components drawn from student responses to the student questionnaire.
The two reading-explanatory scales in turn are employed in a path model (Exhibit 3-7) that relates students’ socio-economic background and gender mediated through reading habits and approaches to learning to reading performance. For example, the difference that a unit change on the socio-economic background scale has on the reading performance TIMSS scale score consists of three effects. These are: a direct effect estimate of 25.6 reading score points, an indirect effect working through approaches to learning to reading performance of .02 x 30.7 reading score points, and another indirect effect that comes from SES working through reading habits of 0.2 x 25.8 reading score points. The total effect then is 25.6 + (0.2 x 30.7 = 6.4) + (0.2 x 25.8 = 5.4) = 37.4. One conclusion is that 68% (25.6/37.4) of the SES effect is direct.
The importance of particular indicator scales for learning is quantified in Exhibit 3-8, which shows the estimated net impact of socio-economic status if students were equally aware of effective reading approaches (i.e. summarizing strategies) as are students in the top quarter of the socio-economic scale. Thus, if students at the bottom quarter of the PISA index of economic, social and cultural status index had the same summarizing strategies for reading as students in the top quarter (Exhibit 3-8), they would diminish by one-third the total difference in reading levels between the bottom and third quartile of the socio-economic index.
Domestic: U.S. Department of Education’s Priority Performance Goals

Under the Government Performance and Results Act (GPRA), the United States Department of Education (USED) is required to measure and report the performance of each of its Congressionally authorized programs. While many of the USED program specific performance indicators are not relevant for NAEP indicator-development, at least two areas are.

One is the high-priority performance goals. As part of this annual performance indicator activity, federal agencies are required to identify a sub-set of high-priority performance goals that are achievable within an 18 to 24-month time frame, with strong execution. The following are the current USED K-12 high-priority indicators:

- **Evidence Based Policy:** Measuring Effectiveness and Investing in What Works: Implementation of a comprehensive approach to using evidence to inform the Department’s policies and major initiatives, including:
  - Increase by 2/3 the number of Department discretionary programs that use evaluation, performance measures and other program data for continuous improvement.
  - Implement rigorous evaluations for all of the Department’s highest priority programs and initiatives.
  - Ensure all newly authorized Department discretionary programs include a rigorous evaluation component.

- **Struggling Schools Reform:** National Models for School Reform: Identify as nationwide models 500 of the persistently lowest achieving schools initiating high-quality intensive reform efforts (e.g., turnarounds, restarts, transformations, or closures).

- **Effective Teaching:** World-Class Teaching and Learning: Improve the quality of teaching and learning by:
  - Increasing by 200,000 the number of teachers for low income and minority students who are being recruited or retained to teach in hard-to-staff subjects and schools with rigorous, transparent and fair processes for determining teacher effectiveness.
  - Ensuring that all States have in place comprehensive teacher evaluation systems, based on multiple measures of effectiveness including student growth, that may be used for professional development, retention, tenure, promotion, and compensation decisions.

- **Data Driven Decisions:** Improved Achievement and Decision-Making through Statewide Data Systems: All States implementing comprehensive statewide longitudinal data systems that link student achievement and teacher data and link K-12 with higher education data and, to the extent possible, with
pre-K and workforce data.

- **College and Career Ready Standards:** World-Class College- and Career-Ready Standards: All States collaborating to develop and adopt internationally benchmarked college- and career-ready standards.

While these identified high priority goals apply to federal programs, they, for the most part, deal with fundamental U.S. educational issues, ones that NAEP background questionnaire could inform. For example, NAEP can develop an indicator for data driven decisions. This would explore whether and how teachers use the data from their statewide longitudinal student outcome system to improve their practice. Principals could be queried whether they use the outcomes from the state longitudinal system to evaluate teachers. With respect to college and career ready standards, NAEP can go beyond adoption of these standards and track implementation challenges. In all three examples, NAEP can further ask about whether use of these measures has produced quantitative evidence of continuous improvement in the quality of teaching and the outcomes of students and how data use itself might be continually improved.

A second point of note is that USED has developed a set of school-level leading indicators for monitoring one of its new and controversial programs, School Improvement Grants (SIG) to support turning around low-performing schools. The idea of leading indicators is that they are precursors to turning around outcomes, in this case low-performing schools. NAEP indicators generated from its background variables might also be differentiated to single out those that are leading indicators of outcome changes based on research on student outcomes.

**Domestic: Education Week’s Quality Counts**

Quality Counts is an annual report prepared by Education Week describing education performance indicators state-by-state. The state-by-state focus, unlike national indicators, allows for comparisons of education indicator values among states with similar student populations. A state can also compare its education indicator values to states with the highest education performance. As we shall discuss below, much of Quality Counts data on educational quality is based on state-level laws and requirements. NAEP, by obtaining data directly from principals, teachers, and students in the field, NAEP could provide unique complimentary information about how the provisions Quality Counts identifies are being implemented state-by-state and for many major urban districts.

Quality Counts annually reports on State education performance in six state education system areas. These are performance indicators and, unlike the previously discussed reports, Quality Counts gives each state a letter grade to gauge their performance in each of the six topic areas:
- **Chance for success** indicators looks at the connection between education and beneficial outcomes at each stage of a person’s life. It covers indicators of education access, education outcomes and beneficial outcomes for early childhood, school years and adult educational outcomes.

- **School finance** indicators examine the level and equitable distribution of financial support within a state.

- **Transitions and alignment** indicators track state-policy efforts to coordinate the connections of K-12 schooling at three stages of education transition: early-childhood education, college readiness, and career readiness.

- **K-12 achievement** evaluates a state’s student performance on three dimensions: current state performance, improvements over time, and equity as measured by poverty-based achievement gaps. The achievement indicators are all drawn from NAEP and complemented by rates for high school and advanced placement.

- **Standards, assessment and accountability** develops quality criteria for state implementation of these results-focused elements of state education policy.

- **The teaching profession** covers three aspects of state policy: accountability for teacher quality; incentives and allocation; and efforts to build and support the capacity of the teaching workforce.

Each of the six topic areas is composed of a number of components. For example, the standards indicator is composed of two components: states having course or grade-specific standards and providing supplementary resources or guides to implement the standards. The School Accountability indicator is composed of five components including school ratings, statewide student-identification system, rewards for high performing schools and assistance to low-performing schools, and sanctions for low-performing schools. Like the scales described above for TIMSS, Quality Counts creates an overall average, in this case an equal weighting of the scores given to each component and then assigns a grade.

The Quality Counts data are state level and for the most part do not report on actual implementation at the school and classroom level of the indicators. Quality Counts does annually supplement the State reports with an online field survey of registered users of the Education Week website with responses including views of teachers, instructional specialists, principals and other building administrators. While very informative, this annual survey examines a special topic, only, which in 2013 spotlighted school social and disciplinary environment and is not state representative. NAEP data could fill indicator gaps regarding implementation. For example, how schools and teachers are actually responding to the Common Core standards or how school finance differences might translate into real differences in
Domestic: NCES’s Condition of Education

The Congress mandated that NCES produce an annual report on the *Condition of Education* to inform policymakers and the public about the current state and progress in key areas of education. The specifics of the report are left up to the Commissioner of Statistics. The latest 2013 report covers 42 indicators focused on four areas: population characteristics, participation in education, elementary and secondary education school characteristics and climate, and postsecondary education.

The following identifies the subset among the 42 indicators of most relevance for NAEP as those covering early childhood and elementary and secondary education.

**Indicator Area 1. Population Characteristics**
- Indicator 5, Percentage of children under 18 living in poverty (state and race ethnicity)

**Indicator Area 2. Participation in education**
- Spotlight on Preprimary Education: Kindergarten Entry Status: On-Time, Delayed-Entry, and Repeating Kindergartners
- Indicator 6. Enrollment Trends by Age
- Indicator 7. Early Education and Child Care Arrangements of Young Children (parents educational attainment)

**Elementary/Secondary Enrollment**
- Indicator 8. Public School Enrollment (state-level breakouts)
- Indicator 9. Charter School Enrollment (state-level breakouts)
- Indicator 10. Private School Enrollment
- Indicator 11. Racial/Ethnic Enrollment in Public Schools
- Indicator 12. English Language Learners (state-level breakouts)
- Indicator 13. Children and Youth With Disabilities

**Indicator Area 3. Elementary and Secondary Education**

**School Characteristics and Climate**
- Indicator 16. Characteristics of Public Elementary and Secondary Schools (Urbanicity data)
- Indicator 17. Concentration of Public School Students Eligible for Free or Reduced-Price Lunch (school poverty, urbanicity breakouts)
- Indicator 18. Rates of School Crime
- Indicator 19. Teachers and Pupil/Teacher Ratios

**Finance**
• Indicator 20. Public School Revenue Sources (state-level breakouts)
• Indicator 21. Public School Expenditures
• Indicator 22. Education Expenditures by Country

Assessments
• Indicator 23. Reading Performance
• Indicator 24. Mathematics Performance
• Indicator 25. Reading and Mathematics Score Trend
• Indicator 26. International Assessments (states participating in international assessments breakouts)

Student Effort, Persistence, and Progress
• Indicator 27. High School Course-taking
• Indicator 28. Public High School Graduation Rates (state, race/ethnicity breakouts)
• Indicator 29. Status Dropout Rates (race/ethnicity breakouts)

Transition to College
• Indicator 30. Immediate Transition to College (family income breakouts)

Several points about these indicators are relevant for NAEP. One, the Condition of Education indicators cover the entire education system and as such NCES limits their numbers to the most strategic indicators. These strategic indicator areas for K-12 are focused primarily on student characteristics, finances and educational outcomes. However, for those interested in the quality of elementary and secondary education, NAEP surveys of principals, teachers and students can provide a great deal of information about school-level processes and students attitudes, learning out-of-school and use of time not contained in the Condition of Education report.

Two, the Condition of Education as a mandated report to Congress tends to have a national focus, with only a few of the indicators at the State level and no data for specific urban districts. NAEP’s rich state-by-state and coverage of many urban districts can potentially add important disaggregated data to the Condition of Education content.

Third, NAEP has the potential to use its student assessment data to breakout Condition of Education indicators such as participation in education, school characteristics and climate and teacher characteristics by student proficiency levels.

Domestic: National Academy of Sciences’ Key National Education Indicators

The National Academy of Sciences is in the process of establishing a list of key Education Indicators to be carried out by the Congressional Commission on Key
National Indicators. As part of their work in progress, they have published the results of a workshop on prospective frameworks along with a candidate list of key national indicators.

### Exhibit 3-9 National Academy of Sciences Draft Indicators Framework

**TABLE 1-1 Framework for Education Indicators Developed to Guide the Workshop**

<table>
<thead>
<tr>
<th>Stages of Education/Learning</th>
<th>Indicators About Institutions, Service Providers, and Resources</th>
<th>Indicators About Individual-Level Behaviors, Engagement, and Outcomes</th>
<th>Indicators About Contextual Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth to age 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other forms of postsecondary education and training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lifelong, informal learning</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


The draft framework emerging from the National Academy of Education Indicators Workshop (Exhibit 3-9) covers five stages of learning: preschool, K-12 education, higher education, other postsecondary education and training, and lifelong or informal learning (learning that occurs outside the formal structures of the education system). It also identifies three sectors of education: institutions, service providers, and resources; individual-level behaviors, engagement, and outcomes; and contextual factors that influence learning. This is similar to a production function process where column one combines inputs and processes, column two is outcomes of the education process; and column three is the interaction of the first two factors with the context in which education takes place.

The first two of the five stages in Exhibit 3-9 are most relevant for NAEP’s pre-college grade focus. The *indicators suggested for the preschool stage (Exhibit 3-10)* are organized according to the Exhibit 3-9 framework into the providers and resources of education, individual student outcomes and contexts. At the preschool level, the institutions are the providers of early childhood education outside the home, the outcomes are both academic and social skills, and the context is the home environment including learning experiences from families and other caregivers. The NRC notes that the early childhood outcome measures could be through a NAEP assessment.
The indicators for K-12 education (Exhibit 3-11) also align with the Workshop
framework (Exhibit 3-9). These indicators include some common to the indicators above that report readily measured school services (teacher/pupil ratio) or student outcomes. But they also include a number of research-based school processes that require on-the-ground measurement from surveys or direct observations.
Examples of K-12 indicators proposed by the NAS workshop include:

- At the institution level include surveys of the safety and orderliness of the school climate and of the collaborative school community focused on student learning (using surveys of teachers, parents and students)
- At the teacher level of their mastery of content knowledge, quality of teacher-student interactions and proportion of teachers whose evaluations distinguish them from a basic standard, using measures of their contribution to student achievement and their professional practice.

NAEP surveys have the potential to provide address these rich in-depth information requirements.

**Implications**

The review of current indicator or indicator like reports yields a distillation of current practices that offer a range of possibilities for NAEP’s indicator development design. Important practices derived from prior reports include:

- **Indicator frameworks** that include organization by a combination of age/grade range, production function-like models, policy issues or provider level in system.
- **Indicator selection** that may focus on regularly reported information, enacted policies or may instead describe services, processes and climates by in-depth school, teacher and classroom surveys.
- Development of indicators in the form of *indicator scales* that statistically combine results from multiple aspects about an indicator.
- The estimation of an indicator’s importance in contributing to learning through *path models* or *other multivariate techniques* based on rigorous longitudinal research or more descriptive modeling based on cross-sectional data.
- The focus of indicator reporting at the *national level or at disaggregated sub-national state and major urban district* levels.
- Similarly, data can be reported nationally for all students or *disaggregated by student groups*.

The choices should be evidence based through direct and indirect relationships with key outcome measures.

These factors along with traditional considerations of validity and reliability of indicator measures are considered in the following chapter on the choices and recommendations in designing and implementing a NAEP indicator framework.
4. Proposed Design of Key Indicators Framework

A NAEP indicators framework specifies five key features in designing a NAEP key indicators system: the organizing structure for the indicators framework; the specification of particular indicators; the development of indicator measures; survey and sampling considerations; and the reporting of indicator results.

Indicator Organizing Structure

The indicator organizational structure guides the selection of indicators by specifying the categories that sort out the selection of the most important factors affecting student learning. Further, the organizational structure clarifies the relationships among factors to learning, which aids in choosing measures and drawing implications.

Several indicator organizational structures were presented in the section 3 review of current domestic and international indicator and large-scale assessment systems. These offer different perspectives on the education system from which to choose potential components. At the outset, it should be noted that no one indicator organization is necessarily best, but it depends upon the purpose and focus of indicator systems. Looking across these indicator perspectives, and in the context of NAEP’s focus on instruction, five desirable design features emerge in specifying an organizational structure applicable to the NAEP. These five features have been incorporated to form the proposed organizational structure for NAEP indicators in Exhibit 4-1, as follows:

1. Explicitly including indicators that represent the key education drivers emerging in response to changing education or workplace conditions. Indicators are most useful when they are used to monitor and continuously improve the education system to help achieve end outcomes and track responses in areas of major education change. Examples of drivers for the U.S. system might be preparing students with 21st century workplace skills, Common Core Standards, or instructional technology.

Focusing on major education drivers is consistent with the prior reports. Education At a Glance explicitly identifies policy issues to which the indicators relate. The NRC report begins its selection of K-12 indicators with a statement “that this system is the focus of many expectations, from producing responsible and productive citizens to boosting the nation’s standing in science and technology and its position with respect to its economic competitors.” EDWEEK explicitly builds tracking implementation of major policy reform areas, such as standards, assessment and accountability. Also note, that tracking education conditions in policy areas is not an endorsement of a policy approach but only a
consideration of the importance of monitoring and understanding responses to policy changes.

2. *Differentiating age/grade of instruction and learning.* The NRC report is explicitly organized around different stages of learning, from early childhood to adult and life-long learning. While NAEP is focused on K-12, the indicators framework should offer the possibility of including pre-school outcomes and experiences of students entering kindergarten. The framework may also want to differentiate early elementary, middle school and secondary school, as each has its own unique education aims and intervention priorities.

3. *Recognizing that the locus of education activity occurs at different levels at which education takes place* – student, teacher, school/classroom and system – and develops indicators for each level. Monitoring key education conditions translates into monitoring the key conditions at each of these levels. This is similar to the “Actors” identified in OECD’s Education At a Glance.

4. *Describing key education conditions in terms of education results (outcomes or outputs); the enablers which are the most important education factors producing education results; and the context and constraints within the education system that affects education results.* This focus on the broad elements in producing education is similar to the organizing structure used by Education At a Glance and the NRC analyses.

5. *Focusing on an indicator framework organization consistent with NAEP’s emphasis on instruction and learning* as contributors to the NAEP assessment results. This covers instruction and learning in both formal and informal settings. This emphasis on describing instruction and learning conditions is similar to the implicit indicators in TIMSS and PISA and to some extent the NRC proposed indicators. This focus differs from those of Education At a Glance or the Condition of Education, which tend to focus on the results of education rather than on instructional processes.
Indicator Selection

Indicator selection is the process of identifying key indicators that are essential to monitor on a regular basis. Parsimony in indicator selection is critical to prevent the indicator user from becoming overwhelmed in data with a loss of focus on priorities.

In selecting indicators for measurement, consideration should be given to:

- *Measuring what matters most.* Identifying what matters most should be guided by focusing on the education and context factors that research has shown bear an important relationship to results (Walberg, 2002). Variables that have high variance (e.g., differences in students’ family socioeconomic
status) are ones that often have high contributions to outcomes.

Also, measuring what matters can give priority to informing policies. For instance, a widespread impression is that U.S. schools are having difficulty filling vacancies for mathematics teachers. Exhibit 4-2 drawn from TIMSS suggests that at least for U.S. middle schools, only about 12 percent of U.S. principals are having at least some difficulty filling vacancies for mathematics teachers. This compares with other Western English-speaking countries of 41 percent of the principals having difficulty hiring math teachers in Australia, 37 percent in England, and 44 percent in New Zealand. The United States numbers are slightly lower than those of Korea, but higher than Singapore. Of course, principal responses across countries may differ in the criteria they use for determining the adequacy of a candidate.

- **Focusing on changing conditions.** Continually measuring a factor that is an important contributor to results but changes infrequently produces little information gain from regular monitoring. This factor might be measured every other time NAEP is administered and the most recent value would be built into indicators that use multiple factors. On the other hand, education conditions undergoing rapid change such as because of technology or new policies should be weighted high for regular indicator development.

*Taking a pyramid approach* to indicator selection. A pyramid perspective on information recognizes that different users have different information needs. The top of the pyramid is the most important measure of a condition, such as national averages. Beneath this *top-level indicator*, additional indicator measures may display results disaggregated such as by population group, state, district or type of secondary school. The pyramid may display further information that shows indicator components, such as numbers, geometry, measurement, algebra and statistics for mathematics.
A set of potential key national indicators are shown in Exhibit 4-2, which is obtained by filling in indicator selections in the cells in Exhibit 4-1. The indicator selections
are drawn from the different international and domestic indicator lists above and represent an organized menu of indicator choices to guide selection of current and potential indicators for NAEP. Consistent with NAEP, the indicator structure is focused primarily around variables at student, teacher and school/classroom and system levels that support learning outcomes across the three aspects of education conditions (Exhibit 4-3):

- **Results** indicators include student assessment outcomes (such as from NAEP), but also teacher evaluations that include student outcomes, and other outcomes such as secondary school completion and parent satisfaction with the school.

- The **enablers** reflect formal learning from different levels of education. These include students exposure to preschool; teachers’ knowledge and skills and their ability to apply them to create a challenging and supportive classroom learning environment; and school instructional time and student engagement in the content areas. Enablers also include system policies and regulations at district, state and national levels regarding teacher certification, standards, assessment and accountability.

- Context/constraints reflect factors not readily manipulable by the education systems, although conditions may be changeable with proper interventions, such as schools intervening in the home learning environment. These factors include learning at home and outside the school in formal and informal settings; factors influencing teacher quality including salaries and working conditions; and factors affecting the school learning environment including school safety, climate and class size.

**Indicator Measurement**

A sound measure for an indicator should meet criteria of validity, reliability, and consistency overtime.

**Validity.** A valid measure is one that adequately captures the underlying education condition of interest. Strong validity also depends on a good level of reliability. Occasionally a key indicator may be validly measured by a response to a single question, but more often a valid and robust indicator will be made up of multiple statistics each of which reflects an aspect of an education condition of interest. Combining multiple statistics such as responses from a number of questions around a topic into a larger comprehensive indicator measure or scale, is not an approach currently incorporated into NAEP background analyses.

*Multiple questions with the same response stem can produce a scale based on response frequencies.* We discussed how TIMSS grade 4 results for the early numeracy activities before beginning primary school are measured by responses to 6
questions about frequency of occurrence of these activities in terms of “often, sometimes or almost never” (Exhibit 3-5 above). The results in Exhibit 4-4 show how TIMSS creates a scale from these responses. For example, the scale for “often engaged in early numeracy activities” corresponds to parents responding to the six questions by indicating they do three of the six activities often and doing the other three sometimes. Within each country in Exhibit 4-4, students in families who on average across the six activities do these activities often score higher than students in families who sometimes do these six activities. They in turn score higher than students in families who never or almost never did these activities. However, these associations do not control for family background or other potentially important correlate factors.

Exhibit 4-4 Development of Indicator Scales From Multiple Questions

Source: IEA, TIMSS, 2011

Another example of creating a scale is used by Education Weeks Quality Counts is to give a a letter grade based on a numeric score to each component forming an indicator and to average these scores to produce the letter grade. For example, the state standards, assessments and accountability indicator category is composed of the three subcategories. The subcategory for assessment consists of four assessment policies and is shown below:
EDWEEK Assessment Policies tracked

- **Types of Test Items**: For each item type, results are reported by school grade span. EPE Research Center review of testing calendars and other materials from state education agency Web sites, as verified by states, 2011.

- **Assessments Aligned to Standards**: Subjects in which state uses assessments aligned to state standards. Results are reported for each core academic-subject area. Ibid.

- **Vertically Equated Assessments**: State tests for the 2011-12 school year have been vertically equated in grades 3-8 so that scores for each grade have been placed on a common metric. Results are reported for English/language arts and mathematics. EPE Research Center annual state policy survey, 2011.

- **Benchmark Assessments**: State provides educators with benchmark assessments or item banks linked to state standards. Assessments or test items may be developed by the state or an external organization.

This subcategory assessment is scored “reflecting the percent of tracked policies a state has implemented” and a numeric score is assigned the subcategory. The scores are then averaged across the state standards, assessment and accountability subcategory and letter grades are assigned based on scores (A=93 to 100, A-minus =90-92). Exhibit 4-5 shows a full Quality Counts display for Maryland, the highest rated state by 2012 Quality Counts.

**Reliability**. A reliable measure is one where the indicator measure produces consistent results when repeatedly measuring the same underlying condition. Of

---

**Exhibit 4-5. Quality Counts Use of Average Indicator Scales**

![Quality Counts Table](image)

Source: EDWeek, Quality Counts, 2012
particular note for measures based on surveys of background variables is that they often use qualitative responses to questions about frequency of occurrence of an activity. For instance, a question might ask for a response in terms of “a lot” or “a little” that are subject to interpretation and may be sensitive to respondent context. For example, in Exhibit 3.5 shows how TIMSS asks families about the frequency of early numeracy activities. A parent with only a high school education or less may interpret “a lot” different than for a parent with a college degree.

Qualitative responses may also be sensitive to the respondent. In the recently completed NAEP background paper on science Exhibit 4-6 was presented showing that teachers were more likely to indicate that resources within a school were “not at all available” than were principals in the same school. This is not surprising, as it is principals who are responsible for school resource availability. Conversely, a strong indicator of a positive school climate might be the degree to which principals and teaches agree on school quality factors including availability of science resources.

**Consistency.** A consistent measure requires using the same measure for an indicator over time. To the extent that measures are changed from time period to time period then it is unclear whether a change in an indicator condition comes about because of a real change in the underlying condition or because of changes in the measure. The Expert Panel report addressed this issue in recommendation 1d:

“Use consistency over time as a criterion to consider for question selection and wording. NAEP’s inconsistent inclusion of background questions weakens its potential to track trends and improvements within a subject area and topic.

For example, the Expert Panel found that only one-third of the 2011 questions asking about course offerings yielded at least a 6-year trend. No 2011 questions about curriculum or school resources were found on the 2005 or earlier questionnaires.
Sources of Indicator Data

Many of the indicator measures across the international and domestic reports are derived from surveys generated during the report process. A second source of indicator data draws upon secondary sources from other surveys. Both are considerations in specifying an indicator framework.

With respect to data from surveys, TIMSS and PISA, unlike NAEP, incorporate a household survey to directly obtain information about parents or guardians socio-economic status and about the home learning environment. TIMSS innovatively combined with PIRLS to develop a joint household survey for grade 4 students. The household survey included questions about:

- Early numeracy activities in the home before beginning primary school (See Exhibit 3-5)
- Early literacy activities in the home before beginning primary school
- Amount of exposure to preschool
- Family perception about child’s literacy and numeracy skills before entering primary school
- Family interaction with the child about school work
- Family perceptions about school
- Family literacy environment
- Family SES

These represent an extensive set of questions about student and family home learning and socio-economic environment conditions compared with NAEP, with its only source of grade 4 information derived from a brief grade 4 student questionnaire. As an example, Exhibit 4-7 displays the results from the home responses on how well their children could do when entering primary school on six numeracy tasks. In every country, average grade 4 mathematics achievement declined as parents reported that their entering primary children could do fewer tasks. This correlation lends external validation to parent responses.
A second characteristic of several of the indicator reports is the pooling of information across different surveys. The Condition of Education and Education At a Glance are drawn almost entirely from data series generated by from other surveys. Quality Counts is also a state-level amalgam of Education Week’s direct analyses of state policies combined with data from other surveys, including prominently featuring the NAEP assessment results. Currently, NAEP background variables only include those from the NAEP student, teacher and school surveys, but combining NAEP background data with data reported from other surveys is a potential source of expanded background reporting.
Reporting on Indicators

A challenge in reporting on indicators is that different audiences need different levels of depth of indicator reporting. In response, the business sector and more recently government have implemented digital dash boards. These provide a click-of-the-mouse approach to presenting different visual perspectives on key performance indicators.
Exhibit 4-8 illustrates the dashboard presented to web site users to the U.S. Department of Education’s performance indicators for teachers and leaders. There are three indicators with national summary results and direction of change shown for each. From the dashboard the interested user can drill down and get a chart of the data, state comparisons and details on an indicator including its specific goal statement, how it is measured, why it is important, etc.


A follow-on to this report will be a second indicator report to NAGB by December 2013 that will contain a recommended set of Key Indicators, examples using current NAEP data, and recommended improvements in NAEP data to strengthen indicator measurement or fill indicator gaps. The second report will build off of the findings in the initial report by addressing the following topics:

- Specify a NAEP Indicators Framework for Background Variables applicable across cognitive assessments.

- Identify indicators that are estimable using current NAEP data (and present a set of examples); indicators that could be developed through changes in the NAEP questionnaires; and indicators that would require a fundamentally new NAEP questionnaire.

- Identify where NAEP offers a unique data or measurement advantage over other indicator sources.

- Explore opportunities for combining NAEP with other NCES indicator-supporting data.

- Explore opportunities for aligning NAEP domestic indicators with the indicators generated by background variables in the international data collections discussed above to yield national and state comparisons with other countries.

- Explore how can NAEP reports best display a pyramid information approach along the lines of an indicator dashboard to provide the user with push-button access to top-level national measures or to more disaggregated measures by indicator component, student characteristics or jurisdictions.
• Assess how consistently the identified key NAEP education indicators have been measured by NAEP over time and identify challenges in fixing these definitions.
References


